

Contents

Preface	6
1 Chapter 1	8
Sets, Functions, Images and Preimages	8
1.0.1 Roadmap	9
1.0.2 Named moves	9
1.1 Sets, Subsets and Russell's Paradox	9
1.1.1 A naive idea, and what goes wrong with it	9
1.1.2 Sets and membership	10
1.1.3 Subsets	11
1.1.3.1 Two inclusions make an equality	14
1.1.3.2 Transitivity, rule by rule	14
1.1.4 Describing sets: listing and specification	16
1.1.5 The empty set	17
1.1.6 Computing with set-builders	18
1.1.7 Russell's paradox as a theorem	19
1.1.8 Exercises	20
1.1.8.1 A. Check your understanding	20
1.1.8.2 B. Practice	21
1.1.8.3 C. Going deeper	22
1.2 Unions, Intersections, Complements and Families	23
1.2.1 Union, intersection and difference	24
1.2.2 Laws of sets are laws of logic	25
1.2.3 Power sets	28
1.2.4 Families of sets	29
1.2.4.1 The smallest set closed under some steps	31
1.2.4.2 Worked examples with a real parameter	32
1.2.5 De Morgan's laws for families	33
1.2.6 Exercises	34
1.2.6.1 A. Check your understanding	34
1.2.6.2 B. Practice	35
1.2.6.3 C. Going deeper	36
1.3 Ordered Pairs, Products and Relations	38
1.3.1 Order matters: ordered pairs	39
1.3.2 Cartesian products	40
1.3.3 Products and the set operations	42
1.3.4 Relations	44
1.3.5 Exercises	47

1.3.5.1	A. Check your understanding	47
1.3.5.2	B. Practice	47
1.3.5.3	C. Going deeper	49
1.4	Functions, Composition, Injections and Surjections	50
1.4.1	What is a function?	51
1.4.2	When are two functions equal?	54
1.4.3	Families, sequences, and the set of all functions	55
1.4.4	Composition	57
1.4.5	Injective, surjective, bijective	58
1.4.5.1	Injective functions	58
1.4.5.2	Surjective functions	59
1.4.5.3	Bijjective functions	59
1.4.5.4	Worked examples	60
1.4.6	Composites of injections and surjections	62
1.4.7	Exercises	63
1.4.7.1	A. Check your understanding	63
1.4.7.2	B. Practice	64
1.4.7.3	C. Going deeper	66
1.5	Left, Right and Two-Sided Inverses	67
1.5.1	Undoing a function	68
1.5.2	Left and right inverses	68
1.5.3	A left inverse exists exactly for injective functions	70
1.5.4	A right inverse forces surjectivity	72
1.5.5	Does every surjection have a right inverse?	73
1.5.6	Two-sided inverses	74
1.5.7	What the section proved	77
1.5.8	Exercises	77
1.5.8.1	A. Check your understanding	77
1.5.8.2	B. Practice	78
1.5.8.3	C. Going deeper	79
1.6	Images and Preimages	81
1.6.1	A question about squares	82
1.6.2	Images and preimages	82
1.6.3	Computing images and preimages	84
1.6.4	Sets of numbers as images	85
1.6.5	Preimages respect every operation	86
1.6.6	Images respect unions	88
1.6.7	Images and intersections	89
1.6.8	There and back	91
1.6.9	Two readings of the inverse symbol	92
1.6.10	Exercises	92
1.6.10.1	A. Check your understanding	92
1.6.10.2	B. Practice	93
1.6.10.3	C. Going deeper	95
1.6.10.4	R. Review	97

2	Chapter 2	101
	Equivalence Relations, Orders, and Induction	101
2.0.1	Roadmap	102
2.0.2	Named moves	102
2.1	The Natural Numbers and Well-Ordering	103
2.1.1	What we assume about the natural numbers	103
2.1.2	The arithmetic of the natural numbers	104
2.1.3	The integers	106
2.1.4	Well-ordering	107
2.1.5	Induction and well-ordering are the same principle	109
2.1.6	Sets of integers	111
2.1.7	Exercises	111
2.1.7.1	A. Check your understanding	111
2.1.7.2	B. Practice	112
2.1.7.3	C. Going deeper	114
2.2	Induction, Strong Induction and the Smallest Counterexample	116
2.2.1	Induction from any integer	116
2.2.2	The smallest counterexample	118
2.2.3	Divisibility and the division algorithm	118
2.2.4	Lowest terms and prime factors	120
2.2.5	How induction proofs go wrong	122
2.2.6	Exercises	124
2.2.6.1	A. Check your understanding	124
2.2.6.2	B. Practice	125
2.2.6.3	C. Going deeper	126
2.3	Definition by Recursion	128
2.3.1	Definitions that define something	129
2.3.2	Starting elsewhere, looking two steps back	131
2.3.3	Sums, products and powers	133
2.3.4	Recursion, then induction	135
2.3.5	When the claim is too weak to carry its step	138
2.3.6	Exercises	139
2.3.6.1	A. Check your understanding	139
2.3.6.2	B. Practice	140
2.3.6.3	C. Going deeper	141
2.4	Properties of Relations	144
2.4.1	Four properties	144
2.4.2	Divisibility	145
2.4.3	A table of examples	146
2.4.4	Exercises	148
2.4.4.1	A. Check your understanding	148
2.4.4.2	B. Practice	148
2.4.4.3	C. Going deeper	150
2.5	Equivalence Relations and Partitions	151
2.5.1	Arithmetic on a clock	152
2.5.2	Equivalence relations	152
2.5.3	Equivalence classes	154

2.5.4	The classes form a partition	155
2.5.5	Every partition comes from one equivalence relation	157
2.5.6	Congruence modulo n	158
2.5.7	Two more sources of equivalence relations	159
2.5.8	Operations on classes	161
2.5.9	Exercises	162
2.5.9.1	A. Check your understanding	162
2.5.9.2	B. Practice	163
2.5.9.3	C. Going deeper	164
2.6	Partial Orders, Maximal and Greatest Elements	166
2.6.1	Partial orders	166
2.6.2	Bounds, greatest and maximal elements	168
2.6.3	Least upper bounds, a first look	171
2.6.4	Well-ordered sets	172
2.6.5	Exercises	173
2.6.5.1	A. Check your understanding	173
2.6.5.2	B. Practice	174
2.6.5.3	C. Going deeper	175
2.6.5.4	R. Review	177
3	Chapter 3	180
	Cardinality: Countable and Uncountable Sets	180
3.0.1	Roadmap	181
3.0.2	Named moves	181
3.1	Equinumerous Sets and Finite Sets	182
3.1.1	Same size without counting	182
3.1.2	Finite sets	184
3.1.3	The pigeonhole principle	185
3.1.4	Subsets and unions of finite sets	187
3.1.5	No finite set is equinumerous with a proper subset	190
3.1.6	Exercises	192
3.1.6.1	A. Check your understanding	192
3.1.6.2	B. Practice	192
3.1.6.3	C. Going deeper	194
3.2	Countable Sets	197
3.2.1	Lists	197
3.2.2	Subsets of the natural numbers	198
3.2.3	Three ways to say countable	199
3.2.4	Closure properties	200
3.2.5	Pairs of natural numbers	202
3.2.6	The integers and the rationals	204
3.2.7	Exercises	205
3.2.7.1	A. Check your understanding	205
3.2.7.2	B. Practice	206
3.2.7.3	C. Going deeper	208
3.3	Countable Unions and the Axiom of Choice	210
3.3.1	A union of countably many lists	210

3.3.2	The obstacle, rule by rule	211
3.3.3	The axiom of choice	211
3.3.4	When choice is not needed	212
3.3.5	Right inverses	213
3.3.6	Countable unions	214
3.3.7	Unions without choice	215
3.3.8	Algebraic numbers	216
3.3.9	Exercises	217
3.3.9.1	A. Check your understanding	217
3.3.9.2	B. Practice	218
3.3.9.3	C. Going deeper	219
3.4	Diagonal Arguments: Uncountable Sets and Cantor's Theorem	221
3.4.1	A sequence that is not on the list	222
3.4.2	Cantor's theorem	223
3.4.3	Decimal expansions	224
3.4.4	The real numbers are uncountable	226
3.4.5	Transcendental numbers	227
3.4.6	Exercises	228
3.4.6.1	A. Check your understanding	228
3.4.6.2	B. Practice	228
3.4.6.3	C. Going deeper	229
3.5	Comparing Sizes: the Schröder–Bernstein Theorem	231
3.5.1	Comparing sizes with injections	231
3.5.2	The Schröder–Bernstein Theorem	233
3.5.3	Intervals of real numbers	235
3.5.4	Exercises	237
3.5.4.1	A. Check your understanding	237
3.5.4.2	B. Practice	238
3.5.4.3	C. Going deeper	239
3.6	Infinite Sets	241
3.6.1	Every infinite set contains a copy of the natural numbers	241
3.6.2	Consequences: the smallest infinite size	243
3.6.3	Infinite sets are the same size as a proper part	243
3.6.4	Exercises	244
3.6.4.1	A. Check your understanding	244
3.6.4.2	B. Practice	245
3.6.4.3	C. Going deeper	246
3.6.4.4	R. Review	248
3.7	Gate 0: Logic and Sets Examination	250
3.7.1	Instructions	250
3.7.2	Definition drill	250
3.7.3	The problems	252
3.7.4	Remediation map	260

Preface

Every definition of analysis is a sentence about sets and functions. *Logic and Proof* supplied the grammar of such sentences: the connectives, the quantifiers, and the prose that carries their rules. This volume supplies the nouns: sets and functions, relations and orders, the natural numbers and induction, and the sizes of infinite sets.

Who it is for. The reader we have in mind has worked through *Logic and Proof*, or has its habits from elsewhere: reading a statement with its quantifiers in place, proving a “for every” by taking an arbitrary element, naming a witness before using it. The chapters point back to *Logic and Proof* where they use one of its rules, so a reader who is unsure of a step can look it up.

What it covers.

Chapter	What it does
1	Sets and the working rules that say which sets exist, Russell’s paradox, families of sets, products and relations, functions and their inverses, images and preimages
2	The arithmetic of $\mathbb{N}$ and $\mathbb{Z}$ from induction, well-ordering, induction as a craft, definition by recursion, equivalence relations and partitions, partial orders
3	Finite and countable sets, the axiom of choice, diagonal arguments and Cantor’s theorem, the Schröder–Bernstein theorem, infinite sets; and Gate 0
4	Zorn’s lemma

Chapters 1–3 are written; Chapter 4 is not yet. The real numbers are used throughout as familiar objects, with their school arithmetic and order, and every use of a school fact is named where it occurs. *Analysis on the Real Line* turns those facts into consequences of axioms, and *Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* builds the numbers out of sets.

How to read it. Read the chapters in order: each uses the ones before it. Every section has Quick checks with folded answers; answer each one before you unfold it. The exercises of a section come in three groups, A to check understanding, B for practice and C to go deeper, and the last section of each chapter adds R items that reach back to earlier chapters. A chapter’s opening page lists what it needs, with links, and the time it takes.

Gate 0. Chapter 3 ends with an examination on *Logic and Proof* and the first three chapters of this volume: a definition drill and six problems, closed book, with a marking scheme and a

map that sends each lost problem back to named sections. Pass it before starting *Analysis on the Real Line*. Chapter 4 is not needed there.

Chapter 1

Sets, Functions, Images and Preimages

Every definition of analysis is a sentence about sets and functions. “The supremum of A is the least element of the set of upper bounds of A .” “A function is continuous if the preimage of every open set is open.” *Logic and Proof*, Chapters 1 to 5 supplied the grammar of such sentences: the connectives, the quantifiers, and the prose that carries their rules. This chapter supplies the nouns: sets and subsets, unions and intersections of whole families of sets, ordered pairs and products, relations, functions, inverses, images and preimages. It is also where the logic of the first five chapters becomes visible inside mathematics. Inclusion is a “for every” around an “implies”, a union is an “or”, the union of a family is a “there exists”, and a function is a “for every there exists exactly one”. *The laws of sets are the laws of logic, read through membership*, and the chapter uses this from its second section to its last.

Two surprises drive the chapter. The first is that the naive rule “any property defines a set” is contradictory (Russell’s paradox), so we must say which sets exist, and every set in this book is cut out of a set already given. The second is the chapter’s motivating problem. *For $f(x) = x^2$, is $f(A \cap B) = f(A) \cap f(B)$? Is $f^{-1}(C \cap D) = f^{-1}(C) \cap f^{-1}(D)$? Why is the second always true and the first not?* The answer is a single quantifier: membership in an image hides a “there exists”, and membership in a preimage hides nothing. Preimages therefore respect every set operation, and images only unions. This asymmetry is why analysis defines continuity by preimages.

What you need. *Logic and Proof*, Chapters 1–5. From *Logic and Proof*, Chapters 1 and 2: the rules for the connectives and the derived laws used as rules, above all distributivity (Logic and Proof, Proposition 1.6.1), De Morgan’s laws (Logic and Proof, Proposition 2.5.5), an iff as two implications (Logic and Proof, Proposition 2.4.1), the contrapositive (Logic and Proof, Proposition 2.3.2), vacuous implication (Logic and Proof, Proposition 2.1.2), and derived rules (Logic and Proof, Definition 2.3.1). From *Logic and Proof*, Chapter 3: the standard equivalences (Logic and Proof, Proposition 4.5.1) and completeness (Logic and Proof, Theorem 4.8.4). From *Logic and Proof*, Chapter 4: the rules for “for every” and “there exists”, bounded quantifiers (Logic and Proof, Definition 5.1.1) and their rules (Logic and Proof, Proposition 6.1.1), the rules for equality (Logic and Proof, Definition 5.7.1), unique existence (Logic and Proof, Definition 5.8.1, Logic and Proof, Proposition 5.8.1), and the laws for negating quantifiers. From *Logic and Proof*, Chapter 5: plain-prose proofs, the convention on which steps may be silent, and the catalogue of what is not allowed. The number sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ are used as familiar sets, with their school arithmetic and order.

Time. About 35 hours on the core track (all of A, one B and one C per section, the exit checks);

about 45 hours for everything.

1.0.1 Roadmap

- **Sets, subsets and Russell's paradox.** What a set is allowed to be, equality by elements, the subset relation and the element-chasing proofs that establish it, describing sets by a property cut out of an ambient set, the empty set, and why no set contains every set.
- **Unions, intersections, complements and families.** The set operations as the connectives read through membership, the algebra of sets as propositional logic, power sets, unions and intersections of arbitrary families, and De Morgan's laws for families, including the empty one.
- **Ordered pairs, products and relations.** An object that remembers order, the product $A \times B$ and how it interacts with the set operations, and relations as sets of pairs.
- **Functions, composition, injections and surjections.** A function as a graph with a declared target, when two functions are equal, composition, and the two questions of injectivity and surjectivity, with the squaring map as three different functions.
- **Left, right and two-sided inverses.** A left inverse exists exactly for injective maps (on a nonempty domain), a right inverse forces surjectivity, bijections are the maps with a two-sided inverse, and the question whether every surjection has a right inverse, which needs a new principle.
- **Images and preimages.** What a function does to sets, why preimages respect every operation and images only unions, where injectivity and surjectivity repair the failures, and sets of numbers such as $-A$ and $A + B$ as images.

1.0.2 Named moves

- **Element-chasing.** To prove $A \subseteq B$: “let $x \in A$ ”, unwind the definition of A , and rewind into the definition of B . To prove $A = B$: two inclusions.
- **Unwind the image.** “ $y \in f(A)$ ” means “there is $a \in A$ with $f(a) = y$ ”: name the witness before using it. A preimage has no witness to name.
- **Refute with a small witness.** A two- or three-element set, the map $x \mapsto x^2$, the empty set.
- **Build the inverse from the definition.** For a left inverse, send $f(x)$ back to x and everything else to one fixed point.

1.1 Sets, Subsets and Russell's Paradox

Logic and Proof, Chapters 1 to 5 built the grammar of mathematical statements: the connectives, the quantifiers, and the prose that carries their rules. The statements themselves have so far been about numbers, and occasionally about “collections” taken in their everyday sense. Analysis talks about collections all the time: the set of upper bounds of a set of numbers, the set of points where a function is positive, the set of all formulas. This section says what a set is allowed to be, what it means for one set to lie inside another, and how sets are described. One new notion carries the weight, the **subset**; the rest is the vocabulary it needs.

We begin with the simplest possible idea of a set, and we will see it fail.

1.1.1 A naive idea, and what goes wrong with it

In everyday speech a collection is given by a property. “The students in this room”, “the even numbers”, “the prime numbers below 100”: in each case we say what an object must satisfy,

and the collection is everything that satisfies it. The naive rule is:

For every property, there is a set whose elements are exactly the objects having that property.

It sounds harmless, and for the properties of everyday mathematics it gives the right answer. But a property may talk about sets themselves, and then the rule can be turned against itself. A set may be an element of another set: the set $\{\mathbb{N}, \mathbb{Z}\}$ has two elements, and each of them is a set. So it makes sense to ask of a set x whether $x \in x$, and the property “ $x \notin x$ ” is a property like any other. Most sets we know have it: the set of even numbers is not an even number, so it is not an element of itself.

Apply the naive rule to this property. It promises a set

$$R = \text{the set of all } x \text{ with } x \notin x.$$

Now ask whether $R \in R$. By the way R was made, an object belongs to R exactly when it is not an element of itself; applied to the object R , this says $R \in R$ exactly when $R \notin R$. A statement that is equivalent to its own negation leads to a contradiction by the rules of *Logic and Proof*, Chapter 2 alone; lines 4–11 of the derivation after Russell's theorem, at the end of this section, carry out exactly this argument, with an extra conjunct carried along. So the naive rule is contradictory: it cannot be part of mathematics.

This is **Russell's paradox**. We hold it until the end of the section, where it becomes a theorem, Russell's theorem, about the sets that do exist. First we set out what does work.

1.1.2 Sets and membership

We do not define “set” or “element”. They are the **primitive** notions of this chapter, in the way that “statement” was primitive in *Logic and Proof*, Chapter 1: we say how they are used, not what they are made of. A **set** is a collection of objects, its **elements**, and we write

$$x \in A$$

for “ x is an element of A ”, or “ x belongs to A ”, or “ x lies in A ”. Its negation is written $x \notin A$. The objects may be numbers, or sets, or (later in this chapter) pairs and functions.

In the language of *Logic and Proof*, Chapter 4, “ $\in$ ” is a two-place predicate: $x \in A$ is a statement about the two objects x and A , true or false like $x < y$. So everything *Logic and Proof*, Chapter 4 says about predicates, quantifiers, and their rules applies to it unchanged. A statement such as “every element of A is positive” is $\forall x (x \in A \rightarrow x > 0)$, which the bounded notation of *Logic and Proof*, Definition 5.1.1 shortens to $\forall x \in A (x > 0)$.

Two conventions fix what kind of objects we deal with. First, the number sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ are **given**: we use them as familiar sets, with their school arithmetic and order, as the earlier chapters did. (*Analysis on the Real Line*, Chapter 1 turns the facts we use about $\mathbb{R}$ into consequences of axioms.) Second, **numbers are not sets**: a number has no elements, and no number is equal to a set. So $1 \neq \{1\}$, and $z \in 5$ is false for every z . *Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* , Chapters 1, 2 and 5 explain how numbers can be built out of sets; nothing in this volume depends on how that is done.

The single most important fact about sets is that **a set is determined by its elements**. It has no other features: no order in which the elements come, no count of how often each is listed, no color. Two sets with the same elements are the same set.

Definition 1.1.1. Equality of Sets

Sets A and B are equal, $A = B$, if and only if they have the same elements:

$$\forall x (x \in A \leftrightarrow x \in B).$$

In words: every element of A is an element of B , and every element of B is an element of A . This principle is called **extensionality**: a set is fixed by its extension, the things it contains, and not by the words used to describe it. The sign $=$ here is not new: it is the equality of Logic and Proof, Definition 5.7.1, “names of the same object”. The direction from $A = B$ to the same elements is $=E$, with φ the formula $x \in z$ and z the variable replaced (and symmetry, Logic and Proof, Proposition 5.7.1, for the way back from B to A). The converse direction is the principle just stated, that a set has no features beyond its elements; it is what lets an equation proved from “same elements” be used with $=E$ like any other. The principle is an “if and only if”, so it is used in both directions: to prove $A = B$, prove the $\forall$ statement; from $A = B$, conclude $x \in A \leftrightarrow x \in B$ for any x we like.

Sets with finitely many elements can be written by **listing** their elements between braces: $\{1, 2, 3\}$ is the set whose elements are 1, 2, and 3, and nothing else. Extensionality has two immediate consequences for such lists.

- **Repetition does not matter.** $\{1, 1, 2\} = \{1, 2\}$. An object x belongs to the left side exactly when $x = 1$ or $x = 1$ or $x = 2$, and to the right side exactly when $x = 1$ or $x = 2$; the two conditions are equivalent, since “ $x = 1$ or $x = 1$ ” says no more than “ $x = 1$ ”.
- **Order does not matter.** $\{1, 2\} = \{2, 1\}$. Here the two conditions are “ $x = 1$ or $x = 2$ ” and “ $x = 2$ or $x = 1$ ”, which are equivalent because $\vee$ commutes (Logic and Proof, Example (First derivations with “or”)).

Order and repetition are exactly what a list has and a set does not. When order matters we will need a different object, the ordered pair of Section 1.3.

A set with one element, $\{a\}$, is a **singleton**. A singleton and its element are different things: $\{1\}$ is a set, and 1 is a number, so $1 \neq \{1\}$ by our convention. For sets the difference is just as real. The set $\{\mathbb{N}\}$ has exactly one element, namely $\mathbb{N}$, while $\mathbb{N}$ has infinitely many; so they differ, by extensionality, as $2 \in \mathbb{N}$ and $2 \notin \{\mathbb{N}\}$.

Quick check. Are the sets $\{1, 2, 2, 3\}$, $\{3, 2, 1\}$, and $\{1, \{2\}, 3\}$ equal?

Solution. The first two are equal: each has exactly the elements 1, 2, 3, and repetition and order do not count. The third is different: its elements are 1, $\{2\}$, and 3. The number 2 lies in the first set, but not in the third, since $2 \neq 1$, $2 \neq 3$, and $2 \neq \{2\}$ (a number is not a set). So by Definition 1.1.1 the third set differs from the other two.

1.1.3 Subsets

Most statements about sets in analysis are not equalities but **inclusions**: every Cauchy sequence converges, every continuous function on $[a, b]$ is bounded, every upper bound of a set is at least its supremum. Each has the shape “everything of this kind is also of that kind”. Before a definition, look at one such statement in the language of sets. The multiples of 6 and the multiples of 3

form sets

$$S = \{\dots, -12, -6, 0, 6, 12, \dots\}, \quad T = \{\dots, -6, -3, 0, 3, 6, \dots\},$$

and every element of S is an element of T , because $6k = 3 \cdot (2k)$. The sets are not equal ($3 \in T$ but $3 \notin S$), yet one sits inside the other. This relation between sets deserves a name.

A set is a subset of another if everything in the first is also in the second.

Definition 1.1.2. Subset

Let A and B be sets. A is a **subset** of B , written $A \subseteq B$, if **every** element of A is an element of B :

$$\forall x (x \in A \rightarrow x \in B).$$

We also say that A is **contained in** B , or that B **contains** A . If $A \subseteq B$ and $A \neq B$, then A is a **proper subset** of B , written $A \subsetneq B$.

In words, clause by clause. The quantifier $\forall x$ ranges over every object whatsoever, not over the elements of A only; the implication $x \in A \rightarrow x \in B$ restricts attention to the elements of A , and says nothing about objects outside A . With the bounded notation of Logic and Proof, Definition 5.1.1, the same condition is $\forall x \in A (x \in B)$: “for every x in A , x is in B ”. The symbol $\subseteq$ allows equality, as $\leq$ does for numbers; the symbol $\subsetneq$ excludes it, as $<$ does. This book never writes $\subset$, which some books use for $\subseteq$ and others for $\subsetneq$.

The definition is a $\forall$ around an $\rightarrow$, and its rules are the rules of those two symbols. To **prove** $A \subseteq B$, we prove the $\forall$ statement: by the rule $\forall I$ of *Logic and Proof*, Chapter 4 we open a box for an arbitrary x (in prose, “let x be arbitrary”), and inside it we prove $x \in A \rightarrow x \in B$ by $\rightarrow I$ (Logic and Proof, Definition 1.4.1): assume $x \in A$ and reach $x \in B$. The two openings are always written together, as “let $x \in A$ ”. To **use** $A \subseteq B$, we apply $\forall E$ (*Logic and Proof*, Chapter 4) at the element we care about, and then $\rightarrow E$: from $A \subseteq B$ and $a \in A$, conclude $a \in B$. This gives the chapter’s first named move.

To show $A \subseteq B$, take an arbitrary element of A and show that it lies in B .

We call this **element-chasing**: we follow one element from A to B , unwinding the definition of A to find out what we know about it, and rewinding the definition of B to see what we must show. Here is the move three times: two ordinary inclusions, then the degenerate one. In each, the first sentence is the pair of openings, $\forall I$ and $\rightarrow I$, and the rest reaches $x \in B$.

Example. Three inclusions

- (a) Show that $[1, 2] \subseteq [0, 3]$.
- (b) Let S be the set of integer multiples of 6 and T the set of integer multiples of 3, as above. Show that $S \subseteq T$. Here $d \mid n$, “ d divides n ”, means that $n = dk$ for some integer k , so S consists of the integers x with $6 \mid x$, and T of those with $3 \mid x$.
- (c) Let A be any set. Show that $A \subseteq A$.

Solution.

- (a) Let $x \in [1, 2]$. Then $1 \leq x \leq 2$, by the meaning of the interval. Hence $0 \leq 1 \leq x$ and $x \leq 2 \leq 3$, so $0 \leq x \leq 3$, that is, $x \in [0, 3]$. Since x was arbitrary, $[1, 2] \subseteq [0, 3]$.
- (b) Let $x \in S$, that is, let x be an integer with $6 \mid x$. By the definition of $\mid$, there is an

integer k with $x = 6k$; choose such a k (this is $\exists E$ of *Logic and Proof*, Chapter 4, and the name k is new). Then $x = 3 \cdot (2k)$, and $2k$ is an integer, so $3 \mid x$, with $2k$ as the witness ($\exists I$). Hence $x \in T$. Since x was arbitrary, $S \subseteq T$.

- (c) Let $x \in A$. Then $x \in A$. Since x was arbitrary, $A \subseteq A$. The box of $\rightarrow I$ opens on $x \in A$ and closes on the same line; in a derivation it is the one-line box that proves $x \in A \rightarrow x \in A$.

Part (c) is the degenerate case: every set is a subset of itself. It matters because it shows that $\subseteq$ allows equality, and so “ $A \subseteq B$ ” alone never tells us that B has anything A lacks. To say that A is a **proper** subset, $A \subsetneq B$, we also need an element of B outside A . In part (b) that element is 3: it lies in T , not in S .

A nonexample by minimal change. Replace 6 by 4 and 3 by 6 in (b): is the set of multiples of 4 a subset of the set of multiples of 6? The first sentence of the proof still works (let $x = 4k$), but the last step fails: $4k$ need not be 6 times an integer. One element is enough to show the failure: 4 is a multiple of 4, and it is not a multiple of 6. What fails is the clause “**every** element of A ” of Definition 1.1.2; here it fails at the element 4.

That last observation is worth recording in symbols, because it is how every inclusion is refuted. By the negation laws of *Logic and Proof*, Chapter 4 ($\neg \forall x \varphi$ is equivalent to $\exists x \neg \varphi$, and $\neg(\varphi \rightarrow \psi)$ to $\varphi \wedge \neg \psi$),

$$A \not\subseteq B \quad \text{means} \quad \exists x (x \in A \wedge x \notin B).$$

To refute an inclusion, exhibit one element of the first set that is not in the second.

Why this definition. The definition says nothing about objects outside A ; that is the implication doing its job. If the $\rightarrow$ were replaced by $\wedge$, the condition $\forall x (x \in A \wedge x \in B)$ would say that **every object** lies in A and in B , which is false for every set we will meet (Russell’s theorem at the end of this section shows that no set contains every set). This is the reason why a bounded “for every” always unfolds to an implication (Logic and Proof, Definition 5.1.1).

Warning. Element or subset? The symbols $\in$ and $\subseteq$ relate different things. $x \in A$ says that x is one of the objects collected in A ; $B \subseteq A$ says that every object collected in B is also collected in A . For example $1 \in \{1\}$ and $\{1\} \subseteq \{1\}$, but $\{1\} \notin \{1\}$: the only element of $\{1\}$ is the number 1, which is not the set $\{1\}$. And “ $1 \subseteq \{1\}$ ” is not a comparison worth making: 1 is a number, not a set, so the definition of $\subseteq$ does not apply to it. Before writing either symbol, ask whether the thing on the left is **one of the objects** in the set on the right, or a **collection** of such objects.

Sometimes a set is both an element and a subset of the same set. In $C = \{1, \{1\}\}$, the set $\{1\}$ is an element (it is listed) and also a subset (its only element, 1, is listed too). Change one entry: in $\{1, 2\}$, the set $\{1\}$ is still a subset, but it is no longer an element, since the two elements are the numbers 1 and 2. Whether $\{1\} \in D$ and whether $\{1\} \subseteq D$ are two separate questions, and each is answered by looking at the elements of D .

Warning. Checking some elements proves nothing about all of them. To prove $A \subseteq B$, checking that a few elements of A lie in B is not enough, however many are checked. “The integers 0, 12, 24 and -12 are multiples of 4 and of 3” does not prove that the set of multiples of 4 is a subset of the set of multiples of 3, and that inclusion is false (4 is a witness). The error is entry (N1) of the catalogue of *Logic and Proof*, Chapter 5: verifying

instances and concluding a $\forall$. The rule $\forall I$ needs an **arbitrary** element, one about which nothing is known except that it lies in A .

1.1.3.1 Two inclusions make an equality

The first payoff of the definition connects it with equality. An equality of sets is an $\leftrightarrow$ under a $\forall$, and by Logic and Proof, Proposition 2.4.1 an $\leftrightarrow$ is two implications. So an equality should be two inclusions.

Proposition 1.1.1. Equality Is Two Inclusions

For all sets A and B , $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.

Idea. Unwind all three definitions. $A = B$ is $\forall x (x \in A \leftrightarrow x \in B)$; the right side is $\forall x (x \in A \rightarrow x \in B)$ and $\forall x (x \in B \rightarrow x \in A)$. Inside the $\forall$, the statements are $P \leftrightarrow Q$ on one side and $P \rightarrow Q, Q \rightarrow P$ on the other, with P the statement $x \in A$ and Q the statement $x \in B$; Logic and Proof, Proposition 2.4.1 says exactly that these are interchangeable. So the proof is that proposition applied at an arbitrary x , once in each direction.

Proof. $(\Rightarrow)$ Suppose $A = B$. Let x be arbitrary. By Definition 1.1.1, $x \in A \leftrightarrow x \in B$, so by the first half of Logic and Proof, Proposition 2.4.1 (with P the statement $x \in A$ and Q the statement $x \in B$, used as a derived rule, Logic and Proof, Definition 2.3.1), $x \in A \rightarrow x \in B$ and $x \in B \rightarrow x \in A$. Since x was arbitrary, $\forall x (x \in A \rightarrow x \in B)$ and $\forall x (x \in B \rightarrow x \in A)$; by Definition 1.1.2, $A \subseteq B$ and $B \subseteq A$.

$(\Leftarrow)$ Suppose $A \subseteq B$ and $B \subseteq A$. Let x be arbitrary. By Definition 1.1.2, $x \in A \rightarrow x \in B$ and $x \in B \rightarrow x \in A$, so by the second half of Logic and Proof, Proposition 2.4.1, $x \in A \leftrightarrow x \in B$. Since x was arbitrary, $A = B$ by Definition 1.1.1. ■

The name of the proposition in words is **antisymmetry** of $\subseteq$: as for $\leq$ on numbers, $A \subseteq B$ and $B \subseteq A$ together force equality. It licenses the move that proves almost every set identity in this chapter.

To show $A = B$, show $A \subseteq B$ and $B \subseteq A$: two element-chases, one in each direction.

A proof of this kind is written in two paragraphs, labeled $(\subseteq)$ and $(\supseteq)$, exactly as a proof of an “if and only if” is written in two paragraphs labeled $(\Rightarrow)$ and $(\Leftarrow)$ (*Logic and Proof*, Section 2.4). The two paragraphs are two separate boxes: neither may use the other’s assumption.

1.1.3.2 Transitivity, rule by rule

The second payoff is that inclusions chain, as $\leq$ does. Its proof is the first element-chase of the chapter written with every rule named, and we give it at two levels: first as a derivation, then as prose in which each sentence carries its rule in brackets. (These are levels 0 and 2 of *Logic and Proof*, Chapter 5; from here on the book writes level-3 prose; after one more fully tagged proof, in Section 1.2, brackets appear only on the step a proof is about.)

Proposition 1.1.2. Inclusion Is Transitive

For all sets A , B , and C , if $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

Idea. The goal $A \subseteq C$ is a $\forall$ around an $\rightarrow$, so the goal-directed moves open two boxes: a box for an arbitrary x_0 , and inside it a box on $x_0 \in A$. The new goal is $x_0 \in C$. Each hypothesis is a $\forall$ around an $\rightarrow$ too, and we use each at x_0 : the first carries x_0 from A into B , the second from B into C .

Proof. Unwound by [Definition 1.1.2](#), the hypotheses are $\forall x (x \in A \rightarrow x \in B)$ and $\forall x (x \in B \rightarrow x \in C)$, and the goal is $\forall x (x \in A \rightarrow x \in C)$. The following is a derivation of the goal from the hypotheses. The outer box, flagged x_0 , is the box of $\forall I$ (*Logic and Proof*, Chapter 4); it has no assumption line.

1		$\forall x (x \in A \rightarrow x \in B)$	
2		$\forall x (x \in B \rightarrow x \in C)$	
3		x_0	$x_0 \in A$
4			$x_0 \in A \rightarrow x_0 \in B$ $\forall E$ 1
5			$x_0 \in B$ $\rightarrow E$ 4, 3
6			$x_0 \in B \rightarrow x_0 \in C$ $\forall E$ 2
7			$x_0 \in C$ $\rightarrow E$ 6, 5
8			$x_0 \in A \rightarrow x_0 \in C$ $\rightarrow I$ 3–7
9			$\forall x (x \in A \rightarrow x \in C)$ $\forall I$ 3–8

The same proof in prose, with each step's rule in brackets:

Suppose $A \subseteq B$ and $B \subseteq C$. Let x be arbitrary [$\forall I$, opening], and suppose $x \in A$ [$\rightarrow I$, opening]. Since $A \subseteq B$, in particular $x \in A \rightarrow x \in B$ [$\forall E$], so $x \in B$ [$\rightarrow E$]. Since $B \subseteq C$, in particular $x \in B \rightarrow x \in C$ [$\forall E$], so $x \in C$ [$\rightarrow E$]. Hence $x \in A \rightarrow x \in C$ [$\rightarrow I$, closing], and since x was arbitrary, $A \subseteq C$ [$\forall I$, closing].

This proves the proposition. ■

At level 3, the prose a mathematician writes, the same proof is one line: *let $x \in A$; then $x \in B$ since $A \subseteq B$, and $x \in C$ since $B \subseteq C$* . The silent steps are those *Logic and Proof*, Chapter 5's convention allows: the two instances of $\forall E$ at the obvious term x , and the two closing sentences. What may never be silent is the opening “let $x \in A$ ”, and the two citations, “since $A \subseteq B$ ” and “since $B \subseteq C$ ”.

Quick check. Let $A = \{1, 2\}$, $B = \{1, 2, 3\}$, and $C = \{1, 3\}$. Which of the six statements $A \subseteq B$, $B \subseteq A$, $A \subseteq C$, $C \subseteq A$, $B \subseteq C$, $C \subseteq B$ are true? For each false one, name a witness.

Solution. True: $A \subseteq B$ and $C \subseteq B$ (every listed element of A , and of C , is listed in B).
 False: $B \subseteq A$ (witness $3 \in B$, $3 \notin A$); $A \subseteq C$ (witness 2); $C \subseteq A$ (witness 3); $B \subseteq C$ (witness 2). Both A and C are proper subsets of B , and neither of A , C is a subset of the other.

1.1.4 Describing sets: listing and specification

Listing works for a handful of elements. For a set such as “the real numbers whose square is less than 2” we need to describe the elements by a property, as the naive rule wanted. The repair of the naive rule is small but essential: we may cut a set out **of a set we already have**.

Definition 1.1.3. Specification

Let A be a set and $P(x)$ a property of objects x . Then there is a set, written

$$\{x \in A : P(x)\},$$

whose elements are exactly the elements of A that have the property P : for every y ,

$$y \in \{x \in A : P(x)\} \iff y \in A \text{ and } P(y).$$

The notation is read “the set of x in A such that $P(x)$ ”; the colon is read “such that” (the book never writes a bar here, so that a bar stays free for divisibility and absolute value). The letter x is a bound variable, like the x of $\forall x: \{x \in A : P(x)\}$ and $\{t \in A : P(t)\}$ are the same set. The set A is the **ambient set** of the description, and every set-builder in this book has one, named or fixed by the context; a set written $\{f(a) : a \in A\}$ is the image of A under f (Section 1.6), and any other set the book defines is shown to exist by (W1)–(W6) where it is introduced. By [Definition 1.1.1](#), the equivalence in the definition determines the set completely, so “the” set is justified.

In the language of the rules: to show that y belongs to $\{x \in A : P(x)\}$, show $y \in A$ and $P(y)$; from $y \in \{x \in A : P(x)\}$, conclude both. Each is one $\leftrightarrow E$ (Logic and Proof, Definition 2.4.1) with the definition, followed or preceded by $\wedge I$ or $\wedge E$ (Logic and Proof, Definition 1.2.5); at level 3 all of these are silent, and a proof simply writes “by definition” or nothing at all.

Specification is one of a short list of **working rules** that say which sets exist. We state them once, here, and use them from now on without comment.

- (W1) **Specification.** For a set A and a property P , $\{x \in A : P(x)\}$ is a set ([Definition 1.1.3](#)).
- (W2) **Listing.** Any finitely many objects $a_1, \dots, a_n$ form a set $\{a_1, \dots, a_n\}$, whose elements are exactly $a_1, \dots, a_n$.
- (W3) **Unions.** For sets A and B , there is a set whose elements are exactly the objects lying in A or in B ; more generally, for a family of sets indexed by a set, there is a set whose elements are exactly the objects lying in at least one member of the family (Section 1.2).
- (W4) **Power sets.** For a set X , the subsets of X form a set (Section 1.2).
- (W5) **Products.** For sets A and B , the ordered pairs (a, b) with $a \in A$ and $b \in B$ form a set; more generally, for a set X and $n \in \mathbb{N}$ with $n \geq 2$, the n -tuples $(x_1, \dots, x_n)$ with every $x_k \in X$ form a set (Section 1.3).
- (W6) **Numbers.** $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ are sets.

These rules are enough for everything the book does with sets until Chapter 3, which adds

one more: a principle for making infinitely many choices at once (Section 1.5 shows where it is first needed). They are not axioms in the formal sense, and we do not try to derive them from fewer principles; *Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* , Chapter 1 states the axioms of set theory behind them and shows how (W1)–(W5) and the set $\mathbb{N}$ follow from them, (W3) only for families of subsets of one given set (a family with no such ambient set needs a further axiom, which *Axiomatic Set Theory* adds), and *Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* , Chapters 2 and 5 build $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$, which completes (W6). What matters here is what is **missing** from the list: there is no rule that collects all objects with a property, without an ambient set. The end of this section shows that such a rule would be contradictory.

1.1.5 The empty set

Among all sets, one is special: the set with nothing in it. It turns up constantly: as the set of real solutions of $x^2 = -1$, as the intersection of two sets with nothing in common (Section 1.2), as the domain of a function with no inputs (Section 1.4). Without it, “the set of solutions” would not always name a set.

Definition 1.1.4. Empty Set

An **empty set** is a set with no elements: a set E such that $\forall x (x \notin E)$.

Such a set exists. Take any set we are given, say $\mathbb{N}$, and collect its elements x with $x \neq x$; by specification (Definition 1.1.3), this is a set, and since $x = x$ for every x (=I, Logic and Proof, Definition 5.7.1), nothing has the property, so the set has no elements. There is also at most one empty set, which justifies the article “**the**”.

Proposition 1.1.3. The Empty Set Is Unique

If E and E' are empty sets, then $E = E'$.

Proof. Let x be arbitrary. Since E and E' are empty, $x \notin E$ and $x \notin E'$. By Logic and Proof, Proposition 2.1.2, $x \notin E$ gives $x \in E \rightarrow x \in E'$, and $x \notin E'$ gives $x \in E' \rightarrow x \in E$. So $x \in E \leftrightarrow x \in E'$ by Logic and Proof, Proposition 2.4.1. Since x was arbitrary, $E = E'$ by Definition 1.1.1. ■

We write $\emptyset$ for **the** empty set. The same vacuous truth gives its most useful property at once: $\emptyset \subseteq A$ **for every set** A . Indeed, let x be arbitrary; $x \notin \emptyset$, so $x \in \emptyset \rightarrow x \in A$ by Logic and Proof, Proposition 2.1.2; since x was arbitrary, $\emptyset \subseteq A$. In words: to refute $\emptyset \subseteq A$ we would need an element of $\emptyset$ outside A , and $\emptyset$ has no elements at all.

The empty set is a set with no elements, and a set **containing** it is not empty. The set $\{\emptyset\}$ has exactly one element, namely $\emptyset$; so $\{\emptyset\} \neq \emptyset$, with $\emptyset$ itself as the witness: $\emptyset \in \{\emptyset\}$ and $\emptyset \notin \emptyset$. A box containing an empty box is not an empty box.

Quick check.

- (a) Is $\emptyset \subseteq \emptyset$?

- (b) Is $\emptyset \in \emptyset$?
 (c) Which of the sets $\emptyset$, $\{\emptyset\}$, $\{\{\emptyset\}\}$ are equal?

Solution.

- (a) Yes: $\emptyset \subseteq A$ for every set A , and in particular for $A = \emptyset$. (It also follows from $A \subseteq A$.)
 (b) No: $\emptyset$ has no elements, so nothing, not even $\emptyset$, belongs to it.
 (c) None. $\emptyset$ has no elements, and each of the other two has exactly one, so $\emptyset$ differs from both. The only element of $\{\emptyset\}$ is $\emptyset$, and the only element of $\{\{\emptyset\}\}$ is $\{\emptyset\}$; these two elements are different by the first observation, so $\emptyset \in \{\emptyset\}$ but $\emptyset \notin \{\{\emptyset\}\}$, and the two sets differ by [Definition 1.1.1](#).

1.1.6 Computing with set-builders

The next worked example shows how a set-builder is computed: we guess the answer, then prove the guess by two inclusions ([Proposition 1.1.1](#)).

Example. Computing sets given by a property

- (a) Show that $\{x \in \mathbb{Z} : x^2 < 5\} = \{-2, -1, 0, 1, 2\}$.
 (b) Show that $\{x \in \mathbb{R} : x^2 < 0\} = \emptyset$.

Solution.

- (a) Write S for the left side. Squaring the integers near 0 suggests the guess: $0^2 = 0$, $(\pm 1)^2 = 1$, $(\pm 2)^2 = 4$ are below 5, and $(\pm 3)^2 = 9$ is not.
 ($\subseteq$) Let $x \in S$. Then $x \in \mathbb{Z}$ and $x^2 < 5$, by [Definition 1.1.3](#). We claim $|x| \leq 2$. Otherwise $|x| \geq 3$, since x is an integer, and then $x^2 = |x|^2 \geq 9 > 5$, which contradicts $x^2 < 5$. So $|x| \leq 2$, and the integers with $|x| \leq 2$ are $-2, -1, 0, 1, 2$. Hence $x \in \{-2, -1, 0, 1, 2\}$.
 ($\supseteq$) Let $x \in \{-2, -1, 0, 1, 2\}$. Then x is one of five integers, and we check each case: x^2 is 4, 1, 0, 1, 4 respectively, and each is less than 5. So $x \in \mathbb{Z}$ and $x^2 < 5$, that is, $x \in S$. By [Proposition 1.1.1](#), $S = \{-2, -1, 0, 1, 2\}$.
 (b) ($\subseteq$) Let $x \in \{x \in \mathbb{R} : x^2 < 0\}$. Then x is a real number with $x^2 < 0$; but $x^2 \geq 0$ for every real x , a contradiction. So this case cannot occur, and $x \in \emptyset$ holds in it vacuously ($\perp$ E, Logic and Proof, [Definition 2.1.1](#)).

($\supseteq$) $\emptyset \subseteq \{x \in \mathbb{R} : x^2 < 0\}$, because $\emptyset$ is a subset of every set.

By [Proposition 1.1.1](#), the set is $\emptyset$. A set-builder with an impossible property is a legitimate description of $\emptyset$.

Two remarks on the example. In (a), the ($\supseteq$) direction is a proof by cases over a finite list, which is legitimate: the five cases exhaust the set, because the list says exactly which elements it has. That is quite different from the error of the second warning above, where finitely many cases were checked for an infinite set. In (b), the ($\supseteq$) direction is free, and only ($\subseteq$) needs an

argument. For “ $S = \emptyset$ ” this is always so: the claim amounts to “ S has no elements”.

1.1.7 Russell's paradox as a theorem

We return to the contradiction of the opening. The naive rule failed on $R =$ “all x with $x \notin x$ ”. Specification allows only the version with an ambient set: for any set A ,

$$R_A := \{x \in A : x \notin x\}$$

is a set, by (W1). What happens to the paradox now?

Theorem 1.1.4. Russell's Theorem

For every set A , the set $R_A = \{x \in A : x \notin x\}$ is **not** an element of A . Hence there is no set of which every set is an element.

Try it first (five minutes): suppose $R_A \in A$. Is $R_A \in R_A$? Try each answer in turn, and see where each one leads.

Idea. The attempt runs into the old paradox, but with one extra condition. If $R_A \in R_A$, then by specification $R_A \notin R_A$: the first answer refutes itself. If $R_A \notin R_A$, then specification puts R_A into R_A , **provided** $R_A \in A$: the second answer refutes itself only under that proviso. So the two answers together are impossible exactly when $R_A \in A$, and the conclusion is not a contradiction in mathematics but a fact: $R_A \notin A$.

We do not need to split into the two answers. The first answer is refuted outright, with no proviso, by $\neg I$ (Logic and Proof, Definition 2.2.1); so $R_A \notin R_A$ holds whenever $R_A \in A$. That fact, fed back into specification, gives $R_A \in R_A$, and the two collide.

Proof. Let A be a set, and write R for R_A . By Definition 1.1.3, for every y ,

$$y \in R \iff y \in A \text{ and } y \notin y. \quad (*)$$

Suppose $R \in A$, for $\neg I$. We first show $R \notin R$. Suppose $R \in R$. Then by $(*)$ at $y = R$, $R \in A$ and $R \notin R$; in particular $R \notin R$, which contradicts $R \in R$. Hence $R \notin R$, by $\neg I$. Now $R \in A$ and $R \notin R$, so $R \in R$ by $(*)$ read from right to left. This contradicts $R \notin R$. Hence $R \notin A$, by $\neg I$.

For the second claim, suppose U were a set of which every set is an element. By the first claim with $A = U$, the set R_U is not an element of U ; but R_U is a set, so it is an element of U . This contradiction shows that no such U exists. ■

Here is the first part of the proof as a derivation, with $(*)$ as its premise, written as the universal statement it is, and R for R_A :

1	$\forall y(y \in R \leftrightarrow (y \in A \wedge y \notin y))$	
2	$R \in A$	
3	$R \in R \leftrightarrow (R \in A \wedge R \notin R)$	$\forall E$ 1
4	$R \in R$	
5	$R \in A \wedge R \notin R$	$\leftrightarrow E$ 3, 4
6	$R \notin R$	$\wedge E$ 5
7	$\perp$	$\neg E$ 4, 6
8	$R \notin R$	$\neg I$ 4–7
9	$R \in A \wedge R \notin R$	$\wedge I$ 2, 8
10	$R \in R$	$\leftrightarrow E$ 3, 9
11	$\perp$	$\neg E$ 10, 8
12	$R \notin A$	$\neg I$ 2–11

Every rule in it is one of $\forall E$, $\leftrightarrow E$, $\wedge E$, $\wedge I$, $\neg E$, and $\neg I$. There is **no** RAA, and no excluded middle: the inner box (lines 4–7) refutes one answer outright instead of splitting into two answers. This is worth noticing. The paradox is not an artifact of classical logic, and cannot be escaped by giving up a classical rule; it is a genuine contradiction in the naive rule, and the only repair is to change which sets exist.

What the theorem says about sets. Read with A arbitrary, the theorem says that **no set contains everything**: for every set A we have produced a set, R_A , outside it. The naive rule “any property defines a set” is exactly specification **without** the phrase “ $x \in A$ ”, and the theorem shows that the phrase cannot be dropped: with A ranging over “everything”, R_A would be the paradoxical R . This is why every set-builder in this book has an ambient set, and why (W1)–(W6) contain no “set of all sets”. When we want “all objects with a property”, we first say where to look.

The argument has a shape that will return: a set built from A so that it disagrees with each candidate on the candidate itself. Chapter 3 proves by the same shape of argument that no function from a set X to the set of its subsets reaches every subset.

1.1.8 Exercises

1.1.8.1 A. Check your understanding

Exercise. A1

- (a) Negate “ $A \subseteq B$ ” in symbols, and then in words, with the negation pushed through the quantifier and the implication.
- (b) True or false: $\{1\} \in \{\{1\}, 2\}$; $\{1\} \subseteq \{\{1\}, 2\}$; $\emptyset \subseteq \emptyset$; $\{\emptyset\} \subseteq \{1\}$. Give a reason for each.
- (c) List the elements of $\{n \in \mathbb{Z} : 0 < n^2 \leq 9\}$.
- (d) In an element-chasing proof of $A \subseteq B$, the first sentence is “let $x \in A$ ”. Which rule of logic gives its first clause, and which its second?

Solution.

- (a) By [Definition 1.1.2](#), $A \subseteq B$ is $\forall x (x \in A \rightarrow x \in B)$. By the quantifier negation law of *Logic and Proof*, Chapter 4, its negation is equivalent to $\exists x \neg(x \in A \rightarrow x \in B)$, and by the law for negated implications (also *Logic and Proof*, Chapter 4) to

$$\exists x (x \in A \wedge x \notin B).$$

In words: some element of A is not an element of B .

- (b) True: $\{1\}$ is listed as an element of $\{\{1\}, 2\}$. False: the only element of $\{1\}$ is 1, and $1 \notin \{\{1\}, 2\}$, since $1 \neq \{1\}$ (a number is not a set) and $1 \neq 2$; so 1 is a witness. True: $\emptyset$ is a subset of every set. False: $\emptyset \in \{\emptyset\}$ but $\emptyset \notin \{1\}$, since $\emptyset \neq 1$; so $\emptyset$ is a witness.
- (c) $\{-3, -2, -1, 1, 2, 3\}$. The integer 0 is excluded because $0^2 = 0$ is not > 0 , and integers with $|n| \geq 4$ have $n^2 \geq 16 > 9$.
- (d) “Let $x \in A$ ” opens two boxes. Its first clause, “let x be arbitrary”, is $\forall I$ (*Logic and Proof*, Chapter 4); its second, “suppose $x \in A$ ”, is $\rightarrow I$ (*Logic and Proof*, Definition 1.4.1). The proof must then reach $x \in B$ inside both boxes.

1.1.8.2 B. Practice

Exercise. B1: A set given by an equation

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

Solution. Write S for the left side.

($\subseteq$) Let $x \in S$. Then x is real and $x^2 - 3x + 2 = 0$. Since $x^2 - 3x + 2 = (x - 1)(x - 2)$, we have $(x - 1)(x - 2) = 0$. A product of two real numbers is zero only if one of the factors is zero, so $x - 1 = 0$ or $x - 2 = 0$, that is, $x = 1$ or $x = 2$. In either case $x \in \{1, 2\}$.

($\supseteq$) Let $x \in \{1, 2\}$, so $x = 1$ or $x = 2$. If $x = 1$, then $x^2 - 3x + 2 = 1 - 3 + 2 = 0$. If $x = 2$, then $x^2 - 3x + 2 = 4 - 6 + 2 = 0$. In either case x is a real number satisfying the equation, so $x \in S$.

By [Proposition 1.1.1](#), $S = \{1, 2\}$.

Exercise. B2: A subset of the empty set

Let A be a set with $A \subseteq \emptyset$. Prove that $A = \emptyset$.

Solution. We have $A \subseteq \emptyset$ by hypothesis, and $\emptyset \subseteq A$ because the empty set is a subset of every set. By [Proposition 1.1.1](#), $A = \emptyset$.

Directly: if $x \in A$, then $x \in \emptyset$ because $A \subseteq \emptyset$, which is impossible; so A has no elements, and it is **the** empty set by [Proposition 1.1.3](#).

Exercise. B3: Equal or not?

For each pair, determine whether the two sets are equal. Justify your answer: prove each equality, and give a witness element for each inequality.

- (a) $\{x \in \mathbb{Z} : 4 \mid x\}$ and $\{x \in \mathbb{Z} : 2 \mid x\}$.
- (b) $\{x \in \mathbb{R} : x^2 = 4\}$ and $\{-2, 2\}$.
- (c) $\{\emptyset, \{\emptyset\}\}$ and $\{\{\emptyset\}, \emptyset, \emptyset\}$.
- (d) $\{x \in \mathbb{Z} : x^2 < 2\}$ and $\{x \in \mathbb{N} : x < 2\}$.

Solution.

- (a) Not equal. The integer 2 lies in the second set ($2 = 2 \cdot 1$) but not in the first, since $2 = 4k$ would need $k = \frac{1}{2}$, which is not an integer. (The first set is a proper subset of the second: $x = 4k$ gives $x = 2 \cdot (2k)$.)
- (b) Equal. ($\subseteq$) If x is real with $x^2 = 4$, then $(x - 2)(x + 2) = x^2 - 4 = 0$, so $x = 2$ or $x = -2$. ($\supseteq$) $2^2 = 4$ and $(-2)^2 = 4$.
- (c) Equal. Both sets have exactly the elements $\emptyset$ and $\{\emptyset\}$: an object lies in the first exactly when it is $\emptyset$ or $\{\emptyset\}$, and in the second exactly when it is $\{\emptyset\}$ or $\emptyset$ or $\emptyset$. Order and repetition do not count.
- (d) Not equal. The first set is $\{-1, 0, 1\}$: an integer with $|x| \geq 2$ has $x^2 \geq 4$, and $(-1)^2 = 1$, $0^2 = 0$, $1^2 = 1$ are below 2. The second set is $\{0, 1\}$, since $\mathbb{N}$ starts at 0. The witness -1 lies in the first set and not in the second ($-1 \notin \mathbb{N}$).

1.1.8.3 C. Going deeper**Exercise. C1: Every set has something outside it**

Let A be a set, and let $R_A = \{x \in A : x \notin x\}$ as in [Theorem 1.1.4](#).

- (a) Let A^+ be the set whose elements are the elements of A together with the object R_A (it exists by (W2) and (W3)). Show that $A \subsetneq A^+$.
- (b) Deduce from [Theorem 1.1.4](#) that no set A has the property “every subset of A is an element of A ”.

Solution (Hint 1). In (a), a proper inclusion needs two things: the inclusion, and a witness in the larger set that is not in the smaller one. The theorem supplies the witness. In (b), look for a subset of A that the theorem says is not an element of A .

Solution (Hint 2). For (b): suppose, for $\neg$ I, that every subset of A is an element of A . First show that $R_A \subseteq A$.

Solution.

- (a) First, $A \subseteq A^+$: an element of A is an element of A^+ by the description of A^+ . Second,

$A^+ \neq A$: the object R_A lies in A^+ , and $R_A \notin A$ by [Theorem 1.1.4](#). So R_A is a witness, and $A \subsetneq A^+$.

- (b) Suppose, for $\neg I$, that A is a set such that every subset of A is an element of A . The set R_A is a subset of A : if $y \in R_A$, then $y \in A$ (and $y \notin y$) by [Definition 1.1.3](#). So by the assumption, $R_A \in A$, which contradicts [Theorem 1.1.4](#). Hence no set has the property.

Part (a) says that no set is “as large as possible”: there is always a set with one more element. Part (b) rules out a set containing all of its own subsets; for instance, no set A has as its elements exactly the subsets of A .

Marking scheme. (10 points) - 2: (a) proves the inclusion $A \subseteq A^+$ from the description of A^+ . - 3: (a) names R_A as the witness to $A^+ \neq A$, citing [Theorem 1.1.4](#) for $R_A \notin A$. - 3: (b) proves $R_A \subseteq A$ by element-chasing through [Definition 1.1.3](#). - 2: (b) combines the assumption with $R_A \subseteq A$ to get $R_A \in A$, and cites the theorem for the contradiction.

Fatal errors (0 for the step): using “the set of all sets” or any set-builder without an ambient set; taking $A^+ \neq A$ as obvious “because an element was added” (if R_A were already in A , nothing would have been added).

Exercise. C2: Antisymmetry for sets of sets

Can there be sets A and B , whose elements are themselves sets, such that $A \subseteq B$, $B \subseteq A$, and $A \neq B$? Explain why the answer does not depend on what the elements of A and B are.

Solution. No. By [Proposition 1.1.1](#), $A \subseteq B$ and $B \subseteq A$ give $A = B$, for **all** sets A and B .

The answer does not depend on the elements, because nothing in the proof of [Proposition 1.1.1](#) looks at them. The proof works with the statements $x \in A$ and $x \in B$ for an arbitrary object x , and uses only [Definition 1.1.2](#), [Definition 1.1.1](#), and the propositional law Logic and Proof, Proposition 2.4.1 applied to those two statements. Whether x is a number, a set, or a set of sets never enters. For example, with $A = \{\emptyset, \{\emptyset\}\}$ and $B = \{\{\emptyset\}, \emptyset\}$, the two inclusions hold and indeed $A = B$.

The point of the question is a habit: an element-chasing proof treats its element as a black box, and so it proves the statement for sets of anything at all.

1.2 Unions, Intersections, Complements and Families

Section 1.1 said what sets are, when two of them are equal, and when one lies inside another. It gave us ways to describe a single set, by listing or by a property cut out of an ambient set. This section makes new sets from old ones. Each connective of *Logic and Proof*, Chapter 1 gives a set operation: “or” gives the union, “and” gives the intersection, “and not” gives the difference. That observation is the thesis of the chapter: *the laws of sets are the laws of logic, read through membership*. We state it, prove the standard laws of set algebra from it, and then take the step analysis needs constantly, from two sets to a whole family of them, where “or” becomes “there exists” and “and” becomes “for all”.

1.2.1 Union, intersection and difference

Here is the standard example of a new set. Let A be the set of integers divisible by 2 and B the set of integers divisible by 3. Three natural questions about an integer x are: is it divisible by 2 **or** by 3? by 2 **and** by 3? by 2 **and not** by 3? Each question collects integers into a set, and each set is built from A and B by one connective applied to the statements $x \in A$ and $x \in B$. The pattern does not depend on divisibility at all, so we name the three constructions once, for any sets.

The union collects what lies in either set, the intersection what lies in both, the difference what lies in the first and not the second.

Definition 1.2.1. Union, Intersection, Difference

Let A and B be sets.

- (a) The **union** $A \cup B$ is the set whose elements are exactly the objects lying in A **or** in B :
 $x \in A \cup B \iff (x \in A \vee x \in B)$.
- (b) The **intersection** $A \cap B$ is the set $\{x \in A : x \in B\}$: so $x \in A \cap B \iff (x \in A \wedge x \in B)$.
- (c) The **difference** $A \setminus B$ is the set $\{x \in A : x \notin B\}$: so $x \in A \setminus B \iff (x \in A \wedge x \notin B)$.

Sets A and B are **disjoint** if $A \cap B = \emptyset$. If $A \subseteq X$, the difference $X \setminus A$ is called the **complement of A in X** .

In words, clause by clause. The union exists by the working rule (W3) of Section 1.1; it is not cut out of an ambient set, since its elements come from two sets at once. The intersection and the difference **are** cut out of an ambient set, namely A , so they exist by specification (W1), and the description $\{x \in A : \dots\}$ names its ambient set as Section 1.1 requires. The “or” of the union is inclusive, as $\vee$ always is: an object in both sets is in the union. “Disjoint” means “no common element”.

The complement is always written $X \setminus A$, with the ambient set X named. Many books write A^c and leave X to the reader, and a reader who guesses X wrongly gets a different set: the complement of $[0, 1]$ in $[0, 2]$ is $(1, 2]$, and in $\mathbb{R}$ it is $(-\infty, 0) \cup (1, \infty)$. There is no “complement of A ” on its own: by Russell’s theorem ([Theorem 1.1.4](#)) there is no set of everything outside A (such a set, joined with A , would contain every set). So this book never writes A^c .

Examples, simplest first, each checked against the membership conditions.

- **Finite sets.** For $A = \{1, 2, 3\}$ and $B = \{2, 4\}$: $A \cup B = \{1, 2, 3, 4\}$, $A \cap B = \{2\}$, $A \setminus B = \{1, 3\}$, $B \setminus A = \{4\}$. For instance $3 \in A \setminus B$ because $3 \in A$ and $3 \notin B$; and $4 \notin A \setminus B$ because $4 \notin A$.
- **Intervals.** For $A = [0, 2]$ and $B = [1, 3]$: $A \cup B = [0, 3]$, $A \cap B = [1, 2]$, $A \setminus B = [0, 1)$. The last one shows how an endpoint behaves: $1 \in B$, so $1 \notin A \setminus B$, while every x with $0 \leq x < 1$ lies in A and not in B .
- **Divisibility.** With A and B the multiples of 2 and of 3 in $\mathbb{Z}$, as in the opening, $A \cap B$ is the set of multiples of 6 (this uses the school fact that an integer divisible by 2 and by 3 is divisible by 6), and $\mathbb{Z} \setminus A$ is the set of odd integers, the complement of A in $\mathbb{Z}$.
- **Degenerate cases.** For every set A : $A \cup \emptyset = A$, $A \cap \emptyset = \emptyset$, $A \setminus \emptyset = A$, and $A \setminus A = \emptyset$. Each is one line: for example, $x \in A \setminus A$ would mean $x \in A$ and $x \notin A$, which is impossible, so $A \setminus A$ has no element. They matter because $\emptyset$ behaves for $\cup$ as 0 does for $+$, and for $\cap$ as 0 does for $\times$.

Three inclusions hold for all sets and are used without comment from now on:

$$A \cap B \subseteq A \subseteq A \cup B, \quad A \setminus B \subseteq A.$$

Each is one element-chase that drops or adds a disjunct or a conjunct: if $x \in A \cap B$, then $x \in A$ ($\wedge E$); if $x \in A$, then $x \in A \cup B$ ($\vee I$); if $x \in A \setminus B$, then $x \in A$ ($\wedge E$).

A nonexample by minimal change: the “difference” of two sets is **not** “everything in one but not in both”. The set $\{x \in A \cup B : \text{not both } x \in A \text{ and } x \in B\}$ is a different set: for $A = \{1, 2, 3\}$ and $B = \{2, 4\}$ it is $\{1, 3, 4\}$, while $A \setminus B = \{1, 3\}$. The element 4 is in the first and not in $A \setminus B$, because the difference keeps only elements **of** A .

Warning. The difference is not symmetric. In general $A \setminus B \neq B \setminus A$. With $A = \{1\}$ and $B = \{2\}$, $A \setminus B = \{1\}$ and $B \setminus A = \{2\}$, and the witness 1 lies in the first and not the second. In fact $A \setminus B$ and $B \setminus A$ are always disjoint, since an element of both would lie in A and not in A ; they are equal only when both are empty, that is, when $A = B$.

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

Solution. $A \cup B = \{1, 2, 3, 4, 6\}$ and $A \cap B = \{2\}$. Then $X \setminus (A \cup B) = \{5\}$. Also $X \setminus A = \{4, 5, 6\}$ and $X \setminus B = \{1, 3, 5\}$, whose intersection is $\{5\}$. The last two answers agree. This is the two-set case of De Morgan’s law for sets, which is proved for arbitrary families later in this section.

1.2.2 Laws of sets are laws of logic

Unwinding the three definitions turns every question about $\cup$, $\cap$, and $\setminus$ into a question about $\vee$, $\wedge$, and $\neg$. Here is the dictionary, with the notions of Section 1.1 included.

On sets	Unwound through membership	Logic
$x \in A \cup B$	$x \in A \vee x \in B$	$\vee$
$x \in A \cap B$	$x \in A \wedge x \in B$	$\wedge$
$x \in A \setminus B$	$x \in A \wedge \neg(x \in B)$	$\wedge$ and $\neg$
$A = \emptyset$	$\forall x \neg(x \in A)$	$\neg$ under $\forall$
$A \subseteq B$	$\forall x (x \in A \rightarrow x \in B)$	$\rightarrow$ under $\forall$
$A = B$	$\forall x (x \in A \leftrightarrow x \in B)$	$\leftrightarrow$ under $\forall$

The dictionary makes an idea precise. Fix an object x and write p, q, r for the statements $x \in A$, $x \in B$, $x \in C$. Then membership of x in any set built from A, B, C by $\cup, \cap, \setminus$ is a propositional formula in p, q, r : for instance,

$$x \in A \cap (B \cup C) \iff p \wedge (q \vee r).$$

Two such sets are equal if their membership formulas can each be derived from the other, for then x lies in one exactly when it lies in the other, and x was arbitrary. So a law of propositional

logic, with letters replaced by membership statements, is a law of sets. That is the thesis of this chapter.

The laws of sets are the laws of logic, read through membership.

Here are the laws used most.

Theorem 1.2.1. Algebra of Sets

For all sets A , B , and C :

- (a) (commutativity) $A \cup B = B \cup A$ and $A \cap B = B \cap A$;
- (b) (associativity) $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$;
- (c) (distributivity) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
- (d) (differences) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$ and $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$.

Idea. Each law is a propositional equivalence under $\forall x$. *Logic and Proof*, Chapter 3 gives a uniform license for all of them at once: an equivalence that holds in every row of a truth table is derivable in both directions, by completeness, and a derived sequent may be used with its letters replaced by any statements. So one argument covers every law, and the table after the proof names, for each law, the equivalence it comes from. We also write out two laws in full, as element-chases, so that the reader sees what the uniform argument does: the first distributive law, and the second difference law, the one law whose element-chase needs a classical rule (the uniform argument uses one for every law, inside the proof of *Logic and Proof*, Theorem 4.8.4).

Proof. The uniform argument. Fix one of the eight equations, $L = R$, and let φ and ψ be the propositional formulas in the atoms p, q, r that describe membership in L and in R , as in the dictionary: for the first distributive law, $\varphi = p \wedge (q \vee r)$ and $\psi = (p \wedge q) \vee (p \wedge r)$. For (a)–(c), the semantic equivalence $\varphi \equiv \psi$ is a clause of *Logic and Proof*, Proposition 4.5.1, with p, q, r in the places of φ, ψ, χ : (f) for (a), (h) for (b), (g) for (c). For (d), it is checked by the truth table displayed below. By *Logic and Proof*, Definition 4.5.4, $\varphi \models \psi$ and $\psi \models \varphi$, so by *Logic and Proof*, Theorem 4.8.4, $\varphi \vdash \psi$ and $\psi \vdash \varphi$. Now let x be arbitrary. By *Logic and Proof*, Definition 2.3.1, each of these two sequents may be used with p, q, r replaced by the statements $x \in A$, $x \in B$, $x \in C$; with [Definition 1.2.1](#), the first gives $x \in L \rightarrow x \in R$ and the second $x \in R \rightarrow x \in L$. Hence $x \in L \leftrightarrow x \in R$, and since x was arbitrary, $L = R$ by [Definition 1.1.1](#).

The truth table for (d), with the atoms in the book's order (T before F, the last atom

alternating fastest):

p	q	r	$p \wedge \neg(q \vee r)$	$(p \wedge \neg q) \wedge (p \wedge \neg r)$	$p \wedge \neg(q \wedge r)$	$(p \wedge \neg q) \vee (p \wedge \neg r)$
T	T	T	F	F	F	F
T	T	F	F	F	T	T
T	F	T	F	F	T	T
T	F	F	T	T	T	T
F	T	T	F	F	F	F
F	T	F	F	F	F	F
F	F	T	F	F	F	F
F	F	F	F	F	F	F

The fourth and fifth columns agree in every row, and so do the sixth and seventh. This proves all eight laws.

The first distributive law, in full. ($\subseteq$) Let $x \in A \cap (B \cup C)$. Then $x \in A$ and $x \in B \cup C$, that is, $x \in A$ and ($x \in B$ or $x \in C$). By Logic and Proof, Proposition 1.6.1, with P, Q, R replaced by $x \in A, x \in B, x \in C$, we get $(x \in A \text{ and } x \in B)$ or $(x \in A \text{ and } x \in C)$; that is, $x \in A \cap B$ or $x \in A \cap C$, so $x \in (A \cap B) \cup (A \cap C)$. ($\supseteq$) The same chain read backwards, using the second sequent of Logic and Proof, Proposition 1.6.1. By Proposition 1.1.1 the two sets are equal.

The second difference law, in full. ($\subseteq$) Let $x \in A \setminus (B \cap C)$. Then $x \in A$ and $\neg(x \in B \wedge x \in C)$. By Logic and Proof, Proposition 2.5.5 (d), $x \notin B$ or $x \notin C$ [RAA, inside (d)]. So $x \in A$ and ($x \notin B$ or $x \notin C$), and by Logic and Proof, Proposition 1.6.1, $(x \in A \text{ and } x \notin B)$ or $(x \in A \text{ and } x \notin C)$; that is, $x \in (A \setminus B) \cup (A \setminus C)$. ($\supseteq$) Let $x \in (A \setminus B) \cup (A \setminus C)$. By the second sequent of Logic and Proof, Proposition 1.6.1, $x \in A$ and ($x \notin B$ or $x \notin C$). By Logic and Proof, Proposition 2.5.5 (c), $\neg(x \in B \wedge x \in C)$, so $x \notin B \cap C$, and $x \in A \setminus (B \cap C)$. By Proposition 1.1.1 the two sets are equal. ■

Each law is a propositional equivalence, and the table records where each comes from: the clause of Logic and Proof, Proposition 4.5.1, and the derivations of *Logic and Proof*, Chapters 1 and 2, where they exist, that prove one or both directions directly.

Law	Clause of Logic and Proof, Proposition 4.5.1	Derived in <i>Logic and Proof</i> , Chapters 1–2
$A \cap B = B \cap A$	(f)	Logic and Proof, Proposition 1.2.1 (either direction, letters exchanged)
$A \cup B = B \cup A$	(f)	Logic and Proof, Example (First derivations with “or”) (likewise)
$(A \cap B) \cap C = A \cap (B \cap C)$	(h)	Logic and Proof, Proposition 1.2.2 (left to right)
$(A \cup B) \cup C = A \cup (B \cup C)$	(h)	Logic and Proof, Example (Regrouping a disjunction) (right to left)

Law	Clause of Logic and Proof, Proposition 4.5.1	Derived in <i>Logic and Proof</i> , Chapters 1–2
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ (g)		Logic and Proof, Proposition 1.6.1 (both directions)
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (g)		none in the text (both directions are exercises in <i>Logic and Proof</i> , Chapters 1 and 2)
$A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$ (b), then regrouping of $\wedge$		Logic and Proof, Proposition 2.5.5 (a), (b)
$A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$ (b), then (g)		Logic and Proof, Proposition 2.5.5 (c), (d); Logic and Proof, Proposition 1.6.1

The two difference laws look alike, but their element-chases are not alike. The first needs De Morgan’s laws (a) and (b), which *Logic and Proof*, Section 2.5 derives without RAA, followed by regrouping the conjunction $x \in A \wedge (x \notin B \wedge x \notin C)$ into $(x \in A \wedge x \notin B) \wedge (x \in A \wedge x \notin C)$, which takes only $\wedge E$ and $\wedge I$. The second needs De Morgan (d) in its $(\subseteq)$ direction, and (d) is the one of the four De Morgan laws that *Logic and Proof*, Section 2.5 derives with RAA; its $(\supseteq)$ direction uses (c), derived without it. Knowing that x is not in both B and C does not by itself tell us which of the two it misses; De Morgan (d) supplies the answer “one of them”, and it is the classical rule that supplies it.

The laws let us calculate with sets as with numbers, but more symmetrically: each of $\cup$ and $\cap$ distributes over the other, whereas $\times$ distributes over $+$ and not conversely. They also let us drop brackets: by associativity, $A \cup B \cup C$ and $A \cap B \cap C$ are unambiguous.

Quick check. Which propositional equivalence proves $A \cap (A \cup B) = A$? Is it one of the clauses of Logic and Proof, Proposition 4.5.1?

Solution. Membership in the left side is $p \wedge (p \vee q)$, and in the right side p . The equivalence $p \wedge (p \vee q) \equiv p$ holds in all four rows: if p is true, both sides are true; if p is false, both are false. It is not one of the clauses of Logic and Proof, Proposition 4.5.1, so the uniform argument applies with a two-atom truth table instead of a clause. (Directly: $(\subseteq)$ is $\wedge E$; $(\supseteq)$ is $\vee I$ followed by $\wedge I$.) This is one of the absorption laws.

1.2.3 Power sets

The operations so far build sets whose elements are elements of the old sets. The next construction collects **sets** as elements. Many questions of analysis quantify over subsets: “every non-empty set of reals that is bounded above has a supremum” is a statement about all subsets of $\mathbb{R}$. So the subsets of a set deserve to be collected into one set.

Definition 1.2.2. Power Set

Let X be a set. The **power set** of X is the set $\mathcal{P}(X)$ whose elements are exactly the subsets of X :

$$S \in \mathcal{P}(X) \iff S \subseteq X.$$

The set exists by the working rule (W4) of Section 1.1. Its elements are sets, and “ $S \in \mathcal{P}(X)$ ” is just another way of writing “ $S \subseteq X$ ”. Two elements are always present: $\emptyset \in \mathcal{P}(X)$, because the empty set is a subset of every set, and $X \in \mathcal{P}(X)$, because $X \subseteq X$.

Example. Small power sets

Compute $\mathcal{P}(\emptyset)$, $\mathcal{P}(\{\emptyset\})$, and $\mathcal{P}(\{1, 2\})$.

Solution. $\mathcal{P}(\emptyset) = \{\emptyset\}$. The only subset of $\emptyset$ is $\emptyset$ itself: if $S \subseteq \emptyset$, then an element of S would be an element of $\emptyset$, so S has no elements, and $S = \emptyset$ by [Proposition 1.1.3](#). And $\emptyset \subseteq \emptyset$. So $\mathcal{P}(\emptyset)$ has exactly one element, $\emptyset$. It is **not** empty: the empty set has one subset.

$\mathcal{P}(\{\emptyset\}) = \{\emptyset, \{\emptyset\}\}$. Let $S \subseteq \{\emptyset\}$. Either $\emptyset \in S$ or $\emptyset \notin S$ (excluded middle, Logic and Proof, Proposition 2.5.3). If $\emptyset \in S$, then S and $\{\emptyset\}$ have the same elements (S has no element other than $\emptyset$, since $S \subseteq \{\emptyset\}$), so $S = \{\emptyset\}$. If $\emptyset \notin S$, then S has no elements, so $S = \emptyset$. Conversely, both $\emptyset$ and $\{\emptyset\}$ are subsets of $\{\emptyset\}$.

$\mathcal{P}(\{1, 2\}) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$. Let $S \subseteq \{1, 2\}$. A subset of $\{1, 2\}$ is determined by which of 1 and 2 it contains, by [Definition 1.1.1](#), and for each of the two there are two cases, $1 \in S$ or $1 \notin S$, and $2 \in S$ or $2 \notin S$. The four combinations give $S = \{1, 2\}, \{1\}, \{2\}, \emptyset$. All four are subsets of $\{1, 2\}$. So $\mathcal{P}(\{1, 2\})$ has exactly four elements.

In each case the number of subsets is a power of two: 1, 2, 4, for sets with 0, 1, 2 elements, because each element is independently in or out. That a set with n elements has exactly 2^n subsets is proved in Chapter 3, where counting is made exact. The name “power set” comes from this count.

Warning. **The elements of $\mathcal{P}(X)$ are subsets, not elements, of X .** For $X = \{1, 2\}$: $1 \in X$ but $1 \notin \mathcal{P}(X)$, since 1 is a number, not a subset; whereas $\{1\} \in \mathcal{P}(X)$ and $\{1\} \notin X$. The correct translations are $a \in X \iff \{a\} \in \mathcal{P}(X)$, and $S \subseteq X \iff S \in \mathcal{P}(X)$.

1.2.4 Families of sets

Analysis rarely stops at two or three sets. It intersects all the intervals $(-t, t)$ with $t > 0$, or all the closed sets containing a given set, or all the subsets of $\mathbb{R}$ with some property. We need unions and intersections of many sets at once, and first a way to name “many sets at once”.

Definition 1.2.3. Indexed Family

Let I be a set. A **family of sets indexed by I** , written $(A_i)_{i \in I}$, assigns to **each** $i \in I$ a set A_i . The set I is the **index set**, and A_i is the **member** of the family with index i . If X

is a set with $A_i \subseteq X$ for every $i \in I$, we speak of a **family of subsets of X** .

Here “assigns” is used in its everyday sense. Section 1.4 makes it exact for a family of subsets of a set X : such a family is a function $I \rightarrow \mathcal{P}(X)$, $i \mapsto A_i$, whose value at i is written A_i . The parentheses follow the book’s rule that ordered or indexed things go in parentheses; the family $(A_i)_{i \in I}$ knows which set has which index, and two indices may carry the same set. Examples: for $I = \{1, 2\}$ the family is just a pair of sets A_1, A_2 ; for $I = (0, \infty)$ and $A_t = (-t, t)$ it is a family of intervals, one for each positive real t ; for $I = \mathbb{N}$ it is a sequence of sets $A_0, A_1, A_2, \dots$ (indexed from 0; the sequences of real numbers in Section 1.4 are indexed from 1), which Chapter 2 treats as a sequence defined by recursion and Chapter 3 uses to form countable unions.

“Or” becomes “for some index”, and “and” becomes “for every index”.

The union of a family collects what lies in at least one member; the intersection collects what lies in every member.

Definition 1.2.4. Union and Intersection of a Family

Let $(A_i)_{i \in I}$ be a family of sets.

- (a) The **union** $\bigcup_{i \in I} A_i$ is the set whose elements are exactly the objects lying in A_i for **some** $i \in I$:

$$x \in \bigcup_{i \in I} A_i \iff \exists i \in I (x \in A_i).$$

- (b) If $(A_i)_{i \in I}$ is a family of subsets of a set X , the **intersection** is

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

For a set $\mathcal{S}$ whose elements are sets, $\bigcup \mathcal{S} := \bigcup_{A \in \mathcal{S}} A$; and for a **non-empty** set $\mathcal{S} \subseteq \mathcal{P}(X)$, $\bigcap \mathcal{S} := \bigcap_{A \in \mathcal{S}} A$. In both, the family is indexed by $\mathcal{S}$ itself, the member with index A being A .

In words, clause by clause. The union exists by (W3), like the union of two sets; when all $A_i \subseteq X$, it may also be written $\{x \in X : \exists i \in I (x \in A_i)\}$, since then every element of the union lies in X , and this form names its ambient set as Section 1.1 requires. The intersection is cut out of X by specification. The last sentence covers sets of sets that come with no index: $\bigcap \mathcal{S}$ is the set of objects lying in every set that belongs to $\mathcal{S}$. For example, *Analysis on the Real Line*, Chapter 1 defines $\mathbb{N}$ inside $\mathbb{R}$ as $\bigcap \mathcal{S}$, where $\mathcal{S}$ is the set of all subsets of $\mathbb{R}$ that contain 0 and are closed under $x \mapsto x + 1$; $\mathcal{S}$ is non-empty because $\mathbb{R}$ itself belongs to it.

With two sets the new definitions give the old ones back. For $I = \{1, 2\}$, “ $\exists i \in \{1, 2\}$ with $x \in A_i$ ” says “ $x \in A_1$ or $x \in A_2$ ”, since an $i \in \{1, 2\}$ is 1 or 2; so $\bigcup_{i \in \{1, 2\}} A_i = A_1 \cup A_2$, and likewise $\bigcap_{i \in \{1, 2\}} A_i = A_1 \cap A_2$. The dictionary gains two rows:

On sets	Unwound through membership	Logic
$x \in \bigcup_{i \in I} A_i$	$\exists i \in I (x \in A_i)$	$\exists$
$x \in \bigcap_{i \in I} A_i$, for $x \in X$	$\forall i \in I (x \in A_i)$	$\forall$

Why the intersection is taken inside X . The ambient set is not decoration; it is forced by the case $I = \emptyset$. If I is empty, the condition $\forall i \in I (x \in A_i)$ is **vacuously true** for every object

x : there is no index at which it could fail (*Logic and Proof*, Chapter 4, vacuous truth for $\forall$). A definition without an ambient set, “ $\bigcap_{i \in I} A_i = \text{all } x \text{ with } \forall i \in I (x \in A_i)$ ”, would then describe the set of all objects, which [Theorem 1.1.4](#) forbids. Inside X , the empty intersection is simply X :

$$\bigcap_{i \in \emptyset} A_i = X, \quad \bigcup_{i \in \emptyset} A_i = \emptyset.$$

The union needs no ambient set, because its condition $\exists i \in \emptyset (x \in A_i)$ is false for every x ; the empty union is empty. For a **non-empty** index set the choice of X does not matter: if some $i_0 \in I$ exists and x lies in every A_i , then in particular $x \in A_{i_0} \subseteq X$, so the condition “ $x \in X$ ” adds nothing. Only the empty family sees X , and there it must. This is also why $\bigcap \mathcal{S}$ above asks $\mathcal{S}$ to be non-empty: a set of sets with no index carries no ambient set with it.

1.2.4.1 The smallest set closed under some steps

One use of the intersection pays a debt to *Logic and Proof*, Chapter 3. There, the formulas were described as the smallest collection of strings closed under (F1)–(F3) of *Logic and Proof*, Definition 4.1.1: the collection that contains the atoms and $\perp$ and is closed under the building steps (F2) and (F3). What does “smallest” mean, and why does such a collection exist? Intersections answer both questions, in any setting of this kind.

Let X be a set, and consider some **starting elements** of X and some **steps**, each of which turns elements of X into an element of X . Call a subset $S \subseteq X$ **closed** if it contains the starting elements and, whenever a step is applied to elements of S , the result lies in S .

Proposition 1.2.2. Smallest Closed Set

Let X be a set with given starting elements and steps, and let $\mathcal{S} \subseteq \mathcal{P}(X)$ be the set of closed subsets of X . Then $\mathcal{S}$ is non-empty, and $M := \bigcap \mathcal{S}$ is closed and lies inside every closed set: M is the **smallest** closed subset of X .

Idea. An element of M is an object lying in every closed set. So M lies inside each closed set by its very definition, and each requirement of being closed, met by every closed set at once, is met by what they have in common.

Proof. The set $\mathcal{S}$ is non-empty, since X itself is closed; so $\bigcap \mathcal{S}$ is defined ([Definition 1.2.4](#)).

- **M lies inside every closed set.** If $S \in \mathcal{S}$ and $x \in M$, then x lies in every member of $\mathcal{S}$, in particular in S .
- **M is closed.** Each starting element lies in every $S \in \mathcal{S}$, hence in M . If a step is applied to elements of M , those elements lie in every $S \in \mathcal{S}$; since each such S is closed, the result lies in every $S \in \mathcal{S}$, hence in M .

This proves the proposition. ■

For *Logic and Proof*, Chapter 3, X is the set of all strings of symbols, the starting elements are the atoms and $\perp$, and the steps are (F2) and (F3) of *Logic and Proof*, Definition 4.1.1; M is then the set of formulas. (That this M consists exactly of the strings with a formation sequence, as clause (F4) says, follows by strong induction along a formation sequence, *Logic and Proof*, Theorem 4.2.2; we do not need it here.) For *Analysis on the Real Line*, Chapter 1, $X = \mathbb{R}$, the

starting element is 0, the step is $x \mapsto x + 1$, and M is $\mathbb{N}$, the set *Analysis on the Real Line*, Chapter 1 works with.

1.2.4.2 Worked examples with a real parameter

Families indexed by an interval of real numbers are the ones analysis meets most. In each example we guess the answer, then prove it by two inclusions; the union needs a **witness index** for each point, and the intersection needs, for each point outside the answer, an index whose member **excludes** it.

Example. Unions and intersections of intervals

Show that:

- (a) $\bigcup_{t \in (0,1)} [t, 1] = (0, 1]$;
- (b) $\bigcap_{t \in (0,\infty)} (-t, t) = \{0\}$, inside $X = \mathbb{R}$;
- (c) $\bigcap_{t \in (0,\infty)} (0, t) = \emptyset$, inside $X = \mathbb{R}$.

Solution.

- (a) ($\subseteq$) Let $x \in \bigcup_{t \in (0,1)} [t, 1]$. Then there is $t \in (0, 1)$ with $x \in [t, 1]$; choose such a t . So $0 < t \leq x \leq 1$, and $x \in (0, 1]$.
 ($\supseteq$) Let $x \in (0, 1]$. We need an index $t \in (0, 1)$ with $t \leq x$. Take $t := x/2$. Then $0 < x/2 \leq 1/2 < 1$, so $t \in (0, 1)$; and $x/2 < x \leq 1$, so $x \in [t, 1]$. Hence x lies in the union. (The witness depends on x , which is allowed: the union asks for **some** index for each point.)
- (b) ($\supseteq$) For every $t > 0$, $-t < 0 < t$, so $0 \in (-t, t)$. Since t was arbitrary, 0 lies in the intersection.
 ($\subseteq$) Let x lie in the intersection. For a real number, $x < 0$ or $x = 0$ or $x > 0$, and we argue by cases. If $x < 0$ or $x > 0$, then $|x| > 0$, so $t := |x|$ is an index, and $x \in (-|x|, |x|)$ would give $|x| < |x|$ (since x is $|x|$ or $-|x|$), which is false; this contradicts the assumption that x lies in every member, so these cases cannot occur. In the remaining case $x = 0$, and $x \in \{0\}$. The index $t = |x|$ depends on the point x it excludes.
- (c) Let x lie in the intersection. With the index $t = 1$, $x \in (0, 1)$, so $x > 0$. Then x is itself an index, and $x \in (0, x)$ would give $x < x$, which is false. So nothing lies in the intersection, and it is empty.

Part (c) is a degenerate case worth remembering. Each interval $(0, t)$ is non-empty, and the intervals are **nested**: $(0, s) \subseteq (0, t)$ whenever $s \leq t$. Yet nothing lies in all of them. *Analysis on the Real Line*, Chapter 1 meets the same phenomenon with a sequence of intervals, and it is why its nested interval theorem needs **closed** intervals.

A nonexample by minimal change. Close the left endpoint of (c): $\bigcap_{t > 0} [0, t) = \{0\}$. The number 0 now lies in every member, since $0 \leq 0 < t$; and any $x \neq 0$ is excluded as before (if $x < 0$, by $[0, 1)$; if $x > 0$, by $[0, x)$). One endpoint changes the answer from $\emptyset$ to $\{0\}$, and the element that (c) loses is exactly the endpoint 0.

Quick check. Compute $\bigcup_{t \in [1,2]} [0, t]$ and $\bigcap_{t \in [1,2]} [0, t]$, inside $X = \mathbb{R}$.

Solution. The union is $[0, 2]$: each $[0, t] \subseteq [0, 2]$ since $t \leq 2$, and $[0, 2]$ is itself a member ($t = 2$). The intersection is $[0, 1]$: $[0, 1] \subseteq [0, t]$ for every $t \geq 1$, and $[0, 1]$ is itself a member ($t = 1$). In general, when the members are nested and there is a largest and a smallest member, the union is the largest and the intersection the smallest.

1.2.5 De Morgan's laws for families

The laws of set algebra have versions for families, and the most used are De Morgan's laws: the complement of a union is the intersection of the complements, and the complement of an intersection is the union of the complements. In analysis they convert statements about open sets into statements about closed sets, and back. By the dictionary, they should be the quantifier negation laws of *Logic and Proof*, Chapter 4, read through membership: “not in some A_i ” is “outside every A_i ”, and “not in every A_i ” is “outside some A_i ”.

Theorem 1.2.3. De Morgan's Laws for Families

Let X be a set and $(A_i)_{i \in I}$ a family of subsets of X , with any index set I , including $I = \emptyset$. Then

$$X \setminus \bigcup_{i \in I} A_i = \bigcap_{i \in I} (X \setminus A_i) \quad \text{and} \quad X \setminus \bigcap_{i \in I} A_i = \bigcup_{i \in I} (X \setminus A_i).$$

All intersections here are taken inside X ; the family $(X \setminus A_i)_{i \in I}$ is again a family of subsets of X , so both sides make sense.

Idea. Unwind both sides of each law through membership, for an element $x \in X$. In the first law, the left side says $\neg \exists i \in I (x \in A_i)$, and the right side says $\forall i \in I \neg(x \in A_i)$. In the second, the left side says $\neg \forall i \in I (x \in A_i)$, and the right side says $\exists i \in I \neg(x \in A_i)$. So each law is a quantifier negation law, and each inclusion is one of its directions. We write the first law out rule by rule, since it is the chapter's model element-chase for families, and the second at level 3, tagging only the step it is about.

Before reading on: which law of *Logic and Proof*, Chapter 4 turns “not in every A_i ” into “outside some A_i ”, and which of its directions is classical?

Proof. First law. ($\subseteq$) Let $x \in X \setminus \bigcup_{i \in I} A_i$ [$\forall$ I and $\rightarrow$ I, opening]. By [Definition 1.2.1](#), $x \in X$ and $x \notin \bigcup_{i \in I} A_i$ [$\leftrightarrow$ E, $\wedge$ E]. Let $i \in I$ [$\forall$ I, Proposition 6.1.1]. Suppose $x \in A_i$ [$\neg$ I, opening]. Then $\exists j \in I (x \in A_j)$, with i as the witness [$\exists$ I], so $x \in \bigcup_{j \in I} A_j$ by [Definition 1.2.4](#). This contradicts $x \notin \bigcup_{j \in I} A_j$ [$\neg$ E]. Hence $x \notin A_i$ [$\neg$ I, closing]. With $x \in X$, this gives $x \in X \setminus A_i$ [$\wedge$ I, $\leftrightarrow$ E]. Since $i \in I$ was arbitrary, $\forall i \in I (x \in X \setminus A_i)$ [$\forall$ I, closing]. With $x \in X$, this is $x \in \bigcap_{i \in I} (X \setminus A_i)$ by [Definition 1.2.4](#). Since x was arbitrary, the inclusion holds [$\rightarrow$ I, $\forall$ I, closing].

($\supseteq$) Let $x \in \bigcap_{i \in I} (X \setminus A_i)$ [$\forall$ I and $\rightarrow$ I, opening]. Then $x \in X$, and $x \in X \setminus A_i$ for every $i \in I$ [$\leftrightarrow$ E, $\wedge$ E]. Suppose $x \in \bigcup_{i \in I} A_i$ [$\neg$ I, opening]. Then there is $i \in I$ with $x \in A_i$; choose such an index i_0 [$\exists$ E, opening, with the new name i_0]. Applying the hypothesis at i_0 , $x \in X \setminus A_{i_0}$ [$\forall$ E], so $x \notin A_{i_0}$ [$\leftrightarrow$ E, $\wedge$ E], which contradicts $x \in A_{i_0}$ [$\neg$ E]. The contradiction $\perp$ does not mention i_0 , so it stands outside the choice [$\exists$ E, closing]. Hence $x \notin \bigcup_{i \in I} A_i$ [$\neg$ I, closing], and with $x \in X$, $x \in X \setminus \bigcup_{i \in I} A_i$ [$\wedge$ I, $\leftrightarrow$ E]. Since x was arbitrary, the inclusion holds [$\rightarrow$ I, $\forall$ I, closing].

closing].

By [Proposition 1.1.1](#), the first law holds.

Second law. ($\subseteq$) Let $x \in X \setminus \bigcap_{i \in I} A_i$, so $x \in X$ and $x \notin \bigcap_{i \in I} A_i$. Then $\forall i \in I (x \in A_i)$ fails: if it held, then together with $x \in X$ it would put x in $\bigcap_{i \in I} A_i$. By the bounded quantifier negation law of *Logic and Proof*, Chapter 4 ($\neg \forall i \in I \varphi$ gives $\exists i \in I \neg \varphi$, the bound kept), there is $i \in I$ with $x \notin A_i$ [RAA, inside the quantifier negation law]. Choose such an index i_0 . Then $x \in X \setminus A_{i_0}$, so $x \in \bigcup_{i \in I} (X \setminus A_i)$, with i_0 as the witness.

($\supseteq$) Let $x \in \bigcup_{i \in I} (X \setminus A_i)$, and choose $i_0 \in I$ with $x \in X \setminus A_{i_0}$. Then $x \in X$ and $x \notin A_{i_0}$. If x were in $\bigcap_{i \in I} A_i$, then $x \in A_{i_0}$ by the definition, applied at i_0 ; so $x \notin \bigcap_{i \in I} A_i$, and $x \in X \setminus \bigcap_{i \in I} A_i$.

By [Proposition 1.1.1](#), the second law holds. This proves the theorem. ■

The answer to the prompt: the bounded quantifier negation law of *Logic and Proof*, Chapter 4 turns $\neg \forall i \in I$ into $\exists i \in I \neg$, and that direction is classical because it rests on the unbounded law $\neg \forall x \varphi \vdash \exists x \neg \varphi$, which needs RAA. It is used once, in the ($\subseteq$) half of the second law. The first law and the ($\supseteq$) half of the second use no classical rule: they are the directions of quantifier negation that *Logic and Proof*, Chapter 4 derives without RAA. The pattern is the one met in the difference laws of [Theorem 1.2.1](#), which are the case $I = \{1, 2\}$, $X = A$, $A_1 = A \cap B$, $A_2 = A \cap C$ (replacing B and C by $A \cap B$ and $A \cap C$ changes neither side): “not in every member” does not by itself name the member it misses.

The empty family. The theorem was stated for every index set, and the conventions of [Definition 1.2.4](#) are what make that true. For $I = \emptyset$, the first law reads $X \setminus \emptyset = X$, since the empty union is $\emptyset$ and the empty intersection is X ; and the second reads $X \setminus X = \emptyset$, since the empty intersection is X and the empty union is $\emptyset$. Both are true. Had the empty intersection been defined as anything other than X , the laws would have needed the hypothesis $I \neq \emptyset$.

Warning. De Morgan swapped the wrong way. “The complement of a union is the union of the complements” is false. Take $X = \{1, 2\}$, $A_1 = \{1\}$, $A_2 = \{2\}$. Then $X \setminus (A_1 \cup A_2) = X \setminus X = \emptyset$, but $(X \setminus A_1) \cup (X \setminus A_2) = \{2\} \cup \{1\} = \{1, 2\}$. Complementing turns $\cup$ into $\cap$ and $\cap$ into $\cup$, just as negation turns $\exists$ into $\forall$ and $\forall$ into $\exists$; nothing stays the same.

1.2.6 Exercises

1.2.6.1 A. Check your understanding

Exercise. A1

These parts are about [Theorem 1.2.3](#) and its proof.

- Which law of *Logic and Proof*, Chapter 4 is used in the ($\subseteq$) half of the second law, and which of its directions is classical?
- Run the first law on $X = \mathbb{Z}$, $I = \{2, 3\}$, and A_i the set of integer multiples of i . Describe both sides in words, and name two integers in each side and two outside it.
- What are the two sides of each law when $I = \emptyset$?

- (d) Let $x \in X$. Translate $x \notin \bigcap_{i \in I} A_i$ into a statement with $\exists$. What is the correct translation if $x \in X$ is **not** assumed?

Solution.

- (a) The bounded quantifier negation law of *Logic and Proof*, Chapter 4: $\neg \forall i \in I \varphi$ gives $\exists i \in I \neg \varphi$, with the bound kept. That direction is classical: it rests on the unbounded law $\neg \forall x \varphi \vdash \exists x \neg \varphi$, which needs RAA.
- (b) The left side $\mathbb{Z} \setminus (A_2 \cup A_3)$ is the set of integers divisible by neither 2 nor 3. The right side $(\mathbb{Z} \setminus A_2) \cap (\mathbb{Z} \setminus A_3)$ is the set of integers that are odd and not multiples of 3. By the theorem they are equal. For instance 1 and 5 (also -1 , 7) lie in both, and 3 and 4 (also 6) lie in neither.
- (c) For $I = \emptyset$, the empty union is $\emptyset$ and the empty intersection is X . First law: $X \setminus \emptyset = X$ and $\bigcap_{i \in \emptyset} (X \setminus A_i) = X$. Second law: $X \setminus X = \emptyset$ and $\bigcup_{i \in \emptyset} (X \setminus A_i) = \emptyset$.
- (d) By [Definition 1.2.4](#), $x \in \bigcap_{i \in I} A_i$ means $x \in X \wedge \forall i \in I (x \in A_i)$. For $x \in X$, its negation is $\neg \forall i \in I (x \in A_i)$, which by the law of (a) is $\exists i \in I (x \notin A_i)$. Without $x \in X$, the negation of the conjunction is, by *Logic and Proof*, Proposition 2.5.5 (d), “ $x \notin X$ or $\exists i \in I (x \notin A_i)$ ”: an object may fail to be in the intersection simply because it is not in X , since the intersection is taken inside X .

1.2.6.2 B. Practice

Exercise. B1: Intersection and difference

Prove that $A \cap (B \setminus C) = (A \cap B) \setminus (A \cap C)$ for all sets A, B, C , and name the propositional equivalence behind it.

Solution. With p, q, r for $x \in A, x \in B, x \in C$, the membership conditions are $p \wedge (q \wedge \neg r)$ and $(p \wedge q) \wedge \neg(p \wedge r)$, and the equivalence behind the law is $p \wedge (q \wedge \neg r) \equiv (p \wedge q) \wedge \neg(p \wedge r)$. Both directions have short direct derivations, which the element-chase follows.

($\subseteq$) Let $x \in A \cap (B \setminus C)$. Then $x \in A, x \in B$, and $x \notin C$. So $x \in A \cap B$. If x were in $A \cap C$, then $x \in C$, a contradiction; so $x \notin A \cap C$. Hence $x \in (A \cap B) \setminus (A \cap C)$.

($\supseteq$) Let $x \in (A \cap B) \setminus (A \cap C)$. Then $x \in A, x \in B$, and $x \notin A \cap C$. If x were in C , then, since $x \in A, x \in A \cap C$, a contradiction; so $x \notin C$. Hence $x \in B \setminus C$, and $x \in A \cap (B \setminus C)$.

By [Proposition 1.1.1](#) the sets are equal. Both halves use $\neg$ I and no classical rule.

Exercise. B2: Three families of intervals

For each family, indexed by $t \in (0, \infty)$, compute the union and the intersection (inside $X = \mathbb{R}$), and prove each answer by two inclusions.

- (a) $A_t = [t, 1 + t]$;
 (b) $B_t = (-t, 1/t]$;
 (c) $C_t = (t, 2t)$.

Solution.

- (a) *Union* $(0, \infty)$. ($\subseteq$) If $x \in [t, 1+t]$ for some $t > 0$, then $x \geq t > 0$. ($\supseteq$) If $x > 0$, take $t := x$: then $x \in [x, 1+x]$. *Intersection* $\emptyset$. Let x be any real number and take $t := |x| + 1 > 0$. Then $x \leq |x| < t$, so $x \notin [t, 1+t]$. So no real number lies in every A_t .
- (b) *Union* $\mathbb{R}$. ($\subseteq$) Every B_t is a set of reals. ($\supseteq$) Let $x \in \mathbb{R}$. If $x \leq 0$, take $t := 1 - x \geq 1$: then $-t = x - 1 < x$, and $1/t > 0 \geq x$, so $x \in B_t$. If $x > 0$, take $t := 1/(2x)$: then $-t < 0 < x$ and $1/t = 2x > x$, so $x \in B_t$. *Intersection* $\{0\}$. ($\supseteq$) For every $t > 0$, $-t < 0 < 1/t$. ($\subseteq$) Let x lie in every B_t . If $x < 0$, take $t := -x > 0$: $x \in B_t$ would need $-t < x$, that is, $x < x$. If $x > 0$, take $t := 1/x$: $x \in B_t$ would need $x < 1/t = x$. Both cases are impossible, so the remaining case of $x < 0$, $x = 0$, $x > 0$ holds: $x = 0$.
- (c) *Union* $(0, \infty)$. ($\subseteq$) If $t < x < 2t$ for some $t > 0$, then $x > 0$. ($\supseteq$) If $x > 0$, take $t := 2x/3$: then $t = 2x/3 < x$ and $2t = 4x/3 > x$, so $x \in C_t$. *Intersection* $\emptyset$. Already two members are disjoint: $C_1 = (1, 2)$ and $C_2 = (2, 4)$ have no common element, since $x < 2$ and $x > 2$ cannot both hold. An element of every C_t would lie in both.

In each union the witness t depends on the point x . In the intersections of (a) and (b) the index that excludes x depends on x too; in (c) two fixed disjoint members, C_1 and C_2 , already exclude every x . All of these are allowed by the definitions.

Exercise. B3: Canceling a union

Give sets A, B, C with $A \cup B = A \cup C$ and $B \neq C$. Then list the properties that any correct answer must have, so that a different answer can be checked against them.

Solution. Take $A = \{1\}$, $B = \emptyset$, $C = \{1\}$. Then $A \cup B = \{1\} = A \cup C$, and $B \neq C$, with the witness $1 \in C$, $1 \notin B$.

Properties of any answer. Call an element **odd one out** if it lies in exactly one of B and C .

- $B \neq C$ holds exactly when there is at least one odd one out (by [Definition 1.1.1](#); getting such an element from $B \neq C$ uses the classical direction of quantifier negation, *Logic and Proof*, Chapter 4).
- $A \cup B = A \cup C$ holds exactly when **every** odd one out lies in A . Indeed, if $x \in B \setminus C$ and $x \notin A$, then $x \in A \cup B$ but $x \notin A \cup C$; symmetrically for $x \in C \setminus B$. Conversely, if every odd one out lies in A , then an element of $A \cup B$ either lies in A , or lies in B and then, by excluded middle on $x \in C$ (*Logic and Proof*, Proposition 2.5.3), in C or in A ; so it lies in $A \cup C$, and symmetrically.

So an answer is correct exactly when some element lies in exactly one of B, C , and every such element lies in A . The union with A “hides” the difference between B and C ; the lesson is that $\cup$ cannot be canceled.

1.2.6.3 C. Going deeper

Exercise. C1: Power sets and operations

Let A and B be sets.

- (a) Decide whether $\mathcal{P}(A \setminus B) = \mathcal{P}(A) \setminus \mathcal{P}(B)$ can hold. Then prove that every **non-empty** element of $\mathcal{P}(A \setminus B)$ lies in $\mathcal{P}(A) \setminus \mathcal{P}(B)$.
- (b) Show by an example that $\mathcal{P}(A) \cup \mathcal{P}(B) = \mathcal{P}(A \cup B)$ can fail. Then prove that it holds if and only if $A \subseteq B$ or $B \subseteq A$.

Solution (Hint 1). In (a), test the empty set against both sides. In (b), one inclusion always holds; for the other, ask which subset of $A \cup B$ could fail to be a subset of A and of B .

Solution (Hint 2). For (b) ($\Rightarrow$), argue by the contrapositive: if $A \not\subseteq B$ and $B \not\subseteq A$, choose $a \in A \setminus B$ and $b \in B \setminus A$.

Solution.

- (a) **Never.** $\emptyset \in \mathcal{P}(A \setminus B)$, since $\emptyset$ is a subset of every set. But $\emptyset \in \mathcal{P}(B)$ as well, so $\emptyset \notin \mathcal{P}(A) \setminus \mathcal{P}(B)$. The empty set is a witness for every A and B .
Now let $S \in \mathcal{P}(A \setminus B)$ be non-empty, and choose $s \in S$. Since $S \subseteq A \setminus B \subseteq A$, we have $S \in \mathcal{P}(A)$. Since $s \in A \setminus B$, $s \notin B$; so s is an element of S outside B , and $S \not\subseteq B$, that is, $S \notin \mathcal{P}(B)$. Hence $S \in \mathcal{P}(A) \setminus \mathcal{P}(B)$.
- (b) *Failure.* For $A = \{1\}$, $B = \{2\}$: $\{1, 2\} \in \mathcal{P}(A \cup B)$, but $\{1, 2\}$ is a subset of neither A nor B , so $\{1, 2\} \notin \mathcal{P}(A) \cup \mathcal{P}(B)$.

One inclusion always holds. If $S \in \mathcal{P}(A)$, then $S \subseteq A \subseteq A \cup B$ ([Proposition 1.1.2](#)), so $S \in \mathcal{P}(A \cup B)$; likewise for $S \in \mathcal{P}(B)$. So $\mathcal{P}(A) \cup \mathcal{P}(B) \subseteq \mathcal{P}(A \cup B)$.

($\Leftarrow$) Suppose $A \subseteq B$. Then $A \cup B = B$ (if $x \in A \cup B$, then $x \in A \subseteq B$ or $x \in B$; and $B \subseteq A \cup B$). Let $S \in \mathcal{P}(A \cup B)$. Then $S \subseteq B$, so $S \in \mathcal{P}(B) \subseteq \mathcal{P}(A) \cup \mathcal{P}(B)$. With the inclusion above, equality holds. The case $B \subseteq A$ is the same with A and B exchanged.

($\Rightarrow$) We prove the contrapositive. Suppose that neither $A \subseteq B$ nor $B \subseteq A$. By the negation of $\subseteq$ ([Exercise \(A1\)](#) (a) records it), there is $a \in A$ with $a \notin B$, and there is $b \in B$ with $b \notin A$; choose such a and b . Let $S := \{a, b\}$. Then $S \subseteq A \cup B$, so $S \in \mathcal{P}(A \cup B)$. But $S \not\subseteq A$, because $b \in S$ and $b \notin A$; and $S \not\subseteq B$, because $a \in S$ and $a \notin B$. So $S \notin \mathcal{P}(A) \cup \mathcal{P}(B)$, and the two sets differ. Two steps here are classical: turning “ $A \not\subseteq B$ ” into a witness uses the classical directions of quantifier negation and of the law for negated implications (*Logic and Proof*, Chapter 4; see [Exercise \(A1\)](#) (a)), and concluding the implication from its contrapositive is *Logic and Proof*, Proposition 2.5.2.

Marking scheme. (10 points) - 2: (a) the empty set as the witness that the equality never holds, with both memberships checked. - 2: (a) a non-empty $S \subseteq A \setminus B$ is shown to lie in $\mathcal{P}(A)$ and, via an element outside B , not in $\mathcal{P}(B)$. - 1: (b) a counterexample with the offending subset named. - 2: (b) the inclusion $\mathcal{P}(A) \cup \mathcal{P}(B) \subseteq \mathcal{P}(A \cup B)$ and the direction ($\Leftarrow$). - 3: (b) ($\Rightarrow$): witnesses $a \in A \setminus B$ and $b \in B \setminus A$ named, and $\{a, b\}$ shown to lie in $\mathcal{P}(A \cup B)$ and in neither $\mathcal{P}(A)$ nor $\mathcal{P}(B)$.

A direct proof also earns the 3: assume equality and $A \not\subseteq B$, fix $a \in A \setminus B$ (1 point); for arbitrary $b \in B$, $\{a, b\} \in \mathcal{P}(A \cup B) = \mathcal{P}(A) \cup \mathcal{P}(B)$, and $\{a, b\} \not\subseteq B$ because $a \notin B$, so $\{a, b\} \subseteq A$ and $b \in A$, giving $B \subseteq A$ (2 points). Both routes also end with a classical step: excluded middle on $A \subseteq B$ in the direct route, the contrapositive in the book’s.

Fatal errors (0 for the step): in (a), testing only non-empty subsets and concluding equality; in (b) ($\Rightarrow$), in the contrapositive route, using one witness a for both failures, or a set such as $\{a\}$ that lies in $\mathcal{P}(A)$.

Exercise. C2: Find the error

Here is a “proof” that $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$ for all sets A, B, C .

(1) Let $x \in (A \cup B) \cap (A \cup C)$. (2) Then $x \in A \cup B$ and $x \in A \cup C$. (3) We consider two cases, according to $x \in A \cup B$. (4) *Case* $x \in A$. Then $x \in A \cup (B \cap C)$. (5) *Case* $x \in B$. Since $x \in A \cup C$, $x \in C$. (6) So $x \in B \cap C$, and hence $x \in A \cup (B \cap C)$. (7) In both cases $x \in A \cup (B \cap C)$, which proves the inclusion.

- (a) Is the conclusion true?
- (b) Which line fails, which rule does it break, and which entry of the catalogue of *Logic and Proof*, Chapter 5 does it commit?
- (c) Give a counterexample to the illegal step itself: sets and an element for which the line’s premise holds and its conclusion fails.
- (d) Repair the proof.

Solution.

- (a) Yes. It is the ($\supseteq$) half of the second distributive law of [Theorem 1.2.1](#) (c).
- (b) Line (5). From $x \in A \cup C$, that is, $x \in A$ or $x \in C$, it concludes $x \in C$, dropping the case $x \in A$. A disjunction is used by $\vee E$, which needs **both** cases; the line uses only one. This is a missing case, entry (N7) of the catalogue of *Logic and Proof*, Chapter 5.
- (c) Take $A = \{1\}$, $B = \{1\}$, $C = \emptyset$, and $x = 1$. Then $x \in B$ (the case of line (5)) and $x \in A \cup C$ (the premise of the line), but $x \notin C$. The step is illegal although the conclusion of the whole proof is true: here $x \in A$, so $x \in A \cup (B \cap C)$ by the other case.
- (d) Replace lines (5)–(6) by: *Case* $x \in B$. Since $x \in A \cup C$, there are two sub-cases. If $x \in A$, then $x \in A \cup (B \cap C)$. If $x \in C$, then $x \in B \cap C$, so $x \in A \cup (B \cap C)$. In both sub-cases $x \in A \cup (B \cap C)$. The rest of the proof stands.

1.3 Ordered Pairs, Products and Relations

Sections 1.1 and 1.2 built sets and combined them: a set is fixed by its elements, and $\cup, \cap, \setminus$ are $\vee, \wedge, \wedge \neg$ read through membership. None of this can yet say “the point $(1, 2)$ of the plane”, “3 is less than 5”, or “ y is the square of x ”. Each of these is about two objects **in a given order**, and a set forgets order. This section adds one new kind of object, the ordered pair (with its longer versions, the n -tuples), and then uses it twice: to build the product $A \times B$ of two sets, and to say what a relation between two sets is. The next section needs exactly this much: a function will be given by a relation of a special kind, its graph, together with a declared target set.

1.3.1 Order matters: ordered pairs

A set remembers which objects are its elements and nothing else. By extensionality (Definition 1.1.1), $\{1, 2\} = \{2, 1\}$, since both sets have the elements 1 and 2; and $\{1, 1\} = \{1\}$, since both have the element 1 and nothing else. For a set that is exactly right. For a point of the plane it is exactly wrong. The point $(1, 2)$ lies one unit to the right of the vertical axis and two units above the horizontal one; the point $(2, 1)$ lies elsewhere. And the point $(1, 1)$ on the diagonal is a perfectly good point, with the same number in both places.

So we need an object built from two objects a and b that remembers **which came first**, and that still has two places when $a = b$. What we need to know about such an object is one fact: when two of them are equal.

An ordered pair is equal to another exactly when they agree in the first place and in the second place.

Definition 1.3.1. Ordered pair

For any objects a and b there is an object (a, b) , the **ordered pair** of a and b ; we call a its **first coordinate** and b its **second coordinate**. Ordered pairs have the **characteristic property**: for all a, b, c, d ,

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

In words: to decide whether two pairs are equal, compare first coordinates with first coordinates and second with second; nothing else matters, and both comparisons must succeed. The direction $(\Leftarrow)$ is nothing new: equal inputs give equal pairs, as for any construction ($=$ E, Logic and Proof, Definition 5.7.1). The direction $(\Rightarrow)$ is the content. It says that from the pair we can read off its coordinates, each in its place.

We do not say here **what** the object (a, b) is. Besides the characteristic property, we use one working convention: pairs and n -tuples are a different kind of object from numbers, just as sets are (Section 1.1), so **a pair is not a number**; and tuples of different lengths are different kinds of object, so a triple is not a pair. Nothing else about pairs is used in this chapter. *Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* , Chapter 1 shows that a pair with this property can be built from sets alone, so that nothing new has to be assumed. Until then, the existence of pairs with this property is taken as given, alongside the working rules (W1)–(W6) of Section 1.1.

A convention about parentheses. The symbol (a, b) now has two meanings. For real numbers $a < b$ it is also the open interval $\{x \in \mathbb{R} : a < x < b\}$. The book keeps both, and context decides, as it does in every analysis text. In “ $(0, 1)$ contains $\frac{1}{2}$ ” it is an interval. In “the point $(0, 1)$ lies on the unit circle” it is a pair. Where both readings are near each other, the text says which is meant. This convention is stated once, here.

Example. Solving an equation between pairs

Find all real numbers x and y with $(x + y, x - y) = (3, 1)$. Then decide whether $(1, 2) = (2, 1)$ and whether $(1, 1) = (1, 1)$.

Solution. By the characteristic property, $(x + y, x - y) = (3, 1)$ holds if and only if

$$x + y = 3 \quad \text{and} \quad x - y = 1.$$

Adding the two equations gives $2x = 4$, so $x = 2$; subtracting them gives $2y = 2$, so $y = 1$. Conversely, $x = 2$ and $y = 1$ give $(x + y, x - y) = (3, 1)$. So there is exactly one solution, $x = 2, y = 1$.

The pairs $(1, 2)$ and $(2, 1)$ are **not** equal. If they were, the characteristic property would give $1 = 2$ for the first coordinates. So $(1, 2) \neq (2, 1)$, by $\neg I$. The same test on $(1, 1)$ and $(1, 1)$ passes in both places, so they are equal; and $(1, 1)$ has two places, although both hold the same number.

The first part is typical. One equation between pairs is **two** equations between coordinates, and this is how pairs are used: an equation of pairs is split by the characteristic property, and the coordinates are compared one at a time.

Nonexample by minimal change. Replace the parentheses by braces. The set $\{1, 2\}$ has the same two numbers as the pair $(1, 2)$, but the characteristic property fails for sets: $\{1, 2\} = \{2, 1\}$, and yet $1 \neq 2$. What survives is membership: 1 and 2 are both in $\{1, 2\}$. What is lost is order, the one thing the pair was made to keep.

Why this definition. Only $(\Rightarrow)$ carries content. Without it, the pair could forget its coordinates, as $\{1, 2\}$ forgets which number came first. The direction $(\Leftarrow)$ could be dropped from the definition without loss: $= E$ already gives it for (a, b) , as for any construction. It is stated so that the property reads as a single test for equality of pairs.

Longer lists. The same idea gives ordered triples (a, b, c) , and more generally ordered n -tuples $(x_1, \dots, x_n)$ for $n \in \mathbb{N}, n \geq 2$. Like pairs, n -tuples are taken as given, and the n -tuples with all coordinates in a set form a set, by the working rule (W5) of Section 1.1. Two n -tuples are equal if and only if they agree in every place: $(x_1, \dots, x_n) = (y_1, \dots, y_n)$ means $x_k = y_k$ for every $k \in \{1, \dots, n\}$. A triple is a new kind of object, not a pair whose first coordinate is a pair (by the convention stated after Definition 1.3.1). The pairs $((1, 2), 3)$ and $(1, (2, 3))$ carry the same three numbers in the same order, and it is tempting to call both “the triple $(1, 2, 3)$ ”; that is an identification made by convention, and exercise A1 asks whether the two pairs are equal.

Quick check. For which real numbers t is $(t^2, t) = (1, -1)$? For which is $\{t^2, t\} = \{1, -1\}$?

Solution. The pair equation holds if and only if $t^2 = 1$ and $t = -1$, so exactly for $t = -1$. The set equation holds for $t = -1$ too, since $\{1, -1\} = \{1, -1\}$. It holds for no other t : the element -1 of the right side must be t^2 or t , and $t^2 \geq 0$, so $t = -1$. Here the sets agree exactly where the pairs do, but that is luck. The pair equation $(t, 1) = (1, t)$ holds only for $t = 1$, while the set equation $\{t, 1\} = \{1, t\}$ holds for every t .

1.3.2 Cartesian products

With pairs in hand, the plane is the set of all pairs of real numbers. More generally, from any two sets we can collect all pairs whose first coordinate comes from the first set and whose second

comes from the second. Taking such a collection is the working rule (W5) of Section 1.1: it is a set.

Definition 1.3.2. Cartesian product

Let A and B be sets. Their **Cartesian product** is the set

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

of all ordered pairs with first coordinate in A and second coordinate in B . Written out, an object p is an element of $A \times B$ if and only if

$$\exists a \in A \exists b \in B \ p = (a, b).$$

For a set X we write $X^2 := X \times X$, and, for $n \in \mathbb{N}$ with $n \geq 2$, X^n is the set of all n -tuples $(x_1, \dots, x_n)$ with every $x_k \in X$.

In words: an element of $A \times B$ is a **pair**, and **there exist** coordinates, one in A and one in B , from which it is made. The membership condition hides two existential quantifiers, so using it means naming the coordinates, by $\exists E$ (Logic and Proof, Proposition 6.1.1), and then matching them against any other naming by the characteristic property of pairs. That two-step move is the one move of this section; (*) below packages it.

For a pair the condition simplifies, and we use the simpler form constantly:

$$(x, y) \in A \times B \iff x \in A \text{ and } y \in B. \quad (*)$$

Here is why. If $x \in A$ and $y \in B$, then x and y themselves are coordinates of the required kind, and $(x, y) \in A \times B$ by $\exists I$. Conversely, let $(x, y) \in A \times B$. Then there are $a \in A$ and $b \in B$ with $(x, y) = (a, b)$ [$\exists E$], and the characteristic property (Definition 1.3.1) gives $x = a$ and $y = b$. So $x \in A$ and $y \in B$. The characteristic property is used exactly once, in the direction ($\Rightarrow$); without it, (x, y) might equal (a, b) with the coordinates swapped or lost, and nothing could be concluded.

We write (*) for this equivalence in the rest of the section. It is the reason a product is easy to work with: membership of a pair in a product is checked one coordinate at a time.

Set-builders with a pair pattern. We write $\{(x, y) \in A \times B : P(x, y)\}$ for $\{p \in A \times B : \text{there are } x, y \text{ with } p = (x, y) \text{ and } P(x, y)\}$, a set-builder in the sense of Definition 1.1.3. By (*) and the characteristic property, (x, y) belongs to it if and only if $x \in A$, $y \in B$ and $P(x, y)$. The notation $\{(a, b) : a \in A, b \in B\}$ in Definition 1.3.2 is different: it names no ambient set, so it is not a set-builder. It is notation for the set that (W5) provides, and the displayed membership condition is its definition.

Example. Products, small and large

- (a) Let a, b, c be three different objects, none of them a number. List $\{1, 2\} \times \{a, b, c\}$ and $\{a, b, c\} \times \{1, 2\}$.
- (b) Describe $\mathbb{R}^2$ and $[0, 1] \times [0, 1]$ as sets of points of the plane.
- (c) What is $\{1, 2\} \times \emptyset$?

Solution.

- (a) Each of the 2 first coordinates is paired with each of the 3 second coordinates:

$$\{1, 2\} \times \{a, b, c\} = \{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c)\},$$

six pairs, all different by the characteristic property, since $1 \neq 2$ and a, b, c are different. Swapping the factors gives the six pairs $(a, 1), (a, 2), (b, 1), (b, 2), (c, 1), (c, 2)$. No pair of the first list is in the second: every pair in the first list has a number as its first coordinate, and every pair in the second list has one of a, b, c there, none of which is a number; so by the characteristic property no pair of one list equals a pair of the other.

- (b) $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is the set of all pairs (x, y) of real numbers: the plane, with x the horizontal and y the vertical coordinate. By (*), $(x, y) \in [0, 1] \times [0, 1]$ if and only if $0 \leq x \leq 1$ and $0 \leq y \leq 1$: the closed unit square, with its four sides.
- (c) It is empty. An element p would be (a, b) with $b \in \emptyset$, and $\emptyset$ has no elements. So $\{1, 2\} \times \emptyset$ has no elements, and a set with no elements is $\emptyset$ ([Proposition 1.1.3](#)).

Part (c) is the degenerate case, and it matters: a product is empty as soon as **one** factor is. The general statement is part of the next proposition.

Quick check. List $\{0, 1\} \times \{0, 1, 2\}$. Is $(1, 1) \in \{1\} \times \emptyset$?

Solution. $\{0, 1\} \times \{0, 1, 2\} = \{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2)\}$, six pairs. And $(1, 1) \notin \{1\} \times \emptyset$: by (*) it would need $1 \in \emptyset$. In fact $\{1\} \times \emptyset$ has no elements at all, as in part (c) of the example.

Warning. The product is not commutative: $A \times B$ and $B \times A$ are in general different sets. For $A = \{1\}$ and $B = \{2\}$, $A \times B = \{(1, 2)\}$ and $B \times A = \{(2, 1)\}$, and $(1, 2) \neq (2, 1)$ by [Example \(Solving an equation between pairs\)](#). The two products have “the same size”, but they do not have the same elements. (They are equal when $A = B$, and when one of the sets is empty. Exercise C2 finds exactly when even the inclusion $A \times B \subseteq B \times A$ holds.)

1.3.3 Products and the set operations

Section 1.2 showed that $\cup$, $\cap$, and $\setminus$ obey the laws of $\vee$, $\wedge$, and $\wedge \neg$. The product interacts with them just as simply, one factor at a time. We use these laws for regions of the plane (exercise B1) and for functions out of the empty set (Section 1.4), so we state them all.

Proposition 1.3.1. Products and Set Operations

For all sets A, B, C :

$$A \times (B \cup C) = (A \times B) \cup (A \times C),$$

$$A \times (B \cap C) = (A \times B) \cap (A \times C),$$

$$A \times (B \setminus C) = (A \times B) \setminus (A \times C),$$

$$A \times \emptyset = \emptyset,$$

and the mirror-image laws, with the varying factor on the left:

$$(B \cup C) \times A = (B \times A) \cup (C \times A),$$

$$(B \cap C) \times A = (B \times A) \cap (C \times A),$$

$$(B \setminus C) \times A = (B \times A) \setminus (C \times A),$$

$$\emptyset \times A = \emptyset.$$

Idea. Every element of each side is a pair, so we name its coordinates by $\exists E$ and then test membership one coordinate at a time, by $(*)$. The first coordinate lies in A on both sides; the second lies in $B \cup C$ on one side and in “ B or C ” on the other, which is the same condition by [Definition 1.2.1](#). Each law is a law of Section 1.2 in the second coordinate, with the first coordinate carried along unchanged.

Proof. *The $\cup$ law.* We prove two inclusions and conclude by [Proposition 1.1.1](#).

($\subseteq$) Let $p \in A \times (B \cup C)$. By [Definition 1.3.2](#) there are $a \in A$ and $d \in B \cup C$ with $p = (a, d)$ [F1, Proposition 6.1.1]. By [Definition 1.2.1](#), $d \in B$ or $d \in C$, and we argue by cases. If $d \in B$, then $p = (a, d) \in A \times B$ by $(*)$, so $p \in (A \times B) \cup (A \times C)$. If $d \in C$, then likewise $p \in A \times C \subseteq (A \times B) \cup (A \times C)$. In either case $p \in (A \times B) \cup (A \times C)$.

($\supseteq$) Let $p \in (A \times B) \cup (A \times C)$. Then $p \in A \times B$ or $p \in A \times C$. If $p \in A \times B$, there are $a \in A$ and $b \in B$ with $p = (a, b)$ [$\exists E$]; since $b \in B \subseteq B \cup C$, $(*)$ gives $p \in A \times (B \cup C)$. If $p \in A \times C$, the same argument with C in place of B gives $p \in A \times (B \cup C)$.

The $\cap$ law. Let $p \in A \times (B \cap C)$, so $p = (a, d)$ with $a \in A$ and $d \in B \cap C$. Then $d \in B$ and $d \in C$, and $(*)$ gives $p \in A \times B$ and $p \in A \times C$. Conversely, let $p \in (A \times B) \cap (A \times C)$. Since $p \in A \times B$, $p = (a, b)$ with $a \in A$ and $b \in B$. Since also $(a, b) = p \in A \times C$, $(*)$ gives $b \in C$. So $b \in B \cap C$, and $p = (a, b) \in A \times (B \cap C)$.

The $\setminus$ law. Let $p \in A \times (B \setminus C)$, so $p = (a, d)$ with $a \in A$, $d \in B$, and $d \notin C$. Then $p \in A \times B$ by $(*)$. If p were in $A \times C$, then $(*)$ would give $d \in C$, which is false; so $p \notin A \times C$, and $p \in (A \times B) \setminus (A \times C)$. Conversely, let $p \in (A \times B) \setminus (A \times C)$. Then $p = (a, b)$ with $a \in A$ and $b \in B$, and $p \notin A \times C$. If b were in C , then $(*)$ would give $p = (a, b) \in A \times C$, which is false; so $b \notin C$. Hence $b \in B \setminus C$ and $p \in A \times (B \setminus C)$.

The $\emptyset$ law. An element of $A \times \emptyset$ would be a pair (a, b) with $b \in \emptyset$, and there is no such b . So $A \times \emptyset$ has no elements, and it equals $\emptyset$ by [Proposition 1.1.3](#).

The mirror-image laws. The same four arguments, with the roles of the two coordinates exchanged throughout, prove them: the varying set now sits in the first coordinate, and $(*)$ is applied to it there. This proves the proposition. ■

The proofs of the $\setminus$ law do not use excluded middle: both “ $\notin$ ” conclusions are reached by assuming membership and deriving a contradiction, which is $\neg$ I. Only the first law was written with every $\exists$ E marked; from here on, “ $p = (a, b)$ with $a \in A$ and $b \in B$ ” is understood to name fresh coordinates.

A proof about products usually runs like the one above: take an element, write it as a pair with named coordinates, and work coordinatewise. Here is the pattern on concrete sets.

Example. Two rectangles

Show that $([0, 2] \times [0, 1]) \cap ([1, 3] \times [0, 2]) = [1, 2] \times [0, 1]$.

Solution. Draw the two rectangles first: the first is 2 wide and 1 high, with its lower left corner at the origin; the second is 2 wide and 2 high, with its lower left corner at $(1, 0)$. They overlap in the rectangle over $[1, 2]$ of height 1. Now the proof, by two inclusions.

($\subseteq$) Let p be in the intersection. Since $p \in [0, 2] \times [0, 1]$, $p = (x, y)$ with $0 \leq x \leq 2$ and $0 \leq y \leq 1$. Since also $(x, y) = p \in [1, 3] \times [0, 2]$, $(*)$ gives $1 \leq x \leq 3$ and $0 \leq y \leq 2$. Together, $1 \leq x \leq 2$ and $0 \leq y \leq 1$, so $p = (x, y) \in [1, 2] \times [0, 1]$ by $(*)$.

($\supseteq$) Let $p \in [1, 2] \times [0, 1]$, so $p = (x, y)$ with $1 \leq x \leq 2$ and $0 \leq y \leq 1$. Then $0 \leq x \leq 2$, so $p \in [0, 2] \times [0, 1]$; and $1 \leq x \leq 3$, $0 \leq y \leq 2$, so $p \in [1, 3] \times [0, 2]$. Hence p is in the intersection.

By [Proposition 1.1.1](#), the two sets are equal. The picture found the answer; the element-chase proved it.

The degenerate law of the proposition deserves a second look, because it is easy to misuse.

Remark. Combining the two $\emptyset$ laws with one more step gives the full description of an empty product: $A \times B = \emptyset$ **if and only if** $A = \emptyset$ or $B = \emptyset$. The direction ($\Leftarrow$) is the pair of $\emptyset$ laws. For ($\Rightarrow$), suppose $A \times B = \emptyset$. If A has an element a and B has an element b , then $(a, b) \in A \times B = \emptyset$, which is impossible. So it is not the case that both sets have an element; by excluded middle (Logic and Proof, Proposition 2.5.3) applied to “ A has an element”, either A has none, and then $A = \emptyset$ by [Proposition 1.1.3](#), or A has one, and then B has none, so $B = \emptyset$.

Warning. “ $A \times B = \emptyset$, so $A = \emptyset$ ” misses a case. The product $\{1, 2\} \times \emptyset$ is empty, and $\{1, 2\}$ is not. The correct conclusion is the disjunction “ $A = \emptyset$ or $B = \emptyset$ ” of the remark, and an argument that continues from it must treat both cases. Dropping one is the missing case of the catalogue in *Logic and Proof*, Chapter 5, entry (N7): a $\forall$ E with one of its boxes left out.

1.3.4 Relations

“Is less than”, “divides”, “is the square of”, “lies on the unit circle with”: each of these connects an object of one set with an object of another, or of the same set. Each can be described by a rule, a formula, or a sentence. But two different descriptions can pick out exactly the same connections: “ $x < y$ ” and “ $y - x > 0$ ” are different sentences about real numbers, and they hold

for exactly the same pairs (x, y) . In mathematics we want these to count as the **same** relation.

The way to make them the same is to forget the description and keep what it picks out: the set of pairs (x, y) for which it holds. Two descriptions that pick out the same pairs then give the same set, by extensionality.

A relation is the set of pairs it relates.

Definition 1.3.3. Relation

Let A and B be sets. A **relation from A to B** is a subset $R \subseteq A \times B$. For $a \in A$ and $b \in B$ we write $a R b$, read “ a is R -related to b ”, to mean $(a, b) \in R$. A **relation on A** is a relation from A to A , that is, a subset of A^2 .

The **domain** of R is the set of first coordinates that occur, $\{a \in A : \exists b \in B \ a R b\}$, and its **range** is the set of second coordinates that occur, $\{b \in B : \exists a \in A \ a R b\}$. Turning every pair of R around gives a relation from B to A , the **inverse relation**

$$R^{-1} := \{(b, a) \in B \times A : (a, b) \in R\}.$$

In words: a relation **is** a set, namely a set of pairs whose coordinates come from the two given sets. The notation $a R b$ is only a way of writing “ $(a, b) \in R$ ” that reads like “ $a < b$ ”. The domain collects the elements of A that are related to **at least one** element of B ; the range collects those of B that something is related to. The inverse relation turns each pair around: $b R^{-1} a$ exactly when $a R b$.

Two relations are equal exactly when they relate the same pairs, since relations are sets (Definition 1.1.1). So the relation “ $x < y$ ” on $\mathbb{R}$ and the relation “ $y - x > 0$ ” on $\mathbb{R}$ are one relation, as we wanted. That is the whole reason for the definition: a relation is determined by **what** it relates, not by **how** it is described.

Example. Relations as sets of pairs

Write each relation as a set, and find its domain and range.

- (a) The relation $<$ on $\{1, 2, 3\}$.
- (b) Divisibility on $\{1, \dots, 6\}$: $a D b$ when $a \mid b$.
- (c) The unit circle, as a relation on $\mathbb{R}$.
- (d) The graph of squaring, as a relation on $\mathbb{R}$.
- (e) The two extreme relations from A to B : $\emptyset$ and $A \times B$.

Solution.

- (a) The pairs (a, b) of elements of $\{1, 2, 3\}$ with $a < b$ are

$$L := \{(1, 2), (1, 3), (2, 3)\},$$

three of the nine elements of $\{1, 2, 3\}^2$. Its domain is $\{1, 2\}$, since nothing in the set is larger than 3; its range is $\{2, 3\}$, since nothing is smaller than 1.

- (b) Write $a \mid b$ when $b = ak$ for some integer k . For each first coordinate a , the second coordinates $b \in \{1, \dots, 6\}$ with $a \mid b$ are the multiples of a up to 6:

a	the $b \in \{1, \dots, 6\}$ with $a \mid b$
1	1, 2, 3, 4, 5, 6
2	2, 4, 6
3	3, 6
4	4
5	5
6	6

So D has $6 + 3 + 2 + 1 + 1 + 1 = 14$ elements, the pairs (a, b) read off the rows. Its domain and range are both all of $\{1, \dots, 6\}$, because every a divides itself: $(a, a) \in D$.

(c) The unit circle is

$$S := \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\},$$

a relation on $\mathbb{R}$; for instance $0 S 1$, $0 S (-1)$, and $1 S 0$. Its domain is $[-1, 1]$. If $x S y$, then $x^2 = 1 - y^2 \leq 1$, so $-1 \leq x \leq 1$. Conversely, for $-1 \leq x \leq 1$ the number $1 - x^2$ is nonnegative and so has a square root $y = \sqrt{1 - x^2}$, which satisfies $x^2 + y^2 = 1$. By the symmetry of the equation, the range is $[-1, 1]$ too.

(d) The graph of squaring is $Q := \{(x, y) \in \mathbb{R}^2 : y = x^2\}$. Its domain is all of $\mathbb{R}$: every x is related to x^2 . Its range is $[0, \infty)$: a square is nonnegative, and every $y \geq 0$ is the square of $\sqrt{y}$.

(e) Both are subsets of $A \times B$, so both are relations. The **empty relation** $\emptyset$ relates nothing, and its domain and range are empty. The **full relation** $A \times B$ relates everything to everything. If A and B both have elements, its domain is A and its range is B .

The circle and the graph of squaring look alike, both being curves in the plane, but they differ in a way that the next section makes precise. In Q every x is related to **exactly one** y . In S the number 0 is related to two numbers, 1 and -1 , and the number 2 to none. A relation of the first kind will be the graph of a function.

Quick check. Write the inverse relation L^{-1} of part (a) as a set. Which familiar relation on $\{1, 2, 3\}$ is it?

Solution. Turning each pair of L around gives $L^{-1} = \{(2, 1), (3, 1), (3, 2)\}$: the pairs (b, a) with $b > a$. So L^{-1} is the relation $>$ on $\{1, 2, 3\}$. Its domain $\{2, 3\}$ is the range of L , and its range $\{1, 2\}$ is the domain of L .

Nonexample by minimal change. The set $\{(1, 2), (2, 3), 3\}$ looks like the relation “is one less than” on $\{1, 2, 3\}$, with one pair damaged. It is not a relation on $\{1, 2, 3\}$: the element 3 is a number, not a pair (the convention after [Definition 1.3.1](#)), so the set is not a subset of $\{1, 2, 3\}^2$. Remove the 3, or replace it by the pair $(3, 4)$ and enlarge the second set to $\{1, 2, 3, 4\}$, and it becomes a relation. The clause that failed is “ $R \subseteq A \times B$ ”: every element of a relation must be a pair with coordinates in the right sets.

Why this definition. A relation could be defined as a rule, a sentence, or a formula. Each of those raises the question of when two rules are “the same”, and that question has no good answer: “ $x < y$ ” and “ $y - x > 0$ ” are different sentences. The definition as a set answers it

by extensionality. The price is small: a relation may be any set of pairs whatever, such as $\{(1, 3), (2, 2)\}$ on $\{1, 2, 3\}$, and the definition never asks for a rule that produces it. That is a gain, not a loss: in analysis we meet relations and functions that no formula describes.

Warning. A relation “from A to B ” has its direction built in: $R \subseteq A \times B$, and the first coordinate comes from A . The relation L of part (a) contains $(1, 2)$ and not $(2, 1)$. Reading a relation backwards, “since $1 L 2$, also $2 L 1$ ”, treats it as a symmetric relation; relations with that property, and other properties such as transitivity, are studied in Chapter 2. Relating b to a is the job of the inverse relation R^{-1} , a different set.

This definition is all the next section needs. The graph of a function from X to Y will be a relation from X to Y in which every element of X is related to exactly one element of Y ; the function is that graph together with the declared codomain Y . The graph Q qualifies and the circle S does not. The properties that make a relation an order or an equivalence (reflexive, symmetric, antisymmetric, transitive) are the subject of Chapter 2.

1.3.5 Exercises

1.3.5.1 A. Check your understanding

Exercise. A1

- (a) Define $A \times B$ in symbols, with its membership condition written out using quantifiers.
- (b) Is $((1, 2), 3) = (1, (2, 3))$? Justify your answer.
- (c) Write the inverse of the relation $<$ on $\{1, 2, 3\}$ as a set of pairs.
- (d) Is $\emptyset$ a relation on every set A ? Justify your answer.

Solution.

- (a) $A \times B = \{(a, b) : a \in A, b \in B\}$; that is, $p \in A \times B$ if and only if $\exists a \in A \exists b \in B \ p = (a, b)$ (Definition 1.3.2).
- (b) No. Both are ordered pairs, so by the characteristic property (Definition 1.3.1) they are equal if and only if their first coordinates are equal and their second coordinates are equal. The first coordinates are the pair $(1, 2)$ and the number 1. A pair is not a number, by the working convention stated after Definition 1.3.1 (*Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* , Chapter 1, which builds both from sets, bears it out). So the first coordinates differ, and the pairs are not equal. Calling both “the triple $(1, 2, 3)$ ” is an identification made by convention, not an equality.
- (c) $\{(2, 1), (3, 1), (3, 2)\}$, which is the relation $>$ on $\{1, 2, 3\}$.
- (d) Yes. A relation on A is a subset of $A \times A$, and $\emptyset \subseteq A \times A$: the condition “ $x \in \emptyset \rightarrow x \in A \times A$ ” holds for every x , because its hypothesis is false (Logic and Proof, Proposition 2.1.2, Definition 1.1.2).

1.3.5.2 B. Practice

Exercise. B1: An L-shaped region

Let $E := ([0, 1] \times [0, 2]) \cup ([1, 2] \times [0, 1])$. Describe E as a region of the plane. Then use [Proposition 1.3.1](#) to write E as $([0, 2] \times [0, 1]) \cup (P \times Q)$ for intervals P and Q , and prove the equality.

Solution. The first rectangle is 1 wide and 2 high, standing on the origin; the second is 1 wide and 1 high, standing on $[1, 2]$. Together they form an “L”: a bottom strip $[0, 2] \times [0, 1]$ of width 2, with a square $[0, 1] \times [1, 2]$ on top of its left half. So we expect $P = [0, 1]$ and $Q = [1, 2]$.

Now the proof. Since $[0, 2] = [0, 1] \cup [1, 2]$, the first law of [Proposition 1.3.1](#) gives

$$[0, 1] \times [0, 2] = ([0, 1] \times [0, 1]) \cup ([0, 1] \times [1, 2]),$$

and the mirror-image $\cup$ law gives

$$[0, 2] \times [0, 1] = ([0, 1] \times [0, 1]) \cup ([1, 2] \times [0, 1]).$$

Substituting the first equation into the definition of E and regrouping the unions ($\cup$ is commutative and associative by [Theorem 1.2.1](#)),

$$\begin{aligned} E &= ([0, 1] \times [0, 1]) \cup ([0, 1] \times [1, 2]) \cup ([1, 2] \times [0, 1]) \\ &= ([0, 2] \times [0, 1]) \cup ([0, 1] \times [1, 2]), \end{aligned}$$

where the second line uses the mirror-image equation. So $P = [0, 1]$ and $Q = [1, 2]$. (The equality $[0, 2] = [0, 1] \cup [1, 2]$ is an order fact: $0 \leq x \leq 2$ holds exactly when $0 \leq x \leq 1$ or $1 \leq x \leq 2$.)

Exercise. B2: The intersection law, written with two pairs

Here is a proof of the inclusion ($\supseteq$) of $A \times (B \cap C) = (A \times B) \cap (A \times C)$ that works from the definition of the product and the characteristic property directly, without the shortcut (*). Its last sentence is blank. *Name the move, then complete it.*

Let $p \in (A \times B) \cap (A \times C)$. Since $p \in A \times B$, there are $a \in A$ and $b \in B$ with $p = (a, b)$. Since $p \in A \times C$, there are $a' \in A$ and $c \in C$ with $p = (a', c)$; the names a' and c are new, because nothing says that this second pair of coordinates is the first one again. Hence $(a, b) = (a', c)$. _____

Solution. The move is the second half of the section’s move, the characteristic property of pairs ([Definition 1.3.1](#)) in the direction ($\Rightarrow$). An equation between pairs splits into equations between coordinates.

The completed last sentence: *By the characteristic property, $a = a'$ and $b = c$; so $b \in B$ and $b = c \in C$, hence $b \in B \cap C$, and $p = (a, b) \in A \times (B \cap C)$.*

The names a' and c had to be fresh (a second $\exists$ may not reuse the names of the first), and the characteristic property is what shows, afterwards, that the two namings name the same coordinates. The proof of [Proposition 1.3.1](#) took the same step through (*), which packages exactly this use of the characteristic property.

Exercise. B3: Two products crossed

Prove that $(A \times B) \cap (B \times A) = (A \cap B) \times (A \cap B)$ for all sets A and B .

Solution. We prove two inclusions and conclude by [Proposition 1.1.1](#).

($\subseteq$) Let $p \in (A \times B) \cap (B \times A)$. Since $p \in A \times B$, $p = (x, y)$ with $x \in A$ and $y \in B$. Since also $(x, y) = p \in B \times A$, (*) gives $x \in B$ and $y \in A$. So $x \in A \cap B$ and $y \in A \cap B$, and $p = (x, y) \in (A \cap B) \times (A \cap B)$ by (*).

($\supseteq$) Let $p \in (A \cap B) \times (A \cap B)$, so $p = (x, y)$ with $x, y \in A \cap B$. Then $x \in A$ and $y \in B$, so $p \in A \times B$; and $x \in B$ and $y \in A$, so $p \in B \times A$. Hence $p \in (A \times B) \cap (B \times A)$.

1.3.5.3 C. Going deeper**Exercise. C1: The square of a union**

- (a) Determine whether $(A \cup B) \times (A \cup B) = (A \times A) \cup (B \times B)$ for all sets A and B . Justify your answer.
- (b) Prove the inclusion $(A \times A) \cup (B \times B) \subseteq (A \cup B) \times (A \cup B)$.
- (c) Prove, from [Proposition 1.3.1](#), that

$$(A \cup B) \times (A \cup B) = (A \times A) \cup (A \times B) \cup (B \times A) \cup (B \times B).$$

Solution.

- (a) No. Take $A = \{1\}$ and $B = \{2\}$. Then $(1, 2) \in \{1, 2\} \times \{1, 2\} = (A \cup B) \times (A \cup B)$. But $(A \times A) \cup (B \times B) = \{(1, 1), (2, 2)\}$, and $(1, 2)$ is neither $(1, 1)$ nor $(2, 2)$, since in each case the second (or first) coordinates differ. So the two sets are different.
- (b) Let $p \in (A \times A) \cup (B \times B)$. If $p \in A \times A$, then $p = (x, y)$ with $x, y \in A$; so $x, y \in A \cup B$, and $p \in (A \cup B) \times (A \cup B)$ by (*). If $p \in B \times B$, the same argument with B in place of A applies. In either case $p \in (A \cup B) \times (A \cup B)$.
- (c) The mirror-image $\cup$ law, with $A \cup B$ as the fixed factor on the right, gives

$$(A \cup B) \times (A \cup B) = (A \times (A \cup B)) \cup (B \times (A \cup B)).$$

The first $\cup$ law, applied to each of the two products, gives

$$A \times (A \cup B) = (A \times A) \cup (A \times B), \quad B \times (A \cup B) = (B \times A) \cup (B \times B).$$

Substituting, and using that $\cup$ is associative ([Theorem 1.2.1](#)), so that no brackets are needed, gives the formula. Part (a)'s counterexample is the term $A \times B$ at work: $(1, 2) \in A \times B$.

Exercise. C2: When one inclusion forces equality

Let A and B be sets. Prove that $A \times B \subseteq B \times A$ if and only if $A = B$ or $A = \emptyset$ or $B = \emptyset$.

Solution (Hint 1). For $(\Rightarrow)$, the empty cases are easy, so assume both sets have elements. To show $A \subseteq B$, you need to test an element $a \in A$; put it into a pair that is sure to lie in $A \times B$.

Solution (Hint 2). Fix $b_0 \in B$. For $a \in A$, the pair (a, b_0) lies in $A \times B$.

Solution. $(\Leftarrow)$ If $A = B$, then $A \times B = B \times A$, and in particular $A \times B \subseteq B \times A$. If $A = \emptyset$, then $A \times B = \emptyset$ by [Proposition 1.3.1](#), and $\emptyset$ is a subset of every set. If $B = \emptyset$, likewise $A \times B = A \times \emptyset = \emptyset$.

$(\Rightarrow)$ Suppose $A \times B \subseteq B \times A$. By excluded middle (Logic and Proof, Proposition 2.5.3), each of A and B either has an element or has none, and a set with no elements is $\emptyset$ ([Proposition 1.1.3](#)). If $A = \emptyset$ or $B = \emptyset$, we are done. In the remaining case both have elements: fix $a_0 \in A$ and $b_0 \in B$ ($\exists$ E twice, with fresh names).

$A \subseteq B$. Let $a \in A$. Then $(a, b_0) \in A \times B$ by $(*)$, so $(a, b_0) \in B \times A$ by the hypothesis. By $(*)$ again, $a \in B$ (and $b_0 \in A$).

$B \subseteq A$. Let $b \in B$. Then $(a_0, b) \in A \times B \subseteq B \times A$, and $(*)$ gives $b \in A$ (and $a_0 \in B$).

By [Proposition 1.1.1](#), $A = B$. This proves the equivalence.

Both witnesses are needed. With $B = \emptyset$ there is no b_0 , and indeed then $A \times B \subseteq B \times A$ holds for every A , including $A \neq B$: the inclusion $\{1\} \times \emptyset \subseteq \emptyset \times \{1\}$ is $\emptyset \subseteq \emptyset$.

Marking scheme. (10 points) - 2: $(\Leftarrow)$: all three cases, including that an empty factor makes $A \times B$ empty and hence a subset of anything. - 2: $(\Rightarrow)$: splits off the empty cases and, in the remaining case, names witnesses $a_0 \in A$ and $b_0 \in B$. - 2: $A \subseteq B$, through a pair (a, b_0) with an arbitrary $a \in A$. - 2: $B \subseteq A$, through a pair (a_0, b) with an arbitrary $b \in B$. - 2: uses the characteristic property (directly or through $(*)$) to read the coordinates off a pair of $B \times A$, and concludes $A = B$ by two inclusions.

Fatal errors (0 for the step): testing inclusion on particular elements only (entry (N1) of *Logic and Proof*, Chapter 5’s catalogue); using b_0 without having shown that B has an element; concluding $A = B$ from $A \times B = B \times A$ “by canceling” without an element argument. A proof of $(\Rightarrow)$ by contraposition (assume $A \neq B$ and both nonempty, and exhibit a pair of $A \times B$ outside $B \times A$) earns full marks if it finds the pair: 2 points for the witnesses and 4 for the pair and the check that it is outside $B \times A$, in place of the two inclusions.

1.4 Functions, Composition, Injections and Surjections

Section 1.3 made “the pairs a description picks out” into a set: a relation from A to B is a subset of $A \times B$. Among its examples, the graph of squaring relates each real number to exactly one number, and the unit circle relates 0 to two numbers and 2 to none. This section gives the first kind its name, **function**, and builds the rest on it: we restrict functions, use them to make families, sequences and the set Y^X precise, compose them, and ask the two questions that the

rest of the chapter turns on: does a function send different inputs to different outputs, and does it reach every point of its target?

1.4.1 What is a function?

Everyone has used functions since school, and the everyday definition is “a rule that assigns to each input one output”. Look at the word “rule”. The rules $x \mapsto x^2$ and $x \mapsto (x + 1)^2 - 2x - 1$ are different: one squares, the other adds, squares, subtracts, and subtracts again. Yet on every real number they give the same answer, and nobody would call them different functions. Other functions have no rule at all that one could write down. A table of measured temperatures, one per day, assigns to each day one number, and no formula produces it.

So the rule is not what matters. What matters is **which output goes with which input**: the set of input–output pairs. That set is a relation, and Section 1.3 said exactly when two relations are the same, namely when they contain the same pairs. So we forget the rule and keep the pairs. One more piece of data is needed, the set in which the outputs are **declared** to lie; the reason is given after the definition.

A function is its graph, with a declared target.

Definition 1.4.1. Function

Let X and Y be sets. A **function from X to Y** , written $f: X \rightarrow Y$, is given by a relation $G \subseteq X \times Y$, the **graph** of f , such that

$$\text{for every } x \in X \text{ there is exactly one } y \in Y \text{ with } (x, y) \in G,$$

together with the set Y , the **codomain** of f . Formally, f is the pair (G, Y) . The set X is the **domain** of f . For $x \in X$, the unique y with $(x, y) \in G$ is written $f(x)$, the **value** of f at x . We write $(x, y) \in f$ for $(x, y) \in G$, and $x \mapsto f(x)$ for the whole function when its domain and codomain are clear.

In symbols, the condition on the graph is

$$\forall x \in X \quad \exists! y \in Y \quad (x, y) \in G,$$

where, as with the bounded quantifiers of Logic and Proof, Definition 5.1.1, $\exists! y \in Y \quad \psi$ abbreviates $\exists! y (y \in Y \wedge \psi)$.

The whole definition is the quantifier pattern $\forall\exists!$. By Logic and Proof, Definition 5.8.1, “ $\exists! y$ ” is an existence claim and a uniqueness claim, and by Logic and Proof, Proposition 5.8.1 the uniqueness claim may be stated without reference to the witness; so the condition on the graph is two clauses:

- **(Fn1)** (existence) for every $x \in X$ there is **at least one** $y \in Y$ with $(x, y) \in G$: every input has an output;
- **(Fn2)** (uniqueness) for every $x \in X$ and all $y, y' \in Y$, if $(x, y) \in G$ and $(x, y') \in G$, then $y = y'$: no input has two outputs.

In words: the graph is a set of input–output pairs, with inputs from X and outputs from Y ; (Fn1) says no input is left without an output, and (Fn2) says no input is given two. The codomain Y is where outputs are allowed to lie. It need not consist of outputs: $x \mapsto x^2$ can be declared as a function from $\mathbb{R}$ to $\mathbb{R}$, although no output is negative. The set of outputs that actually occur is

the **range** of f ,

$$\{y \in Y : \exists x \in X \ f(x) = y\},$$

which Section 1.6 will call the image of X and write $f(X)$. The domain needs no slot of its own in the pair (G, Y) : by (Fn1) every $x \in X$ occurs as a first coordinate in G , and since $G \subseteq X \times Y$, nothing else does. So the graph determines the domain.

The notation $f(x)$ is well defined. Writing “the y with $(x, y) \in f$ ” is legitimate only if there is one, and only one. Clause (Fn1) supplies one, and clause (Fn2) says that any two are equal; together they license the article “the”. This gives the equivalence that every argument with functions uses:

$$(x, y) \in f \iff y = f(x) \quad (x \in X, y \in Y). \quad (**)$$

For $(\Leftarrow)$: $(x, f(x)) \in f$ by the choice of $f(x)$, and if $y = f(x)$ then $(x, y) \in f$ by substituting equals ($=$ E, Logic and Proof, Definition 5.7.1). For $(\Rightarrow)$: both (x, y) and $(x, f(x))$ lie in f , so $y = f(x)$ by (Fn2). Two consequences follow at once. Equal inputs give equal outputs: if $x_1 = x_2$, then $f(x_1) = f(x_2)$. And every value lies in the codomain: $f(x) \in Y$ for every $x \in X$.

Functions given by formulas. Most functions in this book are introduced by a formula: “let $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2$ ”. This means the function whose graph is $\{(x, y) \in \mathbb{R}^2 : y = x^2\}$. Clause (Fn2) holds for any such graph: if $y = x^2$ and $y' = x^2$, then $y = y'$. So for a formula only (Fn1) needs checking, and it says: for each x in the domain, the formula **makes sense** and its value **lies in the codomain**. For x^2 on $\mathbb{R}$ both are clear. For “ $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 1/x$ ” the formula makes no sense at $x = 0$, and (Fn1) fails there; for “ $f: \mathbb{R} \rightarrow [0, \infty)$, $f(x) = x^2$ ” the value must be checked to lie in $[0, \infty)$, which it does, since squares are nonnegative.

A function whose domain is a product $A \times B$ is a **function of two variables**. Its inputs are pairs, and we write $f(a, b)$ rather than $f((a, b))$. Addition $(m, n) \mapsto m + n$ from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ is an example.

Example. First functions

Check (Fn1) and (Fn2), and find the range, for each of the following.

- (a) $f: \{1, 2, 3\} \rightarrow \{a, b\}$ with graph $\{(1, a), (2, b), (3, a)\}$.
- (b) The **identity function** $\text{id}_X: X \rightarrow X$, $\text{id}_X(x) = x$, for a set X .
- (c) A **constant function** $X \rightarrow Y$, $x \mapsto c$, for a fixed $c \in Y$.
- (d) The **indicator function** $\mathbf{1}_A: X \rightarrow \{0, 1\}$ of a subset $A \subseteq X$:

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \in X \setminus A. \end{cases}$$

Solution.

- (a) The graph is a subset of $\{1, 2, 3\} \times \{a, b\}$. Each of 1, 2, 3 occurs as a first coordinate (Fn1), and each occurs exactly once, so no input has two outputs (Fn2). So f is a function, with $f(1) = a$, $f(2) = b$, $f(3) = a$. Its range is $\{a, b\}$, the whole codomain. A function on a small finite set is often given this way, by listing its values; no formula is needed.
- (b) The graph is the **diagonal** $\{(x, x) : x \in X\}$. For each x , the pair (x, x) is in it (Fn1); and if (x, y) is in it, then $(x, y) = (z, z)$ for some $z \in X$, and the characteristic property (Definition 1.3.1) gives $x = z = y$ (Fn2). The range is X .

- (c) The graph is $X \times \{c\}$. For each $x \in X$, (x, c) is in it (Fn1); if (x, y) is in it, then $y \in \{c\}$ by the membership test (*) for pairs in a product (Definition 1.3.2), so $y = c$ (Fn2). If X has an element, the range is $\{c\}$. If $X = \emptyset$, the range is $\emptyset$.
- (d) The graph is $(A \times \{1\}) \cup ((X \setminus A) \times \{0\})$. For (Fn1), let $x \in X$. By excluded middle (Logic and Proof, Proposition 2.5.3), $x \in A$ or $x \notin A$; in the first case $(x, 1)$ is in the graph, in the second $x \in X \setminus A$ and $(x, 0)$ is. For (Fn2), the two cases cannot both hold, so x never receives both 0 and 1, and each case gives only one pair with first coordinate x . The range is $\{0, 1\}$ if A has an element and $X \setminus A$ has an element; it is $\{0\}$ if $A = \emptyset$ and X has an element, $\{1\}$ if $A = X$ and X has an element, and $\emptyset$ if $X = \emptyset$. This is the first function of the chapter whose existence clause needs excluded middle: the definition by cases is a function only because every x falls into one of the cases.

The next example collects relations that fail to be functions. Each failure is one clause, and it pays to say which.

Example. Relations that are not functions

- (a) Is the unit circle $S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ the graph of a function $\mathbb{R} \rightarrow \mathbb{R}$?
- (b) “Send a fraction p/q to its numerator p .” Is this a function from $\mathbb{Q}$ to $\mathbb{Z}$? As a relation it is

$$N := \{(r, p) \in \mathbb{Q} \times \mathbb{Z} : \exists q \in \mathbb{Z}_{>0} \ r = p/q\}.$$

Solution.

- (a) No, and both clauses fail, at different points. Clause (Fn2) fails at $x = 0$: both $(0, 1)$ and $(0, -1)$ lie in S , and $1 \neq -1$. Clause (Fn1) fails at $x = 2$: a pair $(2, y) \in S$ would need $y^2 = 1 - 4 = -3$, and no real square is negative. Restricting the inputs to $[-1, 1]$ repairs (Fn1) but not (Fn2); keeping only the upper half, $y \geq 0$, repairs (Fn2) as well, and gives the function $[-1, 1] \rightarrow \mathbb{R}, x \mapsto \sqrt{1 - x^2}$.
- (b) No. Clause (Fn1) holds: every rational number is p/q for some $p \in \mathbb{Z}$ and $q \in \mathbb{Z}_{>0}$. But (Fn2) fails, because a rational number has many ways of being written as a fraction. Since $\frac{1}{2} = \frac{2}{4}$, both $(\frac{1}{2}, 1)$ and $(\frac{1}{2}, 2)$ lie in N , and $1 \neq 2$. The “rule” gives an output that depends on **how the input is written**, not on the input. A function may not do that, since its value is attached to the input itself. This is why definitions of the form “write x as ..., then send it to ...” must be checked for **well-definedness**, a check that returns in Chapter 2 for functions defined on classes. Here the repair is to fix the way of writing: let q_0 be the least $q \in \mathbb{Z}_{>0}$ with $rq \in \mathbb{Z}$ (such q exist, by the first sentence, so there is a least one by well-ordering, proved in Chapter 2; a least element is unique), and send r to rq_0 . This is a function.

Why this definition. Each clause has a job, and dropping one lets in the failures of Example (Relations that are not functions). Without (Fn1), a “function” could be undefined at some inputs, like $1/x$ at 0; then $f(x)$ would not name anything, and statements “for every $x \in X$ ” about $f(x)$ would lose their meaning. Without (Fn2), $f(x)$ would not name **one** thing, and $f(\frac{1}{2})$ in part (b) would be both 1 and 2. The graph replaces the rule because two rules can describe one function, and sets are compared by their elements.

Why is the codomain part of the data? A relation is a set, so if a function were only its

graph, two functions with the same graph would be equal (Definition 1.1.1), whatever target was declared. The graph of $x \mapsto x^2$ as a function $[0, \infty) \rightarrow \mathbb{R}$ is the same set as its graph as a function $[0, \infty) \rightarrow [0, \infty)$. But these should be different functions, because the second reaches every point of its target and the first does not: this is the property called surjectivity below, and an example later in this section shows it changing with the codomain. Recording the codomain in the pair (G, Y) keeps the two apart. Many books take the same convention, some without saying so.

Warning. A formula is not yet a function. “The function $1/x$ ” names no function until a domain and a codomain are given, and the domain must be chosen so that (Fn1) holds: $x \mapsto 1/x$ is a function $\mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}$, and there is no function $\mathbb{R} \rightarrow \mathbb{R}$ given by this formula. Likewise $\sqrt{x}$ makes sense on $[0, \infty)$ and not on $\mathbb{R}$. When you meet “let $f(x) = \dots$ ”, ask first where x lives and where $f(x)$ is meant to lie.

Quick check. Which of these relations on $\{1, 2, 3\}$ are graphs of functions $\{1, 2, 3\} \rightarrow \{1, 2, 3\}$: $\{(1, 2), (2, 2), (3, 2)\}$, $\{(1, 1), (2, 2)\}$, $\{(1, 3), (2, 1), (3, 2), (2, 3)\}$? For each that is not, name the clause that fails.

Solution. The first is a function: the constant function with value 2. The second fails (Fn1) at 3, which has no output. The third fails (Fn2) at 2, which has the two outputs 1 and 3.

1.4.2 When are two functions equal?

A function is a pair (G, Y) , so two functions are equal when their graphs are equal and their codomains are equal. In practice we compare functions by their values, and the next result says that this is the same test.

Proposition 1.4.1. Equality of Functions

Let $f: X \rightarrow Y$ and $g: X' \rightarrow Y'$ be functions. Then $f = g$ if and only if $X = X'$, $Y = Y'$, and $f(x) = g(x)$ for every $x \in X$.

Idea. Write $f = (G, Y)$ and $g = (G', Y')$. The characteristic property of pairs splits $f = g$ into “same graph” and “same codomain”. Two graphs are equal when they have the same elements, and by (**) an element of a graph is a pair $(x, f(x))$; so “same graph” is “same inputs, same outputs”.

Proof. Write $f = (G, Y)$ and $g = (G', Y')$, as in Definition 1.4.1.

($\Rightarrow$) Suppose $f = g$. By the characteristic property of pairs (Definition 1.3.1), $G = G'$ and $Y = Y'$. The domain is determined by the graph, as the set of first coordinates of its elements, so $X = X'$. Let $x \in X$. By (**), $(x, f(x)) \in G = G'$, so $(x, f(x)) \in g$, and (**) for g gives $f(x) = g(x)$.

($\Leftarrow$) Suppose $X = X'$, $Y = Y'$, and $f(x) = g(x)$ for every $x \in X$. We show $G \subseteq G'$. Let $p \in G$. Since $G \subseteq X \times Y$, $p = (x, y)$ with $x \in X$ and $y \in Y$. By (**), $y = f(x)$, so $y = g(x)$ by the hypothesis, and then $p = (x, y) \in G'$ by (**) for g . The same argument with f and g exchanged gives $G' \subseteq G$, so $G = G'$ by Proposition 1.1.1. Since also $Y = Y'$, the characteristic property gives $(G, Y) = (G', Y')$, that is, $f = g$. ■

So to show two functions equal, check three things: the same domain, the same codomain, and the same value at an arbitrary point. To show them different, one witness suffices: a point where the values differ, or a difference in domain or codomain.

Example. Equal and unequal functions

Decide which of these functions are equal.

- (a) $f, g: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = (x + 1)^2 - 2x - 1$ and $g(x) = x^2$.
- (b) $h, k: \mathbb{R} \rightarrow \mathbb{R}$, $h(x) = |x|$ and $k(x) = \sqrt{x^2}$.
- (c) $s_1: [0, \infty) \rightarrow \mathbb{R}$ and $s_2: [0, \infty) \rightarrow [0, \infty)$, both given by $x \mapsto x^2$.

Solution.

- (a) Equal. The domains are both $\mathbb{R}$ and the codomains both $\mathbb{R}$. For every real x , $(x + 1)^2 - 2x - 1 = x^2 + 2x + 1 - 2x - 1 = x^2$. By Proposition 1.4.1, $f = g$: two rules, one function.
- (b) Equal. Same domain and codomain, and for every real x , $\sqrt{x^2} = |x|$: the number $|x|$ is nonnegative and its square is x^2 , and $\sqrt{x^2}$ is by definition the nonnegative number with square x^2 , which is unique by Logic and Proof, Theorem 2.4.5.
- (c) Not equal. They have the same domain, the same values, and even the same graph; but their codomains $\mathbb{R}$ and $[0, \infty)$ are different sets (-1 is in one and not the other), so by Proposition 1.4.1 $s_1 \neq s_2$.

Restriction. A function can be cut down to a smaller domain. For $f: X \rightarrow Y$ and $A \subseteq X$, the **restriction** of f to A is the function

$$f|_A: A \rightarrow Y, \quad f|_A(a) := f(a),$$

whose graph is $G \cap (A \times Y)$: the pairs of f whose input lies in A . The squaring function $\mathbb{R} \rightarrow \mathbb{R}$ restricted to $[0, \infty)$ is s_1 of part (c). The bar in $f|_A$ is one of the uses of a vertical bar that the book's colon-for-set-builder convention keeps free.

1.4.3 Families, sequences, and the set of all functions

Section 1.2 used families of sets $(A_i)_{i \in I}$ and described them informally, as “a set A_i for each index $i \in I$ ”. We can now say exactly what a family is, once its members are known to lie in some set. A **family of subsets of X indexed by I** is a function $I \rightarrow \mathcal{P}(X)$, $i \mapsto A_i$, with the power set $\mathcal{P}(X)$ as the codomain the definition asks for. More generally, if S is a set whose elements are sets, a family of members of S indexed by I is a function $I \rightarrow S$. A codomain

must be named, because there is no set of all sets ([Theorem 1.1.4](#)). The value of a family at i is written A_i rather than $A(i)$, and the family is written $(A_i)_{i \in I}$, in parentheses because it is indexed. A family is not the same thing as its set of values $\{A_i : i \in I\}$: the family $(A_i)_{i \in \{1,2\}}$ with $A_1 = A_2 = \{0\}$ has two indices and one value.

In the same way, a **sequence** of real numbers $(a_n)_{n \geq 1}$ is a function $\mathbb{Z}_{>0} \rightarrow \mathbb{R}$, $n \mapsto a_n$, with the value at n written a_n . The sequence $(1/n)$ is the function $n \mapsto 1/n$, and its set of values is $\{1/n : n \geq 1\}$. Sequences are the main object of *Analysis on the Real Line*.

Finally, all functions from X to Y form a set, written

$$Y^X := \{f \in \mathcal{P}(X \times Y) \times \{Y\} : f \text{ is a function from } X \text{ to } Y\}.$$

It is a set by the working rules of Section 1.1: every function $X \rightarrow Y$ is a pair (G, Y) with $G \in \mathcal{P}(X \times Y)$, so Y^X is the subset of $\mathcal{P}(X \times Y) \times \{Y\}$ cut out by the condition of [Definition 1.4.1](#), and this product is a set by (W2), (W4), and (W5), so Y^X is a set by specification (W1). The letters are in this order because, for finite sets, Y^X has m^n elements when Y has m and X has n , as the next example suggests; counting is Chapter 3's business. The set $\{0, 1\}^{\mathbb{N}}$ of all sequences of zeros and ones, indexed from 0, plays a leading role in Chapter 3.

Example. All functions between two small sets

List the elements of $\{0, 1\}^{\{1,2\}}$, and say which are indicator functions of subsets of $\{1, 2\}$.

Solution. A function $f: \{1, 2\} \rightarrow \{0, 1\}$ is determined by its two values $f(1)$ and $f(2)$ ([Proposition 1.4.1](#)), and each may be 0 or 1. So there are four:

$f(1)$	$f(2)$	graph	indicator of
0	0	$\{(1, 0), (2, 0)\}$	$\emptyset$
0	1	$\{(1, 0), (2, 1)\}$	$\{2\}$
1	0	$\{(1, 1), (2, 0)\}$	$\{1\}$
1	1	$\{(1, 1), (2, 1)\}$	$\{1, 2\}$

Every one of them is an indicator function: $f = \mathbf{1}_A$ for $A = \{x \in \{1, 2\} : f(x) = 1\}$. The four subsets of $\{1, 2\}$ and the four functions correspond one to one.

The correspondence at the end of the example, between subsets of X and functions $X \rightarrow \{0, 1\}$, holds for every set X : send f to $\{x \in X : f(x) = 1\}$ and A to $\mathbf{1}_A$. So counting the subsets of a finite set and counting the functions from it to $\{0, 1\}$ are one problem, which Chapter 3 solves.

The empty set gives the degenerate cases, and they are worth settling now, since they are the test cases of many later statements.

Proposition 1.4.2. Functions and the Empty Set

- (a) For every set Y there is exactly one function from $\emptyset$ to Y : the pair $(\emptyset, Y)$, called the **empty function**.
- (b) If X is nonempty (it has an element), there is no function from X to $\emptyset$.

Proof.

- (a) *Existence.* The empty relation satisfies $\emptyset \subseteq \emptyset \times Y$, since $\emptyset$ is a subset of every set (Section 1.1). The condition of [Definition 1.4.1](#) reads “for every x , if $x \in \emptyset$, then there is exactly one $y \in Y$ with $(x, y) \in \emptyset$ ”. For each x the hypothesis $x \in \emptyset$ is false, so the implication holds (Logic and Proof, Proposition 2.1.2). So $(\emptyset, Y)$ is a function from $\emptyset$ to Y .
- Uniqueness.* Let (G, Y) be any function from $\emptyset$ to Y . Its graph satisfies $G \subseteq \emptyset \times Y = \emptyset$, by the mirror-image $\emptyset$ law of [Proposition 1.3.1](#). So an element of G would be an element of $\emptyset$; G has no elements, and $G = \emptyset$ by [Proposition 1.1.3](#). Hence $(G, Y) = (\emptyset, Y)$.
- (b) Let $x_0 \in X$, and suppose $f: X \rightarrow \emptyset$ were a function. By (Fn1) there would be $y \in \emptyset$ with $(x_0, y) \in f$; but $\emptyset$ has no elements. So there is no such function $[-I]$. ■

In the notation Y^X : the set $Y^\emptyset$ has exactly one element for every Y , including $Y = \emptyset$; and $\emptyset^X = \emptyset$ whenever X is nonempty. The empty function $\emptyset \rightarrow Y$ returns in Section 1.5 as the witness that a hypothesis there cannot be dropped.

1.4.4 Composition

Two functions can be applied one after the other when the outputs of the first are inputs of the second. Temperature in degrees Celsius, as a function of the day, followed by the conversion to degrees Fahrenheit, gives temperature in Fahrenheit as a function of the day.

Definition 1.4.2. Composition

Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be functions, the codomain of f being the domain of g . Their **composite** is the function

$$g \circ f: X \rightarrow Z, \quad (g \circ f)(x) := g(f(x)),$$

read “ g after f ”.

This is a function by the remark on formulas: for $x \in X$, the value $f(x)$ lies in Y , the domain of g , so $g(f(x))$ makes sense, and it lies in Z . Its graph is $\{(x, z) \in X \times Z : z = g(f(x))\}$. The order of the letters is the order of the formula $g(f(x))$, which is the **reverse** of the order in which the functions act: f acts first.

Example. The order of composition matters

Let $f, g: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x + 1$ and $g(x) = 2x$. Compute $g \circ f$ and $f \circ g$, and decide whether they are equal.

Solution. By [Definition 1.4.2](#), $(g \circ f)(x) = g(x + 1) = 2x + 2$ and $(f \circ g)(x) = f(2x) = 2x + 1$. At $x = 0$ the values are 2 and 1, so by [Proposition 1.4.1](#) $g \circ f \neq f \circ g$. In words: “add one, then double” is not “double, then add one”. (In fact the two composites differ at every x , since $2x + 2 \neq 2x + 1$; one point already suffices.)

Composition can also fail to be defined in one order only. For $f: \{1, 2\} \rightarrow \mathbb{R}$ and $g: \mathbb{R} \rightarrow \mathbb{R}$, the composite $g \circ f: \{1, 2\} \rightarrow \mathbb{R}$ is defined, and $f \circ g$ is not: g 's outputs are real numbers, and f accepts only 1 and 2.

Composition is not commutative, but it is associative, and the identity functions do nothing.

Proposition 1.4.3. Composition Is Associative

Let $f: X \rightarrow Y$, $g: Y \rightarrow Z$, and $h: Z \rightarrow W$ be functions. Then

$$h \circ (g \circ f) = (h \circ g) \circ f,$$

and $f \circ \text{id}_X = f = \text{id}_Y \circ f$.

Proof. Both sides of the first equation are functions from X to W , by Definition 1.4.2. For $x \in X$, applying Definition 1.4.2 twice on each side gives

$$(h \circ (g \circ f))(x) = h((g \circ f)(x)) = h(g(f(x))) = (h \circ g)(f(x)) = ((h \circ g) \circ f)(x).$$

So the two functions agree at every point, and they are equal by Proposition 1.4.1. Likewise $f \circ \text{id}_X$ and $\text{id}_Y \circ f$ are functions from X to Y with values $f(\text{id}_X(x)) = f(x)$ and $\text{id}_Y(f(x)) = f(x)$, so both equal f . ■

Because of associativity, we write $h \circ g \circ f$ without brackets: every way of bracketing gives the same function, $x \mapsto h(g(f(x)))$.

Quick check. For $f, g: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2$ and $g(x) = x - 3$, compute $(g \circ f)(2)$ and $(f \circ g)(2)$.

Solution. $(g \circ f)(2) = g(4) = 1$ and $(f \circ g)(2) = f(-1) = 1$. The values agree at this point, but the functions are different: at $x = 0$, $(g \circ f)(0) = -3$ and $(f \circ g)(0) = 9$. Agreement at one point proves nothing about equality; disagreement at one point proves inequality.

1.4.5 Injective, surjective, bijective

Two questions can be asked of any function $f: X \rightarrow Y$, and much of mathematics turns on them. First: can we recover the input from the output, or can two different inputs give the same output? Second: is every point of the codomain an output? The first question is about how f treats X ; the second is about how it covers Y .

1.4.5.1 Injective functions

An injective function never sends two different inputs to the same output.

Definition 1.4.3. Injective

A function $f: X \rightarrow Y$ is **injective** (an **injection**, or **one-to-one**) if, for all $x_1, x_2 \in X$,

$f(x_1) = f(x_2)$ implies $x_1 = x_2$. In symbols:

$$\forall x_1 \in X \quad \forall x_2 \in X \quad (f(x_1) = f(x_2) \rightarrow x_1 = x_2).$$

In words: if two inputs give the same output, they were the same input. By Logic and Proof, Proposition 2.3.2, the definition gives the **contrapositive form**

$$\forall x_1 \in X \quad \forall x_2 \in X \quad (x_1 \neq x_2 \rightarrow f(x_1) \neq f(x_2)) :$$

different inputs give different outputs. The two forms are equivalent; the way back, from the contrapositive form to the definition, is Logic and Proof, Proposition 2.5.2, which uses RAA. A proof of injectivity usually uses the form of the definition, because an equation $f(x_1) = f(x_2)$ is something to compute with, while an inequation $x_1 \neq x_2$ is not.

The negation is found by the laws of *Logic and Proof*, Chapter 4 for negating bounded quantifiers and implications: “ f is **not** injective” is

$$\exists x_1 \in X \quad \exists x_2 \in X \quad (f(x_1) = f(x_2) \wedge x_1 \neq x_2).$$

So a failure of injectivity is shown by **two witnesses**: two different inputs with the same output.

1.4.5.2 Surjective functions

A surjective function reaches every point of its codomain.

Definition 1.4.4. Surjective

A function $f: X \rightarrow Y$ is **surjective** (a **surjection**, or **onto**) if, **for every** $y \in Y$, **there exists** $x \in X$ with $f(x) = y$. In symbols:

$$\forall y \in Y \quad \exists x \in X \quad f(x) = y.$$

In words: every point of the codomain is an output; equivalently, the range of f is all of Y . An x with $f(x) = y$ is called a **preimage** of y . The definition does not ask for the preimage to be unique. Negating as before, “ f is **not** surjective” is

$$\exists y \in Y \quad \forall x \in X \quad f(x) \neq y,$$

so a failure of surjectivity is shown by **one witness**: a point of the codomain that no input reaches.

Surjectivity depends on the codomain, which is why [Definition 1.4.1](#) records it: s_1 and s_2 of [Example \(Equal and unequal functions\)](#) have the same graph, and we shall see that s_2 is surjective and s_1 is not.

Warning. Surjectivity cannot be checked by trying some y . The function $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2 + x$, reaches $0 = f(0)$, $2 = f(1)$, $6 = f(2)$, and $12 = f(3)$, and a reader who tries these may believe it reaches everything. It does not reach -1 : $x^2 + x = -1$ would mean $x^2 + x + 1 = 0$, but $x^2 + x + 1 = (x + \frac{1}{2})^2 + \frac{3}{4} > 0$ for every real x . Concluding “for every y ” from a few instances is entry (N1) of the catalogue in *Logic and Proof*, Chapter 5.

1.4.5.3 Bijective functions

Definition 1.4.5. Bijective

A function is **bijective** (a **bijection**) if it is both injective and surjective.

A bijection is a function in which every point of the codomain is reached **exactly once**: for every $y \in Y$ there is exactly one $x \in X$ with $f(x) = y$. This is the $\exists!$ of Definition 1.4.1 again, now read from the codomain side, and its two halves are the two properties. Existence of such an x , for every y , is surjectivity. Uniqueness is injectivity: if $f(x_1) = y = f(x_2)$, injectivity gives $x_1 = x_2$; and conversely, if every y has at most one preimage, then $f(x_1) = f(x_2)$ makes x_1 and x_2 two preimages of the same point $y := f(x_1)$, so $x_1 = x_2$.

1.4.5.4 Worked examples

Proofs of these properties have a fixed shape, read off the definitions.

- To prove f **injective**: “Let $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$.” Compute until you reach $x_1 = x_2$.
- To prove f **surjective**: “Let $y \in Y$.” Then produce an x , **say what it is**, check that it lies in X , and check that $f(x) = y$. The x is usually found by solving $f(x) = y$ for x on scrap paper; the proof then states the solution first and checks it, as every existence proof does.
- To prove f **not injective**: name $x_1 \neq x_2$ with $f(x_1) = f(x_2)$. **Not surjective**: name $y \in Y$ and show that $f(x) = y$ has no solution $x \in X$.

Example. Deciding injective and surjective

For each function below, decide with proof whether it is injective and whether it is surjective.

- $f: \mathbb{R} \rightarrow \mathbb{R}, f(x) = 3x - 5$.
- $g: \mathbb{Z} \rightarrow \mathbb{Z}, g(n) = 3n$.
- $h: \{1, 2, 3\} \rightarrow \{a, b\}$, with $h(1) = a, h(2) = b, h(3) = a$.
- $k: \{1, 2\} \rightarrow \{a, b, c\}$, with $k(1) = a, k(2) = c$.

Solution.

- Injective.* Let $x_1, x_2 \in \mathbb{R}$ with $3x_1 - 5 = 3x_2 - 5$. Adding 5 and dividing by 3 gives $x_1 = x_2$.
Surjective. On scrap paper, $3x - 5 = y$ gives $x = (y + 5)/3$. Now the proof: let $y \in \mathbb{R}$, and put $x := (y + 5)/3$. Then $x \in \mathbb{R}$, and $f(x) = 3 \cdot \frac{y+5}{3} - 5 = y$. So f is bijective.
- Injective.* Let $n_1, n_2 \in \mathbb{Z}$ with $3n_1 = 3n_2$. Dividing by 3 gives $n_1 = n_2$.
Not surjective. The witness is $1 \in \mathbb{Z}$. If $n \geq 1$, then $3n \geq 3$; if $n \leq 0$, then $3n \leq 0$. So $3n \neq 1$ for every integer n , and 1 has no preimage. (Solving $3n = 1$ gives $n = \frac{1}{3}$, which is not in the domain $\mathbb{Z}$: the same formula on $\mathbb{Q}$ would be surjective.)
- Not injective:* $h(1) = a = h(3)$ and $1 \neq 3$. *Surjective:* the codomain has two points, and $a = h(1), b = h(2)$. For a finite codomain, checking every point **is** a proof: the instances are all the cases.
- Injective:* the two inputs have the different outputs $a \neq c$, and there are no other pairs of different inputs to check. *Not surjective:* b is not an output, since the only outputs are $k(1) = a$ and $k(2) = c$.

Parts (c) and (d) show two more things. A function from a three-element set to a two-element set cannot be injective, and one from a two-element set to a three-element set cannot be surjective: there are too many inputs in the first case and too few in the second. Chapter 3 proves facts of this kind in general, starting from the pigeonhole principle. And the finite pictures are worth drawing: two columns of dots, with an arrow from each input to its output. Injective means no dot on the right receives two arrows; surjective means every dot on the right receives at least one.

The key example of the section is one formula read as three different functions.

Example. Squaring, three ways

For each of the three functions, settle both questions: injective or not, surjective or not.

- (a) $q_1: \mathbb{R} \rightarrow \mathbb{R}, q_1(x) = x^2$.
- (b) $q_2: [0, \infty) \rightarrow \mathbb{R}, q_2(x) = x^2$.
- (c) $q_3: [0, \infty) \rightarrow [0, \infty), q_3(x) = x^2$.

Solution.

- (a) *Not injective:* $q_1(1) = 1 = q_1(-1)$ and $1 \neq -1$. *Not surjective:* the witness is $-1 \in \mathbb{R}$, since $x^2 \geq 0 > -1$ for every real x . So q_1 is neither.
- (b) *Injective.* Let $x_1, x_2 \geq 0$ with $x_1^2 = x_2^2$. Then $x_1^2 \leq x_2^2$, so $x_1 \leq x_2$ by Logic and Proof, Theorem 2.4.5, which applies since both numbers are nonnegative; and likewise $x_2 \leq x_1$. Hence $x_1 = x_2$. *Not surjective:* the witness -1 still works, for the same reason.
- (c) *Injective*, by the same proof as in (b): injectivity concerns only the domain and the values, and these are the same as for q_2 . *Surjective.* Let $y \in [0, \infty)$. A nonnegative real number has a nonnegative square root, so put $x := \sqrt{y}$. Then $x \in [0, \infty)$ and $q_3(x) = (\sqrt{y})^2 = y$. So q_3 is bijective.

The same rule gives three different functions. Shrinking the domain from $\mathbb{R}$ to $[0, \infty)$ removed the two witnesses ± 1 against injectivity. Shrinking the codomain to $[0, \infty)$ removed the witness -1 against surjectivity. The graphs of q_2 and q_3 are the same set; only the declared codomain differs, and surjectivity changes with it.

Warning. Injectivity is not the defining property of a function read backwards. Every function satisfies “if $x_1 = x_2$, then $f(x_1) = f(x_2)$ ”: equal inputs give equal outputs, which is part of $f(x)$ being well defined. Injectivity is the **converse**, “if $f(x_1) = f(x_2)$, then $x_1 = x_2$ ”. A “proof” of injectivity that starts from $x_1 = x_2$ and reaches $f(x_1) = f(x_2)$ has proved the wrong implication, entry (N5) of *Logic and Proof*, Chapter 5’s catalogue, and it “proves” q_1 injective, which is false.

Quick check. Is $n \mapsto n + 1$ surjective as a function $\mathbb{N} \rightarrow \mathbb{N}$? As a function $\mathbb{Z} \rightarrow \mathbb{Z}$?

Solution. On $\mathbb{N}$: no. The witness is 0, since $n + 1 \geq 1$ for every $n \in \mathbb{N}$, so $n + 1 = 0$ has no solution in $\mathbb{N}$. On $\mathbb{Z}$: yes. For $m \in \mathbb{Z}$, the integer $m - 1$ satisfies $(m - 1) + 1 = m$. Both functions are injective, since $n_1 + 1 = n_2 + 1$ gives $n_1 = n_2$.

1.4.6 Composites of injections and surjections

Both properties survive composition. This is the first theorem about functions whose proof uses the definitions and nothing else, and it shows the two proof shapes of the last subsection working together.

Theorem 1.4.4. Composites of Injections and Surjections

Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be functions.

- (a) If f and g are injective, then $g \circ f$ is injective.
- (b) If f and g are surjective, then $g \circ f$ is surjective.
- (c) If f and g are bijective, then $g \circ f$ is bijective.

Idea. For (a), start from equal outputs of $g \circ f$. They are equal outputs of g , at the points $f(x_1)$ and $f(x_2)$; peel off g by its injectivity, then f by its. For (b), work backwards along the chain $X \rightarrow Y \rightarrow Z$: to reach $z \in Z$, first find a point of Y that g sends to z , then a point of X that f sends there. Part (c) is (a) and (b) together.

Proof.

- (a) Let $x_1, x_2 \in X$ with $(g \circ f)(x_1) = (g \circ f)(x_2)$, that is, $g(f(x_1)) = g(f(x_2))$ by [Definition 1.4.2](#). The points $f(x_1)$ and $f(x_2)$ lie in Y , and g is injective, so $f(x_1) = f(x_2)$ [$\forall E$, $\rightarrow E$: injectivity of g at $f(x_1), f(x_2)$]. Since f is injective, $x_1 = x_2$ [$\forall E$, $\rightarrow E$: injectivity of f at x_1, x_2]. So $g \circ f$ is injective.
- (b) Let $z \in Z$. Since g is surjective, there is $y \in Y$ with $g(y) = z$ [$\exists E$: the name y is fresh]. Since f is surjective, there is $x \in X$ with $f(x) = y$ [$\exists E$: x is fresh]. Then

$$(g \circ f)(x) = g(f(x)) = g(y) = z,$$

so z has the preimage x under $g \circ f$ [$\exists I$, witness x]. Since $z \in Z$ was arbitrary, $g \circ f$ is surjective.

- (c) If f and g are bijective, they are injective and surjective, so $g \circ f$ is injective by (a) and surjective by (b), that is, bijective.

■

Part (a) is short enough to see whole at level 0 of *Logic and Proof*, Section 5.3, a full derivation. Here is its derivation, with the quantifiers bounded and handled by the bounded rules (Logic and Proof, Proposition 6.1.1), and $(g \circ f)(u)$ already unwound to $g(f(u))$ by [Definition 1.4.2](#). Premises 1 and 2 are the injectivity of f and of g ; premise 3 says that f maps X into Y , which the prose used in the clause “the points $f(x_1)$ and $f(x_2)$ lie in Y ”.

1	$\forall u \in X \forall v \in X (f(u) = f(v) \rightarrow u = v)$	
2	$\forall u \in Y \forall v \in Y (g(u) = g(v) \rightarrow u = v)$	
3	$\forall u \in X f(u) \in Y$	
4	$x_1 \mid x_1 \in X$	
5	$x_2 \mid x_2 \in X$	
6	$g(f(x_1)) = g(f(x_2))$	
7	$f(x_1) \in Y$	$\forall E 3, 4$
8	$f(x_2) \in Y$	$\forall E 3, 5$
9	$\forall v \in Y (g(f(x_1)) = g(v) \rightarrow f(x_1) = v)$	$\forall E 2, 7$
10	$g(f(x_1)) = g(f(x_2)) \rightarrow f(x_1) = f(x_2)$	$\forall E 9, 8$
11	$f(x_1) = f(x_2)$	$\rightarrow E 10, 6$
12	$\forall v \in X (f(x_1) = f(v) \rightarrow x_1 = v)$	$\forall E 1, 4$
13	$f(x_1) = f(x_2) \rightarrow x_1 = x_2$	$\forall E 12, 5$
14	$x_1 = x_2$	$\rightarrow E 13, 11$
15	$g(f(x_1)) = g(f(x_2)) \rightarrow x_1 = x_2$	$\rightarrow I 6-14$
16	$\forall v \in X (g(f(x_1)) = g(f(v)) \rightarrow x_1 = v)$	$\forall I 5-15$
17	$\forall u \in X \forall v \in X (g(f(u)) = g(f(v)) \rightarrow u = v)$	$\forall I 4-16$

Compare it with the prose of (a). “Let $x_1, x_2 \in X$ ” opens the two flagged boxes of lines 4 and 5, and “with $g(f(x_1)) = g(f(x_2))$ ” is the assumption of line 6. The first tagged sentence is lines 7–11, and the second is lines 12–14. The closing “so $g \circ f$ is injective” stands for lines 15–17, the closing lines of $\rightarrow I$ and of the two $\forall I$ ’s; by the convention of *Logic and Proof*, Chapter 5 on silent steps, the prose may state their result without naming them. The two tags in the prose mark the only steps where a hypothesis is used, and that is what the proof is about.

Each hypothesis is needed. In (a), dropping the injectivity of g breaks the conclusion: with $f = \text{id}_{\mathbb{R}}$ and $g = q_1$, squaring on $\mathbb{R}$, the composite $g \circ f = q_1$ is not injective, although f is. Dropping the injectivity of f instead, with $f = q_1$ and $g = \text{id}_{\mathbb{R}}$, gives the same composite. In (b), with $f = \text{id}_{\mathbb{R}}$ and $g = q_1$, the composite q_1 is not surjective, although f is; with $f = q_1$ and $g = \text{id}_{\mathbb{R}}$, again $g \circ f = q_1$ is not surjective, although g is.

The converses are more interesting, and they are the exercises of this section. An injective composite $g \circ f$ forces f to be injective and says nothing of the kind about g (exit check C1). A surjective composite forces g to be surjective, and f may miss points (B2). Section 1.5 builds on these facts: a function f with a function g such that $g \circ f = \text{id}_X$ is forced to be injective.

1.4.7 Exercises

1.4.7.1 A. Check your understanding

Exercise. A1

- (a) Define “ $f: X \rightarrow Y$ is injective” in symbols, and write its negation in symbols.

- (b) Is the empty function $\emptyset \rightarrow \{1\}$ injective? Is it surjective? Justify your answers.
- (c) Which of these relations on $\{1, 2, 3\}$ are graphs of functions $\{1, 2, 3\} \rightarrow \{1, 2, 3\}$? For each that is not, name the clause of the definition that fails. $R_1 = \{(1, 2), (2, 2), (3, 1)\}$, $R_2 = \{(1, 1), (2, 3)\}$, $R_3 = \{(1, 1), (1, 2), (2, 2), (3, 3)\}$.
- (d) How many functions $\{1, 2\} \rightarrow \{a, b, c\}$ are there, and how many of them are injective? Answer by listing.

Solution.

- (a) f is injective: $\forall x_1 \in X \forall x_2 \in X (f(x_1) = f(x_2) \rightarrow x_1 = x_2)$ (Definition 1.4.3). Its negation: $\exists x_1 \in X \exists x_2 \in X (f(x_1) = f(x_2) \wedge x_1 \neq x_2)$. The bounds stay as they are, and the negated implication becomes “hypothesis and not conclusion”, by the negation laws of *Logic and Proof*, Chapter 4.
- (b) Injective: yes. The condition “for all $x_1, x_2 \in \emptyset, \dots$ ” holds vacuously, since there are no $x_1 \in \emptyset$ (Logic and Proof, Proposition 2.1.2). Surjective: no. The witness is 1: there is no $x \in \emptyset$ at all, so none with value 1.
- (c) R_1 is a function ($1 \mapsto 2, 2 \mapsto 2, 3 \mapsto 1$). R_2 is not: clause (Fn1) fails at 3, which has no output. R_3 is not: clause (Fn2) fails at 1, which has the two outputs 1 and 2.
- (d) A function $f: \{1, 2\} \rightarrow \{a, b, c\}$ is determined by the pair of values $(f(1), f(2))$ (Proposition 1.4.1), and each value can be any of the three letters. Listing the pairs:

$$\begin{aligned} &(a, a), (a, b), (a, c), \\ &(b, a), (b, b), (b, c), \\ &(c, a), (c, b), (c, c). \end{aligned}$$

There are 9 functions. A function on $\{1, 2\}$ is injective exactly when $f(1) \neq f(2)$, since 1 and 2 are the only two different inputs. Removing the three pairs with equal entries, on the diagonal of the list, leaves 6 injective functions.

1.4.7.2 B. Practice

Exercise. B1: Five functions

Determine which of the following are injective, which are surjective, and which are not functions at all. Justify your answers.

- (a) $\mathbb{Z} \rightarrow \mathbb{Z}, n \mapsto 2n$.
- (b) $\mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}, (m, n) \mapsto m + n$.
- (c) $\mathbb{R} \rightarrow \mathbb{R}, x \mapsto x^3$.
- (d) $\mathbb{R} \rightarrow (-1, 1), x \mapsto x/(1 + |x|)$.
- (e) $\mathbb{Z}_{>0} \rightarrow \mathbb{Z}_{>0}, n \mapsto n + (-1)^n$.

Solution.

- (a) *Injective:* $2n_1 = 2n_2$ gives $n_1 = n_2$. *Not surjective:* the witness is 1, which is odd, while every $2n$ is even, and no integer is both even and odd.
- (b) *Not injective:* $(1, 2) \neq (2, 1)$ (Definition 1.3.1), and both are sent to 3. *Surjective:* for

$k \in \mathbb{N}$, the pair $(0, k)$ lies in $\mathbb{N} \times \mathbb{N}$ and is sent to k .

(c) *Injective.* Let $x_1^3 = x_2^3$. Then

$$0 = x_1^3 - x_2^3 = (x_1 - x_2)(x_1^2 + x_1x_2 + x_2^2), \quad x_1^2 + x_1x_2 + x_2^2 = (x_1 + \frac{1}{2}x_2)^2 + \frac{3}{4}x_2^2.$$

If $x_1 \neq x_2$, the second factor is positive: if $x_2 \neq 0$, then $\frac{3}{4}x_2^2 > 0$; if $x_2 = 0$, it equals $x_1^2 > 0$, because $x_1 \neq 0$. Then the product is nonzero, a contradiction. So $x_1 = x_2$. *Surjective:* for $y \in \mathbb{R}$, every real number has a real cube root, and $x := \sqrt[3]{y}$ satisfies $x^3 = y$. So the function is bijective.

(d) *A function:* for every x , $1 + |x| > 0$, and $|x/(1 + |x|)| = |x|/(1 + |x|) < 1$, so the value lies in $(-1, 1)$. *Injective.* Suppose $y = x/(1 + |x|)$; we solve for x . Taking absolute values, $|y| = |x|/(1 + |x|)$, so $|y|(1 + |x|) = |x|$, that is, $|x|(1 - |y|) = |y|$; since $|y| < 1$ (shown above), $1 - |y| > 0$, and $|x| = |y|/(1 - |y|)$, and then $1 + |x| = 1/(1 - |y|)$. Hence

$$x = y(1 + |x|) = \frac{y}{1 - |y|}.$$

So x is determined by y : two inputs with the same output y are both equal to $y/(1 - |y|)$, hence equal. *Surjective.* Let $y \in (-1, 1)$ and put $x := y/(1 - |y|)$, a real number since $1 - |y| > 0$. Then $|x| = |y|/(1 - |y|)$, so $1 + |x| = 1/(1 - |y|)$, and

$$\frac{x}{1 + |x|} = \frac{y}{1 - |y|} \cdot (1 - |y|) = y.$$

So the function is bijective. (For example, $x = 1 \mapsto \frac{1}{2}$ and $y = \frac{1}{2}$ comes back to $x = \frac{1/2}{1 - 1/2} = 1$.)

(e) *Not a function* $\mathbb{Z}_{>0} \rightarrow \mathbb{Z}_{>0}$: at $n = 1$ the formula gives $1 + (-1) = 0$, and $0 \notin \mathbb{Z}_{>0}$, so (Fn1) fails at 1. This is the planted failure: the question whether it is injective or surjective does not arise. (On $\mathbb{N} = \{0, 1, 2, \dots\}$ the same formula is a function, and a bijection: it swaps $0 \leftrightarrow 1$, $2 \leftrightarrow 3$, and so on.)

Exercise. B2: Put the proof in order

Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be functions. The following lines contain a proof that **if $g \circ f$ is surjective, then g is surjective**, shuffled, together with two lines that do not belong. Write the six correct lines in an order that makes a proof, and say why each of the other two lines fails.

- (i) Then $g(y) = g(f(x)) = (g \circ f)(x) = z$, by [Definition 1.4.2](#).
- (ii) Let $y \in Y$.
- (iii) Since $g \circ f$ is surjective, there is $x \in X$ with $(g \circ f)(x) = z$.
- (iv) Since z was arbitrary, g is surjective.
- (v) Let $y := f(x)$, an element of Y .
- (vi) Since f is surjective, $y = f(x)$ for some $x \in X$.
- (vii) Let $z \in Z$.
- (viii) So z has a preimage under g , namely y .

Solution. The order is (vii), (iii), (v), (i), (viii), (iv):

Let $z \in Z$. Since $g \circ f$ is surjective, there is $x \in X$ with $(g \circ f)(x) = z$. Let $y := f(x)$, an element of Y . Then $g(y) = g(f(x)) = (g \circ f)(x) = z$, by [Definition 1.4.2](#). So z has a preimage under g , namely y . Since z was arbitrary, g is surjective.

No other order is admissible. Each line uses the one before it: (iii) needs the z of (vii); (v) needs the x of (iii); (i) needs the y of (v); (viii) is the $\exists I$ that (i) makes possible; and (iv) closes the $\forall I$ opened by (vii).

The distractor (ii) opens with an arbitrary element of the wrong set. The goal “ g is surjective” is $\forall z \in Z \exists y \in Y g(y) = z$, so the proof must take an arbitrary $z \in Z$; the y is the witness to be found, not something given. The distractor (vi) uses a hypothesis the statement does not have: nothing says f is surjective, and the claim holds without it. For example, with $f: \{1\} \rightarrow \{1, 2\}$, $f(1) = 1$, and $g: \{1, 2\} \rightarrow \{1\}$ constant, $g \circ f$ and g are surjective and f is not.

Exercise. B3: Examples on the natural numbers

Give an example of a function $\mathbb{N} \rightarrow \mathbb{N}$ that is surjective and not injective, and one that is injective and not surjective. Check each property.

Solution. Surjective, not injective. Let $s(0) := 0$ and $s(n) := n - 1$ for $n \geq 1$; for $n \geq 1$, $n - 1 \in \mathbb{N}$, so $s: \mathbb{N} \rightarrow \mathbb{N}$ is a function. It is surjective: for $m \in \mathbb{N}$, $m + 1 \geq 1$, so $s(m + 1) = m$. It is not injective: $s(0) = 0 = s(1)$ and $0 \neq 1$.

Injective, not surjective. Let $t(n) := n + 1$. It is injective: $n_1 + 1 = n_2 + 1$ gives $n_1 = n_2$. It is not surjective: 0 is not an output, since $n + 1 \geq 1$.

Many other answers are correct. An answer to the first part must have: every value in $\mathbb{N}$ (so it is a function into $\mathbb{N}$); at least one value taken at two different inputs, with the two inputs named; and every $m \in \mathbb{N}$ reached, with a preimage given as a formula in m , since checking some m is not enough. An answer to the second part must have: values in $\mathbb{N}$; a proof that equal outputs come from equal inputs; and one named natural number that is never reached, with the reason.

1.4.7.3 C. Going deeper

Exercise. C1: Exit check: composites

Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be functions.

- Prove that if $g \circ f$ is injective, then f is injective.
- Give an example in which $g \circ f$ is injective and g is not injective, and check both claims.

Solution.

- Suppose $g \circ f$ is injective. Let $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$; we must show $x_1 = x_2$.

Applying g to both sides gives $g(f(x_1)) = g(f(x_2))$, because a function gives equal outputs at equal inputs. By [Definition 1.4.2](#), this says $(g \circ f)(x_1) = (g \circ f)(x_2)$. Since $g \circ f$ is injective, $x_1 = x_2$. So f is injective.

- (b) Let $X = \{1\}$, $Y = \{1, 2\}$, $Z = \{1\}$; let $f(1) = 1$, and let $g: \{1, 2\} \rightarrow \{1\}$ be the constant function $g(1) = g(2) = 1$. Then $g \circ f: \{1\} \rightarrow \{1\}$ is injective, because its domain has only one element: if $x_1, x_2 \in \{1\}$, then $x_1 = 1 = x_2$. And g is not injective: $g(1) = g(2)$ with $1 \neq 2$.

Marking scheme. (10 points) - 2: fixes arbitrary $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$ and states the goal $x_1 = x_2$. - 2: applies g to both sides, with the reason that a function gives equal outputs at equal inputs. - 3: rewrites the result as an equation of values of $g \circ f$ and invokes the injectivity of $g \circ f$. - 3: an example with explicit sets and functions, and both claims checked: $g \circ f$ injective, and two different points with the same value under g .

Fatal errors (0 for the step): assuming g is injective; starting from $x_1 = x_2$ and deriving $f(x_1) = f(x_2)$ (the converse, entry (N5) of *Logic and Proof*, Chapter 5's catalogue); an example in which $g \circ f$ is not defined because the codomain of f is not the domain of g . A proof by contrapositive (from $x_1 \neq x_2$, $g \circ f$ injective gives $g(f(x_1)) \neq g(f(x_2))$, hence $f(x_1) \neq f(x_2)$, since equal values of f would give equal values of g) earns full marks: 2 points for fixing $x_1 \neq x_2$ and the goal, 3 for the use of the injectivity of $g \circ f$, and 2 for the step from $g(f(x_1)) \neq g(f(x_2))$ to $f(x_1) \neq f(x_2)$.

Exercise. C2: When the composite is a bijection

Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be functions.

- (a) Deduce from [Exercise \(C1: Exit check: composites\)](#) and [Exercise \(B2: Put the proof in order\)](#) that if $g \circ f$ is bijective, then f is injective and g is surjective.
- (b) Give an example in which $g \circ f$ is bijective and neither f nor g is bijective.

Solution.

- (a) Suppose $g \circ f$ is bijective. Then $g \circ f$ is injective, so f is injective by part (a) of [Exercise \(C1: Exit check: composites\)](#). And $g \circ f$ is surjective, so g is surjective by the proof put in order in [Exercise \(B2: Put the proof in order\)](#).
- (b) Take the functions of [Exercise \(C1: Exit check: composites\)](#) (b): $f: \{1\} \rightarrow \{1, 2\}$, $f(1) = 1$, and $g: \{1, 2\} \rightarrow \{1\}$, $g(1) = g(2) = 1$. Then $g \circ f: \{1\} \rightarrow \{1\}$ sends 1 to 1, so $g \circ f = \text{id}_{\{1\}}$ by [Proposition 1.4.1](#), which is bijective. But f is not surjective, since 2 is not an output; and g is not injective, since $g(1) = g(2)$. So neither is bijective. Part (a) is sharp: it gives exactly the half of each property that survives, and this example shows that the other halves can fail together.

1.5 Left, Right and Two-Sided Inverses

Section 1.4 sorted functions by two questions. Does f ever send two different inputs to the same output? Does f reach every point of its codomain? The answers were the words injective and surjective ([Definition 1.4.3](#), [Definition 1.4.4](#)). This section asks a question that sounds different: can the work of f be **undone**? It turns out that it asks the same two questions again, one for

each side of f . Undoing f after it has acted forces f to be injective, and when X is non-empty injectivity is also enough. Undoing it before, so that f finishes the job, forces f to be surjective. The one new idea in the section is an inverse on one side; an inverse on both sides, written f^{-1} , is simply a map that is both.

Throughout, $f: X \rightarrow Y$ is a function in the sense of Definition 1.4.1, and composites are formed as in Definition 1.4.2: $g \circ f$ is defined when the codomain of f is the domain of g , and $(g \circ f)(x) = g(f(x))$.

1.5.1 Undoing a function

Start with a function we can compute in our heads. Let $d: \mathbb{N} \rightarrow \mathbb{N}$ be the doubling map, $d(n) = 2n$. Suppose someone applies d to a secret number and tells us the result, say 14. We recover the secret by halving: it was 7. Halving undoes doubling.

But halving is not quite a function from $\mathbb{N}$ to $\mathbb{N}$: 7 has no half in $\mathbb{N}$. If we want a function $h: \mathbb{N} \rightarrow \mathbb{N}$ that undoes d , we must give h a value at every odd number as well. Those values do not matter for the undoing, because d never produces an odd number. Any choice will do. So one answer is

$$h(m) := \begin{cases} m/2 & \text{if } m \text{ is even,} \\ 1 & \text{if } m \text{ is odd,} \end{cases}$$

and another is $h'(m) := m/2$ for even m and $h'(m) := m$ for odd m . Both are functions $\mathbb{N} \rightarrow \mathbb{N}$: every $m \in \mathbb{N}$ is even or odd and not both, and an even m is $2k$ for exactly one $k \in \mathbb{N}$, which is $m/2$. Both undo d : for every $n \in \mathbb{N}$, the number $2n$ is even, so

$$h(d(n)) = h(2n) = \frac{2n}{2} = n,$$

and the same computation holds for h' . In the language of composition, $h \circ d = \text{id}_{\mathbb{N}}$ and $h' \circ d = \text{id}_{\mathbb{N}}$.

Now try to undo in the other order. Is there a function $k: \mathbb{N} \rightarrow \mathbb{N}$ with $d \circ k = \text{id}_{\mathbb{N}}$, so that d undoes k ? Such a k would give $d(k(1)) = 1$, that is $2k(1) = 1$, and 1 is odd, not twice a natural number. So there is no such k . The obstacle is visible: 1 is not in the range of d , so no input to d , however chosen, comes out as 1.

Two things happened here, and they are the whole section in miniature. Undoing d **afterwards** was possible, in more than one way, and the reason is that d loses no information: different inputs give different outputs. Undoing it **beforehand** was impossible, and the reason is that d misses a point of its codomain. The two sides behave differently, so they get two names.

1.5.2 Left and right inverses

A left inverse undoes f after it acts; a right inverse is undone by f .

Definition 1.5.1. Left and Right Inverses

Let $f: X \rightarrow Y$ be a function, and let $g: Y \rightarrow X$ be a function in the opposite direction.

- (a) g is a **left inverse** of f if $g \circ f = \text{id}_X$, that is, $g(f(x)) = x$ **for every** $x \in X$.
- (b) g is a **right inverse** of f if $f \circ g = \text{id}_Y$, that is, $f(g(y)) = y$ **for every** $y \in Y$.

In words: a left inverse is written on the **left** of f in the composite, and the composite $g \circ f$ does nothing to points of X . A right inverse is written on the **right**, and $f \circ g$ does nothing to

points of Y . The two phrasings in each clause say the same thing. Two functions $X \rightarrow X$ are equal exactly when they agree at every point ([Proposition 1.4.1](#)), and $g \circ f$ and id_X have the same domain and codomain X ; so $g \circ f = \text{id}_X$ means $(g \circ f)(x) = \text{id}_X(x)$ for every $x \in X$, which is $g(f(x)) = x$.

The phrase “ g is a left inverse of f ” and the phrase “ f is a right inverse of g ” say the same thing: both say $g \circ f = \text{id}_X$. Which word applies depends on which function we are talking about.

Check the definition against the doubling map. With $d(n) = 2n$ on $\mathbb{N}$, the functions h and h' above are left inverses of d : the composite $h \circ d$ sends every n back to n . And d has **no** right inverse, because clause (b) fails at the single point $y = 1$. For a right inverse the “for every $y \in Y$ ” is the demanding part: one bad y is enough to rule it out.

Example. Left inverses on small sets

Let $X = \{1, 2\}$, $Y = \{1, 2, 3\}$, and $e: X \rightarrow Y$, $e(1) = 1$, $e(2) = 2$: the inclusion of X into Y .

- (a) Show that $g: Y \rightarrow X$ with $g(1) = 1$, $g(2) = 2$, $g(3) = 2$ is a left inverse of e , and that e is a right inverse of g .
- (b) Is g a right inverse of e ?
- (c) Let $c: X \rightarrow Y$ be constant, $c(1) = c(2) = 3$. Does c have a left inverse?

Solution.

- (a) Clause (a) of [Definition 1.5.1](#) asks for $g(e(x)) = x$ at each of the two points of X : $g(e(1)) = g(1) = 1$ and $g(e(2)) = g(2) = 2$. So $g \circ e = \text{id}_X$, and g is a left inverse of e . Read with the roles exchanged, the same equation $g \circ e = \text{id}_X$ is clause (b) for $g: Y \rightarrow X$ and $e: X \rightarrow Y$: e is a right inverse of g .
- (b) No. Clause (b) would need $e(g(y)) = y$ for every $y \in Y$, and at $y = 3$ we get $e(g(3)) = e(2) = 2 \neq 3$. The equation $e \circ g = \text{id}_Y$ fails at one point, which is enough.
- (c) No. A left inverse g would need $g(c(1)) = 1$ and $g(c(2)) = 2$, that is $g(3) = 1$ and $g(3) = 2$. A function has one value at 3, and $1 \neq 2$. The information “which input was it” is gone once both inputs land on 3, and nothing applied afterwards can bring it back.

Part (c) is the degenerate case of the example, and it is the one to remember. A left inverse must recover x from $f(x)$, so it cannot exist when two inputs share an output. Part (b) shows the other side: g undoes e , but e does not undo g , because e never produces 3.

Warning. A left inverse need not be a right inverse. For the doubling map $d(n) = 2n$ on $\mathbb{N}$ and its left inverse h above, $h(d(n)) = n$ for every n , but $d(h(1)) = d(1) = 2 \neq 1$. The equation $h \circ d = \text{id}_{\mathbb{N}}$ says nothing about $d \circ h$: composition is not commutative (Section 1.4), and here the two composites are different functions.

Quick check. Let $s: \mathbb{N} \rightarrow \mathbb{N}$ be the successor map, $s(n) = n + 1$. Does s have a left inverse? Does it have a right inverse?

Solution. It has a left inverse, for instance $p(m) := m - 1$ for $m \geq 1$ and $p(0) := 0$. This is a function $\mathbb{N} \rightarrow \mathbb{N}$, since $m - 1 \in \mathbb{N}$ when $m \geq 1$, and $p(s(n)) = p(n + 1) = n$ because $n + 1 \geq 1$. It has no right inverse: a right inverse $k: \mathbb{N} \rightarrow \mathbb{N}$ of s would need $s(k(0)) = 0$, that is $k(0) + 1 = 0$, which no $k(0) \in \mathbb{N}$ satisfies. The number 0 is not reached by s .

1.5.3 A left inverse exists exactly for injective functions

The examples point to a pattern. The doubling map, the inclusion e , and the successor map are injective, and each has a left inverse. The constant map c is not injective, and it has none. Here is the pattern as a theorem. Its hypothesis $X \neq \emptyset$ looks like a technicality; after the proof we will see that it is not.

We need the **range** of f from Section 1.4, and from now on we write it as $f(X)$:

$$f(X) = \{y \in Y : f(x) = y \text{ for some } x \in X\}.$$

Section 1.6 extends this notation to the image of any subset of X .

Theorem 1.5.1. Left Inverses and Injectivity

Let $f: X \rightarrow Y$ be a function.

- (a) If f has a left inverse, then f is injective.
- (b) If $X \neq \emptyset$ and f is injective, then f has a left inverse.

Consequently, if $X \neq \emptyset$, then f has a left inverse **if and only if** f is injective.

Part (a) carries no hypothesis on X , on purpose: the last theorem of this section uses it for every X , the empty set included. The hypothesis $X \neq \emptyset$ sits in (b) only, where it is used. As in [Proposition 1.4.2](#) (b), “ $X \neq \emptyset$ ” is read as “ X is non-empty: it has an element”, and the proof uses it in that form. (From an element, $X \neq \emptyset$ follows at once, since the empty set has none ([Definition 1.1.4](#)). The passage back, from “ X is not the empty set” to “ X has an element”, is the classical law “not all fail, so one holds” of *Logic and Proof*, Chapter 4; reading the hypothesis as “has an element” avoids it.)

Idea. For (a), start from the goal. Injectivity asks: from $f(x_1) = f(x_2)$, get $x_1 = x_2$. A left inverse g turns each $f(x)$ back into x , so apply g to both sides.

For (b), build g point by point, the way halving was extended to the odd numbers. ① A point y of the range has a preimage, and injectivity says it has only one; let $g(y)$ be that one. ② A point outside the range has no preimage; send it to a fixed point $x_0 \in X$. This is where $X \neq \emptyset$ is needed. ③ Check that ① and ② together define a function, and that $g(f(x)) = x$.

Before reading on: where is $X \neq \emptyset$ used, and what would go wrong at that line if X were empty?

Proof.

- (a) Let $g: Y \rightarrow X$ be a left inverse of f . Let $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$. Since g is a function, equal inputs give equal outputs, so $g(f(x_1)) = g(f(x_2))$. By [Definition 1.5.1](#) (a), $g(f(x)) = x$ for every $x \in X$, so

$$x_1 = g(f(x_1)) = g(f(x_2)) = x_2.$$

Since x_1 and x_2 were arbitrary, f is injective by [Definition 1.4.3](#).

- (b) Suppose $X \neq \emptyset$ and f is injective. Since X has an element, fix one and call it x_0 $[\exists E]$. Let G be the subset of $Y \times X$ given by specification ([Definition 1.1.3](#)):

$$G := \{ (y, x) \in Y \times X : (y \in f(X) \text{ and } f(x) = y) \text{ or } (y \notin f(X) \text{ and } x = x_0) \}.$$

We show that for every $y \in Y$ there is exactly one $x \in X$ with $(y, x) \in G$, by checking the two clauses (Fn1) (existence) and (Fn2) (uniqueness) into which Section 1.4 splits the condition of [Definition 1.4.1](#) (Logic and Proof, Proposition 5.8.1). Then G , with codomain X , is a function $g: Y \rightarrow X$. Let $y \in Y$.

Existence (Fn1). By excluded middle (Logic and Proof, Proposition 2.5.3), $y \in f(X)$ or $y \notin f(X)$. *Case 1:* $y \in f(X)$. Then $f(x_1) = y$ for some $x_1 \in X$; name it x_1 $[\exists E]$. The pair (y, x_1) satisfies the first alternative, so $(y, x_1) \in G$. *Case 2:* $y \notin f(X)$. Then (y, x_0) satisfies the second alternative, so $(y, x_0) \in G$. In either case some $x \in X$ has $(y, x) \in G$.

Uniqueness (Fn2). Let $x, x' \in X$ with $(y, x) \in G$ and $(y, x') \in G$. Each pair satisfies one of the two alternatives. If both satisfy the first, then $f(x) = y = f(x')$, so $x = x'$ because f is injective. If both satisfy the second, then $x = x_0 = x'$. If one satisfies the first and the other the second, then $y \in f(X)$ and $y \notin f(X)$, a contradiction, so this case cannot occur. Hence $x = x'$.

So $g: Y \rightarrow X$ is a function, and $g(y) = x$ means $(y, x) \in G$. It remains to show $g \circ f = \text{id}_X$. Let $x \in X$. Then $f(x) \in f(X)$, with x as the witness $[\exists I]$, and $f(x) = f(x)$, so the pair $(f(x), x)$ satisfies the first alternative: $(f(x), x) \in G$, that is, $g(f(x)) = x$. The functions $g \circ f$ and id_X have domain and codomain X and agree at every point, so they are equal by [Proposition 1.4.1](#). Hence g is a left inverse of f .

The last sentence of the theorem combines (a) and (b). This proves the theorem. ■

Look at where each hypothesis did its work. Injectivity was used once, in the uniqueness step: it is what makes “the preimage of y ” a single point. The hypothesis $X \neq \emptyset$ was used once, in the first line of (b), to have a point x_0 for the second alternative. The case split on $y \in f(X)$ used excluded middle, a classical step: the proof does not tell us, for a given y , which case we are in.

Neither hypothesis of (b) can be dropped.

- **Without injectivity.** The constant map c of [Example \(Left inverses on small sets\)](#) (c) sends 1 and 2 to 3, and it has no left inverse. In the proof, the uniqueness step breaks: $(3, 1)$ and $(3, 2)$ both lie in G , so G is not a function.
- **Without $X \neq \emptyset$.** The empty function $\emptyset \rightarrow \{1\}$ is injective: the condition “for all $x_1, x_2 \in \emptyset$...” holds vacuously. A left inverse would be a function $\{1\} \rightarrow \emptyset$, and there is none, since

the point 1 would need a value in the empty set (Proposition 1.4.2). In the proof, the first line fails: there is no x_0 to fix, and the point 1, which lies outside the range, has nowhere to go.

The proof of (b) describes g explicitly in terms of f and x_0 , so on a small example the construction can be run by hand.

Example. Running the construction

Let $f: \{1, 2\} \rightarrow \{a, b, c\}$ be given by $f(1) = a$ and $f(2) = c$, where a, b, c are three different objects. Run the proof of Theorem 1.5.1 (b) with $x_0 = 1$, then with $x_0 = 2$. How many left inverses does f have?

Solution. First the hypotheses: $\{1, 2\} \neq \emptyset$, and f is injective because its two inputs have the different outputs a and c . The range is $f(\{1, 2\}) = \{a, c\}$.

With $x_0 = 1$: the points a and c of the range go back to their unique preimages, $g(a) = 1$ and $g(c) = 2$. The point b lies outside the range, so $g(b) = x_0 = 1$. Check: $g(f(1)) = g(a) = 1$ and $g(f(2)) = g(c) = 2$, so $g \circ f = \text{id}_{\{1, 2\}}$.

With $x_0 = 2$: $g'(a) = 1$, $g'(c) = 2$ as before, and $g'(b) = 2$. The same check shows $g' \circ f = \text{id}_{\{1, 2\}}$.

These are all the left inverses. Any left inverse k must satisfy $k(a) = k(f(1)) = 1$ and $k(c) = k(f(2)) = 2$, so its values on the range are **forced**. Its value at b is free, and there are two points of $\{1, 2\}$ to choose from. So f has exactly 2 left inverses, and the construction with $x_0 = 1$ and with $x_0 = 2$ produces both.

The example shows that a left inverse is **not unique** in general. The values on the range are forced by the equation $g(f(x)) = x$; the values off the range are free, and they may go to any point of X . (They are free but not always varied: if X has only one point, there is only one place to send them.)

1.5.4 A right inverse forces surjectivity

The doubling map has no right inverse because it misses 1. The same obstacle works for every function.

Proposition 1.5.2. Right Inverses Give Surjectivity

Let $f: X \rightarrow Y$ be a function. If f has a right inverse, then f is surjective.

Proof. Let $g: Y \rightarrow X$ be a right inverse of f , and let $y \in Y$. By Definition 1.5.1 (b), $f(g(y)) = y$. So $x := g(y)$ is an element of X with $f(x) = y$ $[\exists I]$. Since y was arbitrary, f is surjective by Definition 1.4.4. ■

The proof names its witness: the right inverse **is** a rule that hands us, for each y , a preimage

$g(y)$. No hypothesis on X or Y is needed. If $Y = \emptyset$, then f is surjective vacuously; and $X = \emptyset$ as well, since there is no function from a non-empty set to $\emptyset$ (Proposition 1.4.2).

Right inverses are not unique either. Here is a small surjection with several.

Example. All right inverses of a small surjection

Let $q: \{1, 2, 3, 4\} \rightarrow \{0, 1\}$ send each number to 0 if it is even and to 1 if it is odd: $q(1) = 1$, $q(2) = 0$, $q(3) = 1$, $q(4) = 0$. Find all right inverses of q .

Solution. A right inverse is a function $k: \{0, 1\} \rightarrow \{1, 2, 3, 4\}$ with $q(k(0)) = 0$ and $q(k(1)) = 1$. The first condition says that $k(0)$ is a preimage of 0, so $k(0) \in \{2, 4\}$. The second says $k(1) \in \{1, 3\}$. The two choices are independent, and each combination gives a function $\{0, 1\} \rightarrow \{1, 2, 3, 4\}$ satisfying both conditions. So q has exactly 4 right inverses:

$$(k(0), k(1)) = (2, 1), \quad (2, 3), \quad (4, 1), \quad (4, 3).$$

Each one picks one preimage of each point of the codomain. By Proposition 1.5.2, the fact that one exists shows again that q is surjective.

1.5.5 Does every surjection have a right inverse?

Theorem 1.5.1 went both ways: for $X \neq \emptyset$, a left inverse exists exactly when f is injective. Does Proposition 1.5.2 go both ways too? Suppose $f: X \rightarrow Y$ is surjective. Must f have a right inverse?

Read what a right inverse must do. For **every** $y \in Y$, $g(y)$ must be a preimage of y : an x with $f(x) = y$. Surjectivity promises that each y has at least one preimage. It does not say which, and there may be many. So the right inverse must **choose** one preimage for every y , all at once, and assemble the choices into one set of pairs.

Now look at what our rules can do.

- **One y .** For a single y , the statement “there is $x \in X$ with $f(x) = y$ ” lets us name a preimage by $\exists E$ (*Logic and Proof*, Section 4.6). That is one choice, made inside one proof.
- **Finitely many y .** For a particular finite $Y = \{y_1, \dots, y_n\}$, with n a fixed number, we can apply $\exists E$ n times, once for each point, and write down the n pairs. (To get this for every finite Y at once, with n arbitrary, one inducts on n , *Logic and Proof*, Axiom 4.2.1.) Example (All right inverses of a small surjection) picked one preimage per point in the same spirit, by listing the preimage sets and choosing from them.
- **Infinitely many y .** A proof is a finite object. It cannot contain one $\exists E$ for each of infinitely many points, and no rule of *Logic and Proof*, Chapter 4 makes infinitely many choices in one step.

Sometimes the difficulty does not arise, because a **rule** picks the preimage. The squaring map from $\mathbb{R}$ onto $[0, \infty)$ is surjective, and each $y \geq 0$ has the preimages $\sqrt{y}$ and $-\sqrt{y}$. The rule “take the non-negative one” defines $g(y) := \sqrt{y}$, and g is a right inverse: $(\sqrt{y})^2 = y$. Likewise the projection $\mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$, $(m, n) \mapsto m$, is surjective, and $m \mapsto (m, 1)$ is a right inverse. In each case the specification of Definition 1.1.3 writes g down, because the rule names one preimage for each y .

For a surjection with no such rule, nothing proved so far produces a right inverse. The missing ingredient is a new principle of set formation, which says that a choice of one element from each set of a family can be made. Chapter 3 states it and proves, with it, that every surjection has a right inverse. Until then, we use only the implication proved above: a right inverse gives surjectivity.

1.5.6 Two-sided inverses

The two halves meet in a single fact about composition: when a function has an inverse on each side, the two inverses coincide.

Proposition 1.5.3. Left and Right Inverses Agree

Let $f: X \rightarrow Y$ be a function. If $g: Y \rightarrow X$ is a left inverse of f and $h: Y \rightarrow X$ is a right inverse of f , then $g = h$.

Proof. By Proposition 1.4.3 (associativity and the identity laws),

$$g = g \circ \text{id}_Y = g \circ (f \circ h) = (g \circ f) \circ h = \text{id}_X \circ h = h,$$

where the first and last equalities are the identity laws, the second uses that h is a right inverse, the third is associativity, and the fourth uses that g is a left inverse. This proves the proposition. ■

The chain is worth reading twice. The composite $g \circ f \circ h$ can be simplified from either end: the left pair $g \circ f$ collapses to id_X , the right pair $f \circ h$ collapses to id_Y . Associativity says both simplifications are of the same function.

Definition 1.5.2. Inverse Function

Let $f: X \rightarrow Y$ be a function. A function $g: Y \rightarrow X$ is a **two-sided inverse**, or an **inverse**, of f if it is both a left inverse and a right inverse of f :

$$g \circ f = \text{id}_X \quad \text{and} \quad f \circ g = \text{id}_Y.$$

A function with an inverse is **invertible**. Its inverse is **unique**: if g and g' are both inverses of f , then g is a left inverse and g' a right inverse, so $g = g'$ by Proposition 1.5.3. We call it **the** inverse of f and write it $f^{-1}: Y \rightarrow X$.

In words: f^{-1} undoes f on both sides. It sends $f(x)$ back to x , and f sends $f^{-1}(y)$ back to y . The article “the” was earned by Proposition 1.5.3; without it, f^{-1} could name several functions, as “a left inverse” does.

The definition is symmetric in f and f^{-1} . The two equations that say f^{-1} is an inverse of f also say that f is an inverse of f^{-1} . So f^{-1} is invertible, and by uniqueness its inverse is f :

$$(f^{-1})^{-1} = f.$$

This is the pattern for every identity about inverses: to show that some function is **the** inverse of f , check the two equations of Definition 1.5.2, and uniqueness does the rest.

Example. Inverses found and checked

Show that each function is invertible by finding its inverse and checking both composites.

- (a) $a: \mathbb{R} \rightarrow \mathbb{R}$, $a(x) = 3x - 2$.
- (b) $r: (0, \infty) \rightarrow (0, \infty)$, $r(x) = 1/x$.
- (c) $s: \mathbb{R} \rightarrow \mathbb{R}$, $s(x) = x|x|$, that is, $s(x) = x^2$ for $x \geq 0$ and $s(x) = -x^2$ for $x < 0$.

Solution. Idea. Solve $y = f(x)$ for x to find a candidate, then check both equations of [Definition 1.5.2](#). The solving is scratch work; the two checks are the proof.

- (a) Solving $y = 3x - 2$ gives $x = (y + 2)/3$. Let $b(y) := (y + 2)/3$, a function $\mathbb{R} \rightarrow \mathbb{R}$. For every $x \in \mathbb{R}$, $b(a(x)) = (3x - 2 + 2)/3 = x$, and for every $y \in \mathbb{R}$, $a(b(y)) = 3 \cdot (y + 2)/3 - 2 = y$. So $b = a^{-1}$.
- (b) For $x > 0$, $1/x > 0$, so r does map $(0, \infty)$ to $(0, \infty)$. Here $r(r(x)) = 1/(1/x) = x$ for every $x > 0$. This one equation is both checks at once, with r in the roles of f and g . So $r^{-1} = r$: the function is its own inverse.
- (c) Solve by cases on the sign of y , since $s(x)$ has the sign of x . For $y \geq 0$ the candidate is $\sqrt{y}$; for $y < 0$ it is $-\sqrt{-y}$. Let

$$t(y) := \begin{cases} \sqrt{y} & \text{if } y \geq 0, \\ -\sqrt{-y} & \text{if } y < 0. \end{cases}$$

First check, $t(s(x)) = x$, by cases on x . If $x \geq 0$, then $s(x) = x^2 \geq 0$, so $t(s(x)) = \sqrt{x^2} = x$, the non-negative square root of x^2 being x . If $x < 0$, then $s(x) = -x^2 < 0$, so $t(s(x)) = -\sqrt{-x^2} = -(-x) = x$, since the non-negative square root of x^2 is $-x$.

Second check, $s(t(y)) = y$, by cases on y . If $y \geq 0$, then $t(y) = \sqrt{y} \geq 0$, so $s(t(y)) = (\sqrt{y})^2 = y$. If $y < 0$, then $t(y) = -\sqrt{-y} < 0$, so $s(t(y)) = -(\sqrt{-y})^2 = -(-y) = y$.

Both checks hold, so $t = s^{-1}$. Notice that each check needs its own case split: the first splits on the sign of the input x , the second on the sign of y .

Now the theorem that connects inverses with the words of Section 1.4.

Theorem 1.5.4. Bijections Are the Invertible Functions

A function $f: X \rightarrow Y$ is bijective if and only if it has a two-sided inverse.

Idea. $(\Leftarrow)$ is already proved in two halves: an inverse is a left inverse, which gives injectivity by [Theorem 1.5.1](#) (a), and a right inverse, which gives surjectivity by [Proposition 1.5.2](#). Neither half asks anything of X .

$(\Rightarrow)$ looks like the open question of the last subsection, since we must pick a preimage of every y . But for a bijection there is nothing to pick. Surjectivity gives each y a preimage and injectivity says there is only one. So the set of pairs (y, x) with $f(x) = y$ already has the $\forall \exists!$ property of [Definition 1.4.1](#).

Proof. ($\Leftarrow$) Let g be a two-sided inverse of f . Then g is a left inverse of f , so f is injective by [Theorem 1.5.1](#) (a); and g is a right inverse, so f is surjective by [Proposition 1.5.2](#). Hence f is bijective by [Definition 1.4.5](#).

($\Rightarrow$) Let f be bijective, and let

$$G := \{(y, x) \in Y \times X : f(x) = y\},$$

a subset of $Y \times X$ by specification ([Definition 1.1.3](#)). Let $y \in Y$. Since f is surjective, there is $x \in X$ with $f(x) = y$, so some x has $(y, x) \in G$. If $(y, x) \in G$ and $(y, x') \in G$, then $f(x) = y = f(x')$, so $x = x'$ since f is injective. So G satisfies (Fn1) and (Fn2) (Logic and Proof, Proposition 5.8.1): for every $y \in Y$ there is exactly one $x \in X$ with $(y, x) \in G$, and G with codomain X is a function $g: Y \rightarrow X$ ([Definition 1.4.1](#)). By construction, $g(y) = x$ holds exactly when $f(x) = y$.

Now let $y \in Y$. Then $g(y) = x$ for the x with $(y, x) \in G$, so $f(g(y)) = y$. And let $x \in X$. The pair $(f(x), x)$ lies in G , since $f(x) = f(x)$; so $g(f(x)) = x$. By [Proposition 1.4.1](#), $f \circ g = \text{id}_Y$ and $g \circ f = \text{id}_X$. So g is a two-sided inverse of f . This proves the theorem. ■

Compare the two constructions of this section. The left inverse of [Theorem 1.5.1](#) (b) needed a fixed point x_0 and a case split by excluded middle, because some y have no preimage. The inverse of a bijection needs neither. Every y has **exactly one** preimage, so the definition “ $g(y)$ is the preimage of y ” names one element for each y without any choosing. This is why no new principle is needed here, while the open question about surjections needs one. The difference between the two is exactly the difference between “exactly one” and “at least one”.

The theorem also covers the degenerate case. A function $f: \emptyset \rightarrow Y$ is surjective only if $Y = \emptyset$, since a point of Y would need a preimage in $\emptyset$. The empty function $\emptyset \rightarrow \emptyset$ is bijective, and it is its own inverse: both composites are the empty function $\emptyset \rightarrow \emptyset$, which is $\text{id}_\emptyset$. The proof of ($\Leftarrow$) runs at $X = \emptyset$ without change, because it cites only part (a) of [Theorem 1.5.1](#), which has no hypothesis on X .

Remark. The notation f^{-1} is reserved, as a function, for a bijection f . For a function that is not bijective, “ $f^{-1}: Y \rightarrow X$ ” names nothing. Section 1.6 uses the symbol in a second way, $f^{-1}(B)$ for a **set** B , which makes sense for every function; it shows that for a bijection the two readings give the same set.

Warning. Do not write f^{-1} for a left inverse. The doubling map d has the two different left inverses h and h' of the first subsection, so “ d^{-1} ” would not say which one is meant. And neither is an inverse: $d(h(1)) = 2 \neq 1$. The notation is earned by [Proposition 1.5.3](#), which needs an inverse on **both** sides.

Quick check. Does $x \mapsto x^3$ from $\mathbb{R}$ to $\mathbb{R}$ have a two-sided inverse? Does $x \mapsto x^2$ from $[0, \infty)$ to $\mathbb{R}$?

Solution. The first does: $y \mapsto \sqrt[3]{y}$, the real cube root, which exists and is unique for every real y (a school fact). Indeed $\sqrt[3]{x^3} = x$ and $(\sqrt[3]{y})^3 = y$. The second does not. It is injective (Example (Squaring, three ways) (b)), but it is not surjective, since -1 has no preimage; by Theorem 1.5.4 it has no two-sided inverse. (It has left inverses, by Theorem 1.5.1, for instance $y \mapsto \sqrt{y}$ for $y \geq 0$ and $y \mapsto 0$ for $y < 0$.)

1.5.7 What the section proved

For a function $f: X \rightarrow Y$, the section proved four facts and left one question open.

- A left inverse makes f injective (Theorem 1.5.1 (a)).
- If $X \neq \emptyset$, an injective f has a left inverse (Theorem 1.5.1 (b)). The empty function $\emptyset \rightarrow \{1\}$ shows that the hypothesis is needed.
- A right inverse makes f surjective (Proposition 1.5.2).
- A two-sided inverse exists exactly when f is bijective (Theorem 1.5.4), and it is unique (Proposition 1.5.3).
- **Open:** a surjective f has a right inverse. Chapter 3 proves it, with the new principle.

1.5.8 Exercises

1.5.8.1 A. Check your understanding

Exercise. A1

These parts are about Theorem 1.5.1 and its proof.

- At which line of the proof of (b) is $X \neq \emptyset$ used, and for what?
- The proof shows that the set G is a function by proving two things about the case split into $y \in f(X)$ and $y \notin f(X)$. What are they, and which step of the proof uses each?
- Run the construction on $f: \{1, 2, 3\} \rightarrow \{1, 2, 3, 4\}$, $f(k) = k + 1$, with $x_0 = 3$. Write out the left inverse g and check $g \circ f = \text{id}_{\{1, 2, 3\}}$.
- Give the idea of the proof of (b) in two sentences.
- Let $f: \{1, 2\} \rightarrow \{1\}$ be the only function between these sets. Which step of the proof of (b) fails for f , and why?

Solution.

- In the first line of (b): “fix one and call it x_0 ”. The point x_0 is the value $g(y)$ for every y outside the range $f(X)$; without it, the second alternative in the definition of G would have nothing to send such y to.
- The two cases **exhaust** every y : by excluded middle, $y \in f(X)$ or $y \notin f(X)$; this is used in the existence step. And they **do not overlap**: a y cannot satisfy both, since $y \in f(X)$ and $y \notin f(X)$ contradict each other; this is used in the uniqueness step, to rule out a pair from each alternative. (Uniqueness inside the first case uses injectivity; inside the second, it is automatic.)
- The hypotheses hold: $\{1, 2, 3\} \neq \emptyset$, and f is injective, since $k + 1 = k' + 1$ gives $k = k'$. The range is $f(\{1, 2, 3\}) = \{2, 3, 4\}$. On the range, g returns the unique preimage: $g(2) = 1$, $g(3) = 2$, $g(4) = 3$. The point 1 is outside the range, so $g(1) = x_0 = 3$.

Check: $g(f(1)) = g(2) = 1$, $g(f(2)) = g(3) = 2$, $g(f(3)) = g(4) = 3$. So $g \circ f = \text{id}_{\{1,2,3\}}$.

- (d) A point of the range has exactly one preimage, by injectivity, and g sends it there; a point outside the range is sent to a fixed point x_0 of the non-empty domain. Then $g(f(x))$ is the unique preimage of $f(x)$, which is x .
- (e) The uniqueness step fails. The point $y = 1$ lies in the range and has the two preimages 1 and 2, so $(1, 1)$ and $(1, 2)$ both lie in G , and G is not a function. The step “ $f(x) = y = f(x')$, so $x = x'$ ” needs injectivity, which this f lacks. (By part (a) of the theorem, f has no left inverse at all.)

1.5.8.2 B. Practice

Exercise. B1: Counting one-sided inverses

- (a) Find all left inverses of $f: \{1, 2\} \rightarrow \{1, 2, 3\}$, $f(k) = k$.
- (b) Find all right inverses of $g: \{1, 2, 3\} \rightarrow \{1, 2\}$, $g(1) = g(2) = 1$, $g(3) = 2$.

Solution.

- (a) A left inverse $h: \{1, 2, 3\} \rightarrow \{1, 2\}$ must satisfy $h(f(1)) = 1$ and $h(f(2)) = 2$, that is $h(1) = 1$ and $h(2) = 2$. The point 3 is outside the range $\{1, 2\}$, so $h(3)$ is free: 1 or 2. Each choice satisfies both equations. So there are exactly 2 left inverses: $h(3) = 1$ and $h(3) = 2$, with $h(1) = 1$, $h(2) = 2$ in both.
- (b) A right inverse $k: \{1, 2\} \rightarrow \{1, 2, 3\}$ must satisfy $g(k(1)) = 1$ and $g(k(2)) = 2$. The preimages of 1 under g are 1 and 2, so $k(1) \in \{1, 2\}$; the only preimage of 2 is 3, so $k(2) = 3$. So there are exactly 2 right inverses: $k(1) = 1$ or $k(1) = 2$, with $k(2) = 3$ in both.

Exercise. B2: The inverse of a composite

Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be bijections. Prove that $g \circ f$ is invertible and that $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Hint: check the two equations of Definition 1.5.2 for the candidate $f^{-1} \circ g^{-1}$.

Solution. The candidate $f^{-1} \circ g^{-1}: Z \rightarrow X$ is defined, since $g^{-1}: Z \rightarrow Y$ and $f^{-1}: Y \rightarrow X$ exist by Theorem 1.5.4. We check the two equations of Definition 1.5.2, using Proposition 1.4.3 to regroup and for its identity laws $k \circ \text{id} = k = \text{id} \circ k$:

$$\begin{aligned} (f^{-1} \circ g^{-1}) \circ (g \circ f) &= f^{-1} \circ (g^{-1} \circ g) \circ f = f^{-1} \circ \text{id}_Y \circ f = f^{-1} \circ f = \text{id}_X, \\ (g \circ f) \circ (f^{-1} \circ g^{-1}) &= g \circ (f \circ f^{-1}) \circ g^{-1} = g \circ \text{id}_Y \circ g^{-1} = g \circ g^{-1} = \text{id}_Z. \end{aligned}$$

So $f^{-1} \circ g^{-1}$ is a two-sided inverse of $g \circ f$. By the uniqueness of inverses (Proposition 1.5.3, as in Definition 1.5.2), it is **the** inverse: $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$. Notice the order: to undo “first f , then g ”, undo g first.

Exercise. B3: A bijection between an interval and the line

Let $h: (-1, 1) \rightarrow \mathbb{R}$, $h(x) = x/(1 - |x|)$, and $k: \mathbb{R} \rightarrow (-1, 1)$, $k(y) = y/(1 + |y|)$. Show that h and k are well defined (each denominator is non-zero, and k takes values in $(-1, 1)$), and that k is the inverse of h , by checking both composites by cases on the sign. Hence h is a bijection.

Solution. Well defined. For $x \in (-1, 1)$, $|x| < 1$, so $1 - |x| > 0$. For $y \in \mathbb{R}$, $1 + |y| \geq 1 > 0$, and $|k(y)| = |y|/(1 + |y|) < 1$ because $|y| < 1 + |y|$; so $k(y) \in (-1, 1)$.

First check, $k(h(x)) = x$ for $x \in (-1, 1)$. If $0 \leq x < 1$, then $h(x) = x/(1 - x) \geq 0$, so

$$k(h(x)) = \frac{x/(1 - x)}{1 + x/(1 - x)} = \frac{x/(1 - x)}{1/(1 - x)} = x.$$

If $-1 < x < 0$, then $|x| = -x$ and $h(x) = x/(1 + x) < 0$, so $|h(x)| = -x/(1 + x)$ and

$$k(h(x)) = \frac{x/(1 + x)}{1 - x/(1 + x)} = \frac{x/(1 + x)}{1/(1 + x)} = x.$$

Second check, $h(k(y)) = y$ for $y \in \mathbb{R}$. If $y \geq 0$, then $k(y) = y/(1 + y) \in [0, 1)$, so

$$h(k(y)) = \frac{y/(1 + y)}{1 - y/(1 + y)} = \frac{y/(1 + y)}{1/(1 + y)} = y.$$

If $y < 0$, then $k(y) = y/(1 - y) \in (-1, 0)$, so $|k(y)| = -y/(1 - y)$ and

$$h(k(y)) = \frac{y/(1 - y)}{1 + y/(1 - y)} = \frac{y/(1 - y)}{1/(1 - y)} = y.$$

Both composites are identities, so $k = h^{-1}$, and h is a bijection by [Theorem 1.5.4](#).

1.5.8.3 C. Going deeper**Exercise. C1: One composite equal to the identity**

Let $f: X \rightarrow Y$ and $g: Y \rightarrow X$ be functions with $g \circ f = \text{id}_X$.

- Prove that f is injective and g is surjective.
- Show by an example that neither f nor g need be bijective.
- Prove that if, in addition, f is surjective, then $g = f^{-1}$.

Solution (Hint 1). Read the equation $g \circ f = \text{id}_X$ twice: once as a statement about f , once as a statement about g .

Solution (Hint 2). For g : let $x \in X$; you need a point of Y that g sends to x .

Solution.

- (a) The equation says that g is a left inverse of f , so f is injective by [Theorem 1.5.1 \(a\)](#). The same equation says that f is a right inverse of g , so g is surjective by [Proposition 1.5.2](#). (Directly: for $x \in X$, the point $f(x) \in Y$ satisfies $g(f(x)) = x$.)
- (b) Let $f: \{1\} \rightarrow \{1, 2\}$, $f(1) = 1$, and $g: \{1, 2\} \rightarrow \{1\}$, $g(1) = g(2) = 1$. Then $g(f(1)) = 1$, so $g \circ f = \text{id}_{\{1\}}$. But f is not surjective (2 has no preimage) and g is not injective ($g(1) = g(2)$ with $1 \neq 2$), so neither is bijective.
- (c) If f is also surjective, it is bijective by (a), so f^{-1} exists by [Theorem 1.5.4](#), and it is in particular a right inverse of f . Also g is a left inverse of f . By [Proposition 1.5.3](#), $g = f^{-1}$.

Marking scheme. (10 points)

- 2: f injective, from the left-inverse equation (citing [Theorem 1.5.1 \(a\)](#) or arguing directly from $f(x_1) = f(x_2)$).
- 2: g surjective, either by citing [Proposition 1.5.2](#) (f is a right inverse of g) or directly, naming the witness $f(x)$ for a given $x \in X$.
- 3: an example with explicit sets and both failures checked (a point missed by f , two points identified by g).
- 3: (c) obtains a right inverse of f (from bijectivity) and applies uniqueness of one-sided inverses, or checks $f \circ g = \text{id}_Y$ directly (for $y = f(x)$, $f(g(y)) = f(g(f(x))) = f(x) = y$, using surjectivity to write $y = f(x)$).

Fatal errors (0 for the step): assuming $f \circ g = \text{id}_Y$ in (a) or (c) without proof; in (c), writing f^{-1} before showing f bijective.

Exercise. C2: When is a left inverse unique?

Let $f: X \rightarrow Y$ be injective, with $X \neq \emptyset$.

- (a) Prove that if f is surjective, or X has exactly one element, then f has exactly one left inverse.
- (b) Prove that if some $y_1 \in Y$ is not in $f(X)$, and X has two different elements $x_0 \neq x_1$, then f has at least two different left inverses.
- (c) Let $f: \{1, 2\} \rightarrow \{1, 2, 3, 4\}$, $f(k) = 2k$. List all left inverses of f , and explain in one sentence why there are $2^2 = 4$ of them.

Solution (Hint 1). For (a), a left inverse exists by [Theorem 1.5.1 \(b\)](#); only “at most one” needs proof. For (b), run the construction of that proof twice.

Solution (Hint 2). For (a), let g and g' be left inverses of f . If f is surjective, write a given $y \in Y$ as $y = f(x)$ and compare $g(y)$ with $g'(y)$.

Solution.

- (a) A left inverse exists by [Theorem 1.5.1](#) (b). For “at most one”, let g and g' be left inverses of f . *Case 1:* f is surjective. Then f is bijective ([Definition 1.4.5](#)), so it has the inverse f^{-1} ([Theorem 1.5.4](#)), which is in particular a right inverse of f . By [Proposition 1.5.3](#), $g = f^{-1}$ and $g' = f^{-1}$, so $g = g'$. *Case 2:* X has exactly one element x_0 . For every $y \in Y$, both $g(y)$ and $g'(y)$ are elements of X , so $g(y) = x_0 = g'(y)$; hence $g = g'$ by [Proposition 1.4.1](#). In either case any two left inverses are equal, so f has exactly one.
- (b) Run the proof of [Theorem 1.5.1](#) (b) once with the fixed point x_0 and once with x_1 ; call the resulting left inverses g_0 and g_1 . The point y_1 is outside the range, so the second alternative in the definition of G applies to it: $g_0(y_1) = x_0$ and $g_1(y_1) = x_1$. Since $x_0 \neq x_1$, the functions g_0 and g_1 differ at y_1 , so $g_0 \neq g_1$.
- (c) The range is $\{2, 4\}$. A left inverse $g: \{1, 2, 3, 4\} \rightarrow \{1, 2\}$ must satisfy $g(f(1)) = 1$ and $g(f(2)) = 2$, that is $g(2) = 1$ and $g(4) = 2$. The values $g(1)$ and $g(3)$ are free, each in $\{1, 2\}$, and every choice gives a left inverse. The four are

$$(g(1), g(2), g(3), g(4)) = (1, 1, 1, 2), (1, 1, 2, 2), (2, 1, 1, 2), (2, 1, 2, 2).$$

The values on the range are forced, and each of the two points 1, 3 outside the range may go to either of the two points of the domain, independently: $2 \cdot 2 = 4$ choices. This is the situation of (b): f is not surjective and its domain has two elements.

Marking scheme. (10 points)

- 1: (a) existence of a left inverse, citing [Theorem 1.5.1](#) (b).
- 2: (a) surjective case: an inverse exists, and uniqueness of one-sided inverses ([Proposition 1.5.3](#)) forces every left inverse to equal it (or a direct argument: for $y = f(x)$, $g(y) = x = g'(y)$).
- 2: (a) one-element case: any two functions into a one-element set agree at every point, so they are equal.
- 3: (b) two left inverses that take the values x_0 and x_1 at y_1 (by the construction or defined directly), each checked to be a left inverse, and shown to differ at y_1 .
- 2: (c) the forced values $g(2) = 1$, $g(4) = 2$, the four left inverses, and the count $2 \cdot 2$.

Fatal errors (0 for the step): in (a), writing f^{-1} before showing f bijective; in (b), concluding $g_0 \neq g_1$ without exhibiting a point where they differ.

1.6 Images and Preimages

A function $f: X \rightarrow Y$ sends points to points. Much of analysis asks what it does to **sets** of points. Where does f send an interval? Which inputs land in a given set of outputs? The first question is answered by the image $f(A)$, the second by the preimage $f^{-1}(B)$, and the two are the subject of this section. They look like mirror images of each other, and they are not: preimages respect every set operation of Section 1.2, while images respect only unions. The reason is one quantifier, and by the end of the section we will be able to point at it.

Throughout, $f: X \rightarrow Y$ is a function ([Definition 1.4.1](#)), and the unions and intersections of families are those of [Definition 1.2.4](#). An intersection of a family of subsets of Y is taken inside Y ; of subsets of X , inside X .

1.6.1 A question about squares

Let $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2$, and take the two intervals $A = [-1, 0]$ and $B = [0, 1]$. Squaring A gives every number from 0 to 1, and so does squaring B . So the set of squares of points of A and the set of squares of points of B are both $[0, 1]$, and the points they share make up $[0, 1]$.

Now intersect first. The intervals A and B share only the point 0, so $A \cap B = \{0\}$, and its only square is 0. Squaring the intersection gives $\{0\}$; intersecting the squares gives $[0, 1]$. The order of the two operations matters.

Try the same thing backwards. Instead of pushing sets forwards through f , pull them back: for a set C of outputs, collect the inputs whose squares lie in C . Take $C = [0, 1]$ and $D = [1, 4]$. The inputs with square in C form $[-1, 1]$, those with square in D form $[-2, -1] \cup [1, 2]$, and the inputs common to both are -1 and 1 . Intersecting first gives $C \cap D = \{1\}$, whose inputs are again -1 and 1 . This time the order does not matter.

Is that an accident of these intervals? No. Pulling back will turn out to respect intersections for every function and all sets, while pushing forwards respects them exactly for the injective functions. To see why, we first need names for the two operations.

1.6.2 Images and preimages

The two sets in the question above keep recurring: “all the values $f(a)$ for a in A ” and “all the inputs x with $f(x)$ in B ”. The first was already used in Section 1.4 for $A = X$, where it is the range of f .

An image hides a “there exists”; a preimage hides nothing.

Definition 1.6.1. Image

Let $f: X \rightarrow Y$ be a function and $A \subseteq X$. The **image** of A under f is the set

$$f(A) := \{y \in Y : \text{there exists } a \in A \text{ with } f(a) = y\}.$$

We also write it $\{f(a) : a \in A\}$. In symbols, for every y ,

$$y \in f(A) \iff y \in Y \wedge \exists a \in A \ f(a) = y.$$

Definition 1.6.2. Preimage

Let $f: X \rightarrow Y$ be a function and $B \subseteq Y$. The **preimage** of B under f , also called its **inverse image**, is the set

$$f^{-1}(B) := \{x \in X : f(x) \in B\}.$$

In symbols, for every x , $x \in f^{-1}(B) \iff x \in X \wedge f(x) \in B$. The preimage is defined for **every** function f , whether or not f has an inverse.

In Section 1.4 a point x with $f(x) = y$ was called a preimage of y ; that is, $x \in f^{-1}(\{y\})$. So the preimages of a point y are the elements of the preimage of the set $\{y\}$.

Read the two definitions clause by clause.

- **The ambient sets.** Both are sets by specification (Definition 1.1.3): $f(A)$ is carved out of the codomain Y , and $f^{-1}(B)$ out of the domain X . So $f(A) \subseteq Y$ and $f^{-1}(B) \subseteq X$ always. The shorthand $\{f(a) : a \in A\}$ names the same subset of Y ; it is a way of writing, not a new rule of set formation.
- **The image.** “ $y \in f(A)$ ” says that y is **hit** by some point of A . The words “there exists” are part of the definition. To **use** a line $y \in f(A)$, we must name a witness: “let $a \in A$ with $f(a) = y$ ”. That is the rule $\exists E$ of *Logic and Proof*, Section 4.6, and the name must be fresh. To **prove** $y \in f(A)$, we exhibit a witness, by $\exists I$ (*Logic and Proof*, Section 4.5).
- **The preimage.** “ $x \in f^{-1}(B)$ ” says that x is sent **into** B . There is no quantifier. For $x \in X$, a line $x \in f^{-1}(B)$ is used and proved by one substitution: it is exchanged for $f(x) \in B$, and back.
- **The symbol f^{-1} .** In $f^{-1}(B)$, the symbol f^{-1} is not a function that is applied to B . The whole expression $f^{-1}(B)$ is the name of a set built from f and B . Section 1.5 used f^{-1} for the inverse function of a bijection; the remark at the end of this section reconciles the two readings.

The notation $f(A)$ puts a set where $f(x)$ puts a point. When the letter in the brackets has been declared a subset of X , the expression is an image; when it has been declared a point, it is a value. As with the pair and the interval of Section 1.3, the declaration decides.

Start with the smallest possible example, where everything can be listed.

Example. Images and preimages by listing

Let $X = \{1, 2, 3, 4\}$, $Y = \{0, 1, 2, 5\}$, and let $f: X \rightarrow Y$ be given by

$$f(1) = 1, \quad f(2) = 2, \quad f(3) = 0, \quad f(4) = 1.$$

Find $f(\{1, 4\})$, $f(\{2, 3\})$, $f(X)$, and $f(\emptyset)$; then $f^{-1}(\{1\})$, $f^{-1}(\{0, 2\})$, $f^{-1}(\{5\})$, $f^{-1}(\emptyset)$, and $f^{-1}(Y)$.

Solution. Images. A point y lies in $f(A)$ when some $a \in A$ has $f(a) = y$, so we apply f to each point of A and collect the results. $f(\{1, 4\}) = \{1\}$: both points go to 1, and a set lists 1 once. $f(\{2, 3\}) = \{0, 2\}$. $f(X) = \{0, 1, 2\}$, the range; 5 is not in it. $f(\emptyset) = \emptyset$: a point y in it would need some $a \in \emptyset$, and there is none.

Preimages. A point $x \in X$ lies in $f^{-1}(B)$ when $f(x) \in B$, so we test each of the four points of X . $f^{-1}(\{1\}) = \{1, 4\}$. $f^{-1}(\{0, 2\}) = \{2, 3\}$. $f^{-1}(\{5\}) = \emptyset$: no point of X is sent to 5. $f^{-1}(\emptyset) = \emptyset$, since $f(x) \in \emptyset$ never holds. $f^{-1}(Y) = X$, since $f(x) \in Y$ holds for every $x \in X$.

The two computations differ in shape. For the image we ran through the points of A and **produced** outputs. For the preimage we ran through the points of X and **tested** each one. That is the “there exists” of the image and the plain condition of the preimage, seen on a four-point set.

The example also contains a non-example by minimal change. The sets $\{1\}$ and $\{5\}$ are both one-point subsets of Y . The first has a two-point preimage; change 1 to 5 and the preimage becomes empty, because 5 lies outside the range. The preimage of a non-empty set can be empty. What fails is the existence of a point of X sent into the set.

Why is the preimage defined for **every** function, when f^{-1} as a function exists only for bijections?

Because the question “which inputs land in B ?” makes sense for any f : its answer may be empty, or contain many points, and neither is a problem for a set. The function f^{-1} of Section 1.5 asks for more: exactly one input for each output. This is why analysis states its definitions with preimages. “The preimage of every open set is open” makes sense for every function; *Metric and Topological Spaces*, Chapter 1 uses it to characterize continuity.

Warning. The preimage $f^{-1}(B)$ does not presuppose that f is invertible. For the squaring map on $\mathbb{R}$, which has no inverse function, $f^{-1}(\{4\}) = \{-2, 2\}$ is a perfectly good set. What the symbol does **not** give is a value “ $f^{-1}(4)$ ”: that would be a point, and there are two candidates.

1.6.3 Computing images and preimages

With an infinite set, listing is impossible, and the definitions are used as they stand. The image asks us to name witnesses; the preimage asks us to solve an inequality.

Example. Images and preimages of intervals under squaring

Let $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2$. Prove that

- (a) $f([-1, 2]) = [0, 4]$;
- (b) $f^{-1}([1, 4]) = [-2, -1] \cup [1, 2]$;
- (c) $f^{-1}(\{-1\}) = \emptyset$;
- (d) $f^{-1}(f([0, 1])) = [-1, 1]$.

Solution. We use school facts about real numbers, $x^2 \geq 0$ and every $y \geq 0$ has a square root $\sqrt{y} \geq 0$, and, by Logic and Proof, Theorem 2.4.5, for $s, t \geq 0$, $s \leq t$ holds exactly when $s^2 \leq t^2$.

- (a) *Idea.* An image is proved equal to a set by two inclusions. The inclusion $\subseteq$ uses $y \in f([-1, 2])$, so it names a witness. The inclusion $\supseteq$ proves $y \in f([-1, 2])$, so it exhibits one.

($\subseteq$) Let $y \in f([-1, 2])$. Then $y = a^2$ for some $a \in [-1, 2]$; name it a . Certainly $y = a^2 \geq 0$. If $a \geq 0$, then $0 \leq a \leq 2$, so $a^2 \leq 4$. If $a < 0$, then $0 < -a \leq 1$, so $a^2 = (-a)^2 \leq 1 \leq 4$. In either case $0 \leq y \leq 4$, that is, $y \in [0, 4]$.

($\supseteq$) Let $y \in [0, 4]$. Take $a := \sqrt{y}$. Then $0 \leq a \leq \sqrt{4} = 2$, so $a \in [-1, 2]$, and $f(a) = y$. Hence $y \in f([-1, 2])$.

The witness in ($\supseteq$) could have been $-\sqrt{y}$ for $y \leq 1$, but not for $y > 1$, where $-\sqrt{y} < -1$. One witness per y is all we need, so we choose the one that always works.

- (b) *Idea.* A preimage is a set of solutions of a condition on x . No witness is needed: each line is an equivalence.

For every $x \in \mathbb{R}$,

$$\begin{aligned} x \in f^{-1}([1, 4]) &\iff 1 \leq x^2 \leq 4 \iff 1 \leq |x| \leq 2 \\ &\iff x \in [-2, -1] \cup [1, 2]. \end{aligned}$$

The first step is [Definition 1.6.2](#). The second uses $x^2 = |x|^2$ and Logic and Proof, Theorem 2.4.5 with s, t among 1, $|x|$, 2. The third splits by the sign of x : for $x \geq 0$,

$|x| = x$, and for $x < 0$, $|x| = -x$. So the two sets have the same elements.

- (c) For every $x \in \mathbb{R}$, $f(x) = x^2 \geq 0 > -1$, so $f(x) \in \{-1\}$ never holds. The preimage is empty, although $\{-1\}$ is not.
- (d) First $f([0, 1]) = [0, 1]$, by the argument of (a): a square of a number in $[0, 1]$ lies in $[0, 1]$, and each $y \in [0, 1]$ is $(\sqrt{y})^2$ with $\sqrt{y} \in [0, 1]$. Then, as in (b), for every $x \in \mathbb{R}$, $x \in f^{-1}([0, 1]) \iff 0 \leq x^2 \leq 1 \iff |x| \leq 1 \iff x \in [-1, 1]$.

Part (d) is worth a second look. We started with $[0, 1]$, pushed it forwards and pulled it back, and got the bigger set $[-1, 1]$. The negative numbers came in because they have the same squares as the positive ones. That is a failure of injectivity, and the subsection “There and back” later in this section proves that it is the only way the round trip can grow a set.

Quick check. For $f(x) = x^2$ on $\mathbb{R}$, find $f^{-1}(\{4\})$, $f^{-1}([-4, -1])$, and $f(f^{-1}([-1, 1]))$.

Solution. $f^{-1}(\{4\}) = \{-2, 2\}$, the solutions of $x^2 = 4$. $f^{-1}([-4, -1]) = \emptyset$, since $x^2 \geq 0 > -1$. For the last, $f^{-1}([-1, 1]) = \{x \in \mathbb{R} : -1 \leq x^2 \leq 1\} = [-1, 1]$, and then $f([-1, 1]) = [0, 1]$, by the argument of [Example \(Images and preimages of intervals under squaring\)](#) (a). So $f(f^{-1}([-1, 1])) = [0, 1]$, which is smaller than $[-1, 1]$: the negative numbers are not squares.

1.6.4 Sets of numbers as images

Analysis on the Real Line, Chapter 1 works with three ways of making new sets of real numbers from old ones: for $A, B \subseteq \mathbb{R}$ and $c \in \mathbb{R}$,

$$-A := \{-a : a \in A\}, \quad cA := \{ca : a \in A\}, \quad A + B := \{a + b : a \in A, b \in B\}.$$

The first two are images in the sense of [Definition 1.6.1](#). In the third, two letters vary, and the notation means, as for [Definition 1.3.2](#),

$$A + B = \{y \in \mathbb{R} : \exists a \in A \exists b \in B \ y = a + b\}.$$

Each is an image, and seeing them as images tells us how to prove things about them: an element of $A + B$ is used by naming two witnesses.

Example. Reflections, scalings and sums as images

Let $A, B \subseteq \mathbb{R}$ and $c \in \mathbb{R}$.

- (a) Let $n: \mathbb{R} \rightarrow \mathbb{R}$, $n(x) = -x$, and $m_c: \mathbb{R} \rightarrow \mathbb{R}$, $m_c(x) = cx$. Show that $-A = n(A)$ and $cA = m_c(A)$.
- (b) Let $s: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, $s(a, b) = a + b$. Show that $A + B = s(A \times B)$.
- (c) Compute $-[0, 1]$, $\{1, 2\} + \{10, 20\}$, and $\{1, 2\} + \{1, 2\}$.

Solution.

- (a) By the shorthand of [Definition 1.6.1](#), $n(A) = \{n(a) : a \in A\} = \{-a : a \in A\}$, which is $-A$ as written above. In the same way, $m_c(A) = \{ca : a \in A\} = cA$. Nothing more is needed: the set-builder $\{-a : a \in A\}$ is the image, written in the form of

Definition 1.6.1.

- (b) Here the domain is the product $\mathbb{R} \times \mathbb{R}$ of Section 1.3, and $A \times B \subseteq \mathbb{R} \times \mathbb{R}$. An element p of $A \times B$ is a pair (a, b) with $a \in A$ and $b \in B$ (Definition 1.3.2).
 $(\subseteq)$ Let $y \in A + B$. Then $y = a + b$ for some $a \in A$ and $b \in B$; name them. The pair $p := (a, b)$ lies in $A \times B$, and $s(p) = a + b = y$. So $y \in s(A \times B)$, with witness p .
 $(\supseteq)$ Let $y \in s(A \times B)$. Then $y = s(p)$ for some $p \in A \times B$; name it p . Since $p \in A \times B$, $p = (a, b)$ for some $a \in A$ and $b \in B$; name them. Then $y = s(a, b) = a + b$, so $y \in A + B$.
 Unwinding $y \in s(A \times B)$ took **two** layers of witnesses: first the pair p from the image, then its coordinates a, b from the product.
- (c) $-[0, 1] = [-1, 0]$: if $y = -a$ with $0 \leq a \leq 1$, then $-1 \leq y \leq 0$; and if $-1 \leq y \leq 0$, then $a := -y \in [0, 1]$ and $y = -a$. Next, $\{1, 2\} + \{10, 20\}$ is the image of the four pairs $(1, 10), (1, 20), (2, 10), (2, 20)$, so it is $\{11, 21, 12, 22\}$, with 4 elements. Finally, $\{1, 2\} + \{1, 2\}$ is the image of $(1, 1), (1, 2), (2, 1), (2, 2)$, which is $\{2, 3, 4\}$, with only 3 elements: the pairs $(1, 2)$ and $(2, 1)$ have the same image 3, because s is not injective.

This pays a debt. *Analysis on the Real Line*, Chapter 1 introduces these sets by set-builder, and part (b) shows that even the two-variable form is an image, of $A \times B$ under s , so Definition 1.6.1 is what all three rest on: an element of $A + B$ is used by naming two witnesses, as in part (b).

1.6.5 Preimages respect every operation

Now the promised theorem. The preimage of a union is the union of the preimages, and the same holds for intersections, complements, and differences. No hypothesis on f is needed.

Theorem 1.6.1. Preimages of Set Operations

Let $f: X \rightarrow Y$ be a function.

- (a) For every family $(B_i)_{i \in I}$ of subsets of Y , $f^{-1}\left(\bigcup_{i \in I} B_i\right) = \bigcup_{i \in I} f^{-1}(B_i)$.
 (b) For every family $(B_i)_{i \in I}$ of subsets of Y , $f^{-1}\left(\bigcap_{i \in I} B_i\right) = \bigcap_{i \in I} f^{-1}(B_i)$, the intersection on the left taken in Y and on the right in X .
 (c) For every $B \subseteq Y$, $f^{-1}(Y \setminus B) = X \setminus f^{-1}(B)$.
 (d) For all $B, C \subseteq Y$, $f^{-1}(B \setminus C) = f^{-1}(B) \setminus f^{-1}(C)$.

Idea. Unwind each side at a point $x \in X$. A preimage turns “ x is in the preimage of something” into “ $f(x)$ is in that something”. The set operation is then applied to **one** point, $f(x)$, and on the other side it is applied to one point, x , whose membership in each $f^{-1}(B_i)$ is again a statement about $f(x)$. The two sides say the same thing about the same point. No point is ever named as a witness for an image. The only “there exists” is over the index i in (a), and its witness is handed straight from $\exists E$ to $\exists I$, so no witness can go wrong, and nothing is asked of f .

Proof. Every set in the four laws is a subset of X : a preimage is one by Definition 1.6.2, a union of subsets of X is one because each of its elements lies in some $f^{-1}(B_i) \subseteq X$, an inter-

section in X and a difference $X \setminus \cdot$ are subsets of X by [Definition 1.2.4](#) and [Definition 1.2.1](#), and $f^{-1}(B) \setminus f^{-1}(C) \subseteq f^{-1}(B)$. For two subsets S, T of X , a point outside X lies in neither, so by [Definition 1.1.1](#), $S = T$ holds as soon as $x \in S \iff x \in T$ for every $x \in X$. Let $x \in X$. In each chain below, the steps labeled (P) are [Definition 1.6.2](#), which applies because $x \in X$; the others are named.

(a)

$$\begin{aligned}
 x \in f^{-1}\left(\bigcup_{i \in I} B_i\right) &\iff f(x) \in \bigcup_{i \in I} B_i && \text{(P)} \\
 &\iff \exists i \in I, f(x) \in B_i && \text{(union)} \\
 &\iff \exists i \in I, x \in f^{-1}(B_i) && \text{(P)} \\
 &\iff x \in \bigcup_{i \in I} f^{-1}(B_i) && \text{(union).}
 \end{aligned}$$

The steps marked “union” are [Definition 1.2.4](#). In the third step, for each fixed $i \in I$ the statements $f(x) \in B_i$ and $x \in f^{-1}(B_i)$ are equivalent, so one holds for some i exactly when the other does: a witness i for one is a witness for the other.

(b)

$$\begin{aligned}
 x \in f^{-1}\left(\bigcap_{i \in I} B_i\right) &\iff f(x) \in \bigcap_{i \in I} B_i && \text{(P)} \\
 &\iff \forall i \in I, f(x) \in B_i && \text{(intersection)} \\
 &\iff \forall i \in I, x \in f^{-1}(B_i) && \text{(P)} \\
 &\iff x \in \bigcap_{i \in I} f^{-1}(B_i) && \text{(intersection).}
 \end{aligned}$$

In the second step, the intersection in Y is $\{y \in Y : \forall i \in I, y \in B_i\}$, and the clause $f(x) \in Y$ holds because f maps into Y . In the last step, the intersection in X is $\{x' \in X : \forall i \in I, x' \in f^{-1}(B_i)\}$, and the clause $x \in X$ holds by assumption. The third step holds for each i , hence for all i at once.

(c)

$$\begin{aligned}
 x \in f^{-1}(Y \setminus B) &\iff f(x) \in Y \setminus B && \text{(P)} \\
 &\iff f(x) \notin B && (f(x) \in Y) \\
 &\iff x \notin f^{-1}(B) && \text{(P, negated)} \\
 &\iff x \in X \setminus f^{-1}(B) && (x \in X).
 \end{aligned}$$

The second and fourth steps are [Definition 1.2.1](#) for $\setminus$, with the clause that is known to hold dropped. The third negates both sides of the equivalence $x \in f^{-1}(B) \iff f(x) \in B$, as Logic and Proof, Example (Negating both sides of an iff) allows.

(d)

$$\begin{aligned}
 x \in f^{-1}(B \setminus C) &\iff f(x) \in B \wedge f(x) \notin C \\
 &\iff x \in f^{-1}(B) \wedge x \notin f^{-1}(C) \\
 &\iff x \in f^{-1}(B) \setminus f^{-1}(C),
 \end{aligned}$$

by (P) and [Definition 1.2.1](#) in the first step, (P) and Logic and Proof, Example (Negating both sides of an iff) in the second, and [Definition 1.2.1](#) in the third.

In each chain the equivalences link up, so the first statement is equivalent to the last, for every $x \in X$. This proves the theorem. ■

The proof never named a witness for an image (the index i in (a) is taken by $\exists E$ and passed unchanged to $\exists I$) and never used a property of f beyond its being a function. That is the whole content of the slogan: a preimage hides nothing. Each law of sets was carried across by the corresponding law of logic, applied to the one point $f(x)$, exactly as in Section 1.2.

The empty family is covered. When $I = \emptyset$, the intersection of the empty family in Y is Y and in X is X (Definition 1.2.4), and (b) says $f^{-1}(Y) = X$, which is true. The union of the empty family is $\emptyset$, and (a) says $f^{-1}(\emptyset) = \emptyset$, also true. This is the payoff of taking intersections inside a named ambient set: the theorem holds with no exception for $I = \emptyset$.

Preimages also behave well under composition, with the order reversed. We will need this in *Metric and Topological Spaces*, Chapter 1, to show that composites of continuous maps are continuous.

Proposition 1.6.2. Preimage under a Composite

Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be functions, and $C \subseteq Z$. Then

$$(g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C)).$$

Proof. Both sides are subsets of X by Definition 1.6.2, so, as in the proof of Theorem 1.6.1, it suffices to compare memberships of each $x \in X$. For $x \in X$,

$$\begin{aligned} x \in (g \circ f)^{-1}(C) &\iff g(f(x)) \in C \iff f(x) \in g^{-1}(C) \\ &\iff x \in f^{-1}(g^{-1}(C)). \end{aligned}$$

The first step is Definition 1.6.2 for $g \circ f$ with Definition 1.4.2; the second is Definition 1.6.2 for g , which applies because $f(x) \in Y$; the third is Definition 1.6.2 for f . This proves the proposition. ■

To pull a set back through “first f , then g ”, pull it back through g first. This is the same reversal as for the inverse of a composite (Exercise B2 of Section 1.5), and here no function needs an inverse.

1.6.6 Images respect unions

Now push forwards. For unions the result is the same as for preimages, but the proof is different in kind: it has to name witnesses.

Theorem 1.6.3. Image of a Union

Let $f: X \rightarrow Y$ be a function, and let $(A_i)_{i \in I}$ be a family of subsets of X . Then

$$f\left(\bigcup_{i \in I} A_i\right) = \bigcup_{i \in I} f(A_i).$$

Idea. Unwind the left side at a point y : there is a point a in the union with $f(a) = y$, and being in the union means there is an index i with $a \in A_i$. So “there is a , and there is

i”. The right side says “there is i , and there is a ”. The two differ only in the order of two “there exists”, and two existential quantifiers can be exchanged: name both witnesses, then reintroduce them so that the quantifiers come out in the other order.

Proof. We prove two inclusions. The union $\bigcup_{i \in I} A_i$ is a subset of X , so its image is defined.

($\subseteq$) Let $y \in f(\bigcup_{i \in I} A_i)$. By Definition 1.6.1, $y \in Y$ and there is $a \in \bigcup_{i \in I} A_i$ with $f(a) = y$; name such a point a [$\exists E$]. By Definition 1.2.4, there is $j \in I$ with $a \in A_j$; name such an index j [$\exists E$]. Now $a \in A_j$ and $f(a) = y$, so $y \in f(A_j)$, with witness a [$\exists I$]. Since $j \in I$, $y \in \bigcup_{i \in I} f(A_i)$, with witness j [$\exists I$].

($\supseteq$) Let $y \in \bigcup_{i \in I} f(A_i)$. There is $j \in I$ with $y \in f(A_j)$; name it j [$\exists E$]. So $y \in Y$, and there is $a \in A_j$ with $f(a) = y$; name it a [$\exists E$]. Since $a \in A_j$ and $j \in I$, $a \in \bigcup_{i \in I} A_i$, with witness j [$\exists I$]. Since $f(a) = y$, $y \in f(\bigcup_{i \in I} A_i)$, with witness a [$\exists I$].

This proves the theorem. ■

Look at the order of the tags in ($\subseteq$): $\exists E$ for a , $\exists E$ for j , $\exists I$ for a , $\exists I$ for j . The witnesses are reintroduced in the same order they were named. Because $\exists I$ builds the formula from the inside out, a , which was the outer witness, ends up as the inner one. To rebuild the quantifiers in their old order, the $\exists I$ steps would have run in reverse, $\exists I$ for j before $\exists I$ for a . That the order is not reversed is the exchange of the two “there exists”. The two $\exists E$ boxes may close at the end because the conclusion $y \in \bigcup_i f(A_i)$ mentions neither a nor j : the side condition of $\exists E$ (*Logic and Proof*, Section 4.6) holds. The case $I = \emptyset$ needs no separate treatment: both sides are $f(\emptyset) = \emptyset$ and $\emptyset$.

1.6.7 Images and intersections

For unions, images behave. For intersections, the question of the opening showed they do not: $f(A \cap B) = \{0\}$ and $f(A) \cap f(B) = [0, 1]$ for the squaring map. One inclusion survives, and the other fails in a way that is instructive, because the failing inclusion has a tempting false proof.

Example. Two witnesses under one name

Here is a “proof” that $f(A) \cap f(B) \subseteq f(A \cap B)$ for every function $f: X \rightarrow Y$ and all $A, B \subseteq X$.

- (1) Let $y \in f(A) \cap f(B)$. (2) Then $y \in f(A)$, so $y = f(x)$ for some $x \in A$. (3) Also $y \in f(B)$, so $y = f(x)$ for some $x \in B$. (4) Hence $x \in A$ and $x \in B$, so $x \in A \cap B$. (5) Therefore $y = f(x) \in f(A \cap B)$.

Show that the conclusion is false, and find the line that fails.

Solution. The conclusion is false. Take $f(x) = x^2$ on $\mathbb{R}$, $A = \{-1\}$, and $B = \{1\}$. Then $f(A) = f(B) = \{1\}$, so $f(A) \cap f(B) = \{1\}$; but $A \cap B = \emptyset$, so $f(A \cap B) = \emptyset$, and $1 \notin \emptyset$.

The failing line is (3). Line (2) is a correct $\exists E$: it names a witness $x \in A$ with $f(x) = y$. Line (3) performs a second $\exists E$, on $y \in f(B)$, and gives its witness the **same** name x . The

rule requires a fresh name (*Logic and Proof*, Section 4.6). The statement “ $y = f(b)$ for some $b \in B$ ” promises some point of B , not the point already named. Reusing the name silently asserts that the two witnesses are equal, and nothing says they are. This is the error “one name for two witnesses”, entry (N4) of the catalogue of what is not allowed (*Logic and Proof*, Section 5.6).

A counterexample to the step. With the data above and $y = 1$: line (2) names $x = -1$, the only point of A . The only point of B with square 1 is $1 \neq -1$, so line (3), with the name x meaning the x of line (2), asserts $-1 \in B$, which is false. Lines (4) and (5) are correct inferences from what (3) wrongly asserts.

Warning. When you unwind two images, give the two witnesses two names: “ $y = f(a)$ with $a \in A$, and $y = f(b)$ with $b \in B$ ”. A single name is a hidden claim that $a = b$, which is exactly what fails for $x \mapsto x^2$ with $a = -1$, $b = 1$. If the proof needs them equal, it must prove it.

With two names, the proof gets as far as $f(a) = y = f(b)$ and stops. To conclude $a = b$ we need exactly the property that injectivity supplies. So the inclusion should hold for injective functions, and only for them.

Theorem 1.6.4. Image of an Intersection

Let $f: X \rightarrow Y$ be a function.

- (a) For all $A, B \subseteq X$, $f(A \cap B) \subseteq f(A) \cap f(B)$.
- (b) The equality $f(A \cap B) = f(A) \cap f(B)$ holds for all $A, B \subseteq X$ **if and only if** f is injective.

Try it first (ten minutes): if f is injective, why should $f(A) \cap f(B) \subseteq f(A \cap B)$ hold? Start from “let $y \in f(A) \cap f(B)$ ” and unwind both images, with two names.

Idea. The attempt names $a \in A$ and $b \in B$ with $f(a) = y = f(b)$, and then has two points where it needs one in $A \cap B$. That is the obstacle [Example \(Two witnesses under one name\)](#) tried to wish away. Injectivity removes it: $f(a) = f(b)$ gives $a = b$, so a lies in both sets.

For the converse, run the failure backwards. If equality holds for **all** A, B , use it on the smallest sets that could show a failure of injectivity: two one-point sets $\{x_1\}$ and $\{x_2\}$ with $f(x_1) = f(x_2)$.

Before reading on: which line of the proof of $(\Leftarrow)$ uses injectivity, and which line of $(\Rightarrow)$ uses that equality holds for all sets rather than for one pair?

Proof.

- (a) Let $y \in f(A \cap B)$. By [Definition 1.6.1](#) there is $c \in A \cap B$ with $f(c) = y$; name it c . Since $c \in A$, $y \in f(A)$, with witness c ; since $c \in B$, $y \in f(B)$, with witness c . So $y \in f(A) \cap f(B)$.

(b) ($\Leftarrow$) Suppose f is injective, and let $A, B \subseteq X$. By (a) it suffices to prove $f(A) \cap f(B) \subseteq f(A \cap B)$. Let $y \in f(A) \cap f(B)$. Since $y \in f(A)$, there is $a \in A$ with $f(a) = y$; name it a $[\exists E]$. Since $y \in f(B)$, there is $b \in B$ with $f(b) = y$; name it b , a new name $[\exists E]$. Then $f(a) = y = f(b)$, so $a = b$ because f is injective (Definition 1.4.3). Hence $a \in B$, so $a \in A \cap B$, and $y = f(a) \in f(A \cap B)$, with witness a $[\exists I]$.

($\Rightarrow$) Suppose $f(A \cap B) = f(A) \cap f(B)$ for all $A, B \subseteq X$. Let $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$, and apply the hypothesis to $A := \{x_1\}$ and $B := \{x_2\}$. Now $f(x_1) \in f(A)$, with witness x_1 , and $f(x_1) = f(x_2) \in f(B)$, with witness x_2 . So $f(x_1) \in f(A) \cap f(B) = f(A \cap B)$. Hence there is $c \in \{x_1\} \cap \{x_2\}$ with $f(c) = f(x_1)$; name it c . Then $c = x_1$ and $c = x_2$, so $x_1 = x_2$. Since x_1, x_2 were arbitrary, f is injective. This proves the theorem. ■

The proof of ($\Rightarrow$) is direct: it proves $x_1 = x_2$ by pushing the one-point sets through the hypothesis, without supposing $x_1 \neq x_2$. Read contrapositively, the same sets give the witness of failure: whenever $f(x_1) = f(x_2)$ with $x_1 \neq x_2$, the sets $\{x_1\}$ and $\{x_2\}$ have $f(\{x_1\} \cap \{x_2\}) = f(\emptyset) = \emptyset$ while $f(\{x_1\}) \cap f(\{x_2\}) = \{f(x_1)\}$. For the squaring map, $x_1 = -1$ and $x_2 = 1$ is such a pair, and the intervals $[-1, 0]$ and $[0, 1]$ of the opening are a larger failure of the same kind.

Compare with Example (Two witnesses under one name). The false proof and the true one begin identically. The false one reuses a name and thereby assumes $a = b$; the true one keeps two names and **proves** $a = b$, from injectivity, on the one line where the hypothesis is used. That line is the whole difference between a theorem and a false statement.

1.6.8 There and back

What happens when we push a set forwards and pull it back, or the other way round? Part (d) of Example (Images and preimages of intervals under squaring) and the quick check after it gave one answer each for the squaring map: $f^{-1}(f([0, 1])) = [-1, 1]$ is bigger than $[0, 1]$, and $f(f^{-1}([-1, 1])) = [0, 1]$ is smaller than $[-1, 1]$. In general, the round trip through Y can only grow a set, and the round trip through X can only shrink one.

Proposition 1.6.5. Image and Preimage Round Trips

Let $f: X \rightarrow Y$ be a function.

- (a) $A \subseteq f^{-1}(f(A))$ for every $A \subseteq X$. Equality holds for all $A \subseteq X$ if and only if f is injective.
- (b) $f(f^{-1}(B)) = B \cap f(X)$ for every $B \subseteq Y$; in particular $f(f^{-1}(B)) \subseteq B$. Equality $f(f^{-1}(B)) = B$ holds for all $B \subseteq Y$ if and only if f is surjective.

Idea. For (a), a point $a \in A$ is sent into $f(A)$, with itself as witness, so it lies in the preimage. A point x of the preimage has $f(x) = f(a)$ for some $a \in A$, and it lies in A when $x = a$: that is injectivity again. For (b), a point of $f(f^{-1}(B))$ is $f(x)$ with $f(x) \in B$: so it lies in B , and it is a value of f . Those two conditions are all there is to it.

Proof.

- (a) Let $A \subseteq X$ and $a \in A$. Then $a \in X$, and $f(a) \in f(A)$, with witness a . So $a \in f^{-1}(f(A))$ by [Definition 1.6.2](#). This proves $A \subseteq f^{-1}(f(A))$.
 $(\Leftarrow)$ Suppose f is injective, and let $x \in f^{-1}(f(A))$. Then $x \in X$ and $f(x) \in f(A)$, so there is $a \in A$ with $f(a) = f(x)$; name it a $[\exists E]$. Since f is injective ([Definition 1.4.3](#)), $a = x$, so $x \in A$. With the inclusion already proved, $A = f^{-1}(f(A))$ by [Proposition 1.1.1](#).
 $(\Rightarrow)$ Suppose $A = f^{-1}(f(A))$ for every $A \subseteq X$. Let $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$. Then $f(x_2) = f(x_1) \in f(\{x_1\})$, with witness x_1 , so $x_2 \in f^{-1}(f(\{x_1\}))$ by [Definition 1.6.2](#). The hypothesis with $A = \{x_1\}$, a subset of X because $x_1 \in X$, gives $f^{-1}(f(\{x_1\})) = \{x_1\}$, so $x_2 \in \{x_1\}$. Hence $x_2 = x_1$, and f is injective.
- (b) Let $B \subseteq Y$. $(\subseteq)$ Let $y \in f(f^{-1}(B))$. There is $x \in f^{-1}(B)$ with $f(x) = y$; name it x . Then $x \in X$ and $f(x) \in B$, so $y \in B$; and $y \in f(X)$, with witness x . $(\supseteq)$ Let $y \in B \cap f(X)$. Since $y \in f(X)$, there is $x \in X$ with $f(x) = y$; name it x . Then $f(x) = y \in B$, so $x \in f^{-1}(B)$, and $y \in f(f^{-1}(B))$, with witness x . This proves $f(f^{-1}(B)) = B \cap f(X)$, and $B \cap f(X) \subseteq B$.

For the last statement: if f is surjective ([Definition 1.4.4](#)), every $y \in Y$ lies in $f(X)$; since $f(X) \subseteq Y$ by [Definition 1.6.1](#), $f(X) = Y$ by [Proposition 1.1.1](#), and $B \cap f(X) = B \cap Y = B$ for every $B \subseteq Y$, because $B \subseteq Y$. Conversely, if $f(f^{-1}(B)) = B$ for every $B \subseteq Y$, take $B = Y$: since $f^{-1}(Y) = X$, this says $f(X) = Y$, so every $y \in Y$ is $f(x)$ for some $x \in X$ ([Definition 1.6.1](#)), and f is surjective ([Definition 1.4.4](#)). This proves the proposition. ■

So the two round trips measure the two failures of Section 1.4. Pushing forwards and pulling back adds to A the points that share a value with a point of A : nothing is added, for every A , exactly when f is injective. Pulling back and pushing forwards removes from B the points that are never hit: nothing is removed, for every B , exactly when f is surjective.

1.6.9 Two readings of the inverse symbol

Remark. The symbol f^{-1} now has two uses. For **every** function $f: X \rightarrow Y$ and every $B \subseteq Y$, $f^{-1}(B)$ is the preimage of [Definition 1.6.2](#). For a **bijection** f , $f^{-1}: Y \rightarrow X$ is also the inverse function of [Definition 1.5.2](#), and then $f^{-1}(B)$ could also be read as the image of B under that function. The two readings give the same set. Write g for the inverse function; both sets are subsets of X . If $x = g(b)$ with $b \in B$, then $f(x) = f(g(b)) = b \in B$, since $f \circ g = \text{id}_Y$, so x lies in the preimage. If $x \in X$ and $f(x) \in B$, then $x = g(f(x)) \in g(B)$, since $g \circ f = \text{id}_X$. So the two readings name the same set, and the notation is never ambiguous. When f is not a bijection only the first reading exists.

For example, $a(x) = 3x - 2$ on $\mathbb{R}$ is a bijection with inverse $a^{-1}(y) = (y + 2)/3$ (Section 1.5). The preimage $a^{-1}([1, 4])$ is $\{x \in \mathbb{R} : 1 \leq 3x - 2 \leq 4\} = [1, 2]$, and the image of $[1, 4]$ under the inverse function is $\{(y + 2)/3 : y \in [1, 4]\} = [1, 2]$ as well.

1.6.10 Exercises

1.6.10.1 A. Check your understanding

Exercise. A1

Parts (a)–(d) are about [Theorem 1.6.1](#) and its proof.

- (a) Suppose we tried to prove the image version of law (b), $f(\bigcap_{i \in I} A_i) = \bigcap_{i \in I} f(A_i)$, by the same kind of chain. At which step would a witness have to be named, and why does the preimage version need none?
- (b) Run law (b) on $f(x) = x^2$ on $\mathbb{R}$, with $I = \{1, 2\}$, $B_1 = [0, 4]$, and $B_2 = [1, 9]$: compute both sides.
- (c) What are both sides of law (b) when $I = \emptyset$?
- (d) State law (c) in symbols, for $f: X \rightarrow Y$ and $B \subseteq Y$.
- (e) True or false, with a reason: for every $A \subseteq X$, $f(X \cap A) = f(A)$.

Solution.

- (a) At the step that unwinds $y \in \bigcap_i f(A_i)$ into “for every $i \in I$, $y \in f(A_i)$ ”, and then each $y \in f(A_i)$ into “there is $a_i \in A_i$ with $f(a_i) = y$ ”. The witnesses a_i may differ from one i to another, and the left side needs **one** point in every A_i . In the preimage version, the corresponding step is “for every i , $f(x) \in B_i$ ”, a statement about the single point $f(x)$ with no “there exists” in it, so nothing needs naming.
- (b) $B_1 \cap B_2 = [1, 4]$, so the left side is $f^{-1}([1, 4]) = [-2, -1] \cup [1, 2]$ ([Example \(Images and preimages of intervals under squaring\)](#) (b)). On the right, $f^{-1}([0, 4]) = [-2, 2]$ and $f^{-1}([1, 9]) = [-3, -1] \cup [1, 3]$, whose intersection is $[-2, -1] \cup [1, 2]$. The two sides agree.
- (c) Both are X . The left side is $f^{-1}(Y) = X$, since the empty intersection in Y is Y ; the right side is the empty intersection in X , which is X .
- (d) $f^{-1}(Y \setminus B) = X \setminus f^{-1}(B)$: a point of X is sent outside B exactly when it is not in the preimage of B .
- (e) True. Since $A \subseteq X$, $X \cap A = A$ (each element of A lies in X), so the two sides are the image of the same set.

1.6.10.2 B. Practice
Exercise. B1: Computing images and preimages

Compute each set, and prove each equality by two inclusions or by a chain of equivalences.

- (a) For $d: \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{Z}$, $d(m, n) = m - n$: the image $d(\{(m, n) \in \mathbb{N} \times \mathbb{N} : m > n\})$ and the preimage $d^{-1}(\{0\})$.
- (b) For $\mathbf{1}_{[0,1]}: \mathbb{R} \rightarrow \{0, 1\}$: the preimages of $\{1\}$ and $\{0\}$, and the image of $[2, 3]$.
- (c) For $g: \mathbb{R} \rightarrow \mathbb{R}$, $g(x) = x^2 - 2x$: the image $g([0, 3])$ and the preimages $g^{-1}(\{0\})$ and $g^{-1}([-1, 0])$.

Solution.

- (a) The image is $\mathbb{Z}_{>0}$. ($\subseteq$) If $y = m - n$ with $m > n$, then y is an integer and $y > 0$, so $y \in \mathbb{Z}_{>0}$. ($\supseteq$) Let $k \in \mathbb{Z}_{>0}$. The pair $(k, 0)$ lies in $\mathbb{N} \times \mathbb{N}$, has $k > 0$, and $d(k, 0) = k$; so

k is in the image. The preimage is $d^{-1}(\{0\}) = \{(m, n) \in \mathbb{N} \times \mathbb{N} : m = n\} = \{(n, n) : n \in \mathbb{N}\}$, since $m - n = 0$ exactly when $m = n$.

- (b) By the definition of the indicator (Section 1.4), $\mathbf{1}_{[0,1]}(x) = 1$ for $x \in [0, 1]$ and 0 otherwise. So $\mathbf{1}_{[0,1]}^{-1}(\{1\}) = [0, 1]$ and $\mathbf{1}_{[0,1]}^{-1}(\{0\}) = \mathbb{R} \setminus [0, 1]$, each a chain of one equivalence. The image $\mathbf{1}_{[0,1]}([2, 3]) = \{0\}$: every point of $[2, 3]$ lies outside $[0, 1]$ and is sent to 0, and 0 is hit, for instance by 2.
- (c) Write $g(x) = (x - 1)^2 - 1$. The image is $g([0, 3]) = [-1, 3]$. ($\subseteq$) If $0 \leq a \leq 3$, then $-1 \leq a - 1 \leq 2$, so $|a - 1| \leq 2$, and hence $0 \leq (a - 1)^2 = |a - 1|^2 \leq 4$ by Logic and Proof, Theorem 2.4.5 applied to $|a - 1|$ and 2, and $-1 \leq g(a) \leq 3$. ($\supseteq$) Let $-1 \leq y \leq 3$, and take $a := 1 + \sqrt{y + 1}$. Then $0 \leq \sqrt{y + 1} \leq 2$, so $a \in [1, 3] \subseteq [0, 3]$, and $g(a) = (y + 1) - 1 = y$. For the preimages: $g(x) = 0 \iff x(x - 2) = 0 \iff x \in \{0, 2\}$, so $g^{-1}(\{0\}) = \{0, 2\}$; and $-1 \leq g(x) \leq 0 \iff 0 \leq (x - 1)^2 \leq 1 \iff |x - 1| \leq 1 \iff x \in [0, 2]$, so $g^{-1}([-1, 0]) = [0, 2]$.

Exercise. B2: Two laws, element by element

Let $f: X \rightarrow Y$ be a function. Prove each law directly, by two inclusions or by a chain of equivalences at a point, tagging each step that names or exhibits a witness with $[\exists E]$ or $[\exists I]$.

- (a) $f(A \cup B) = f(A) \cup f(B)$ for $A, B \subseteq X$. (This is a case of Theorem 1.6.3; prove it without citing it.)
- (b) $f^{-1}(C \setminus D) = f^{-1}(C) \setminus f^{-1}(D)$ for $C, D \subseteq Y$. (This is Theorem 1.6.1 (d); prove it without citing it.) Which steps, if any, need a tag?

Solution.

- (a) ($\subseteq$) Let $y \in f(A \cup B)$. There is $c \in A \cup B$ with $f(c) = y$; name it c $[\exists E]$. By Definition 1.2.1, $c \in A$ or $c \in B$. *Case 1:* $c \in A$. Then $y \in f(A)$, with witness c $[\exists I]$, so $y \in f(A) \cup f(B)$. *Case 2:* $c \in B$. Then $y \in f(B)$, with witness c $[\exists I]$, so $y \in f(A) \cup f(B)$. ($\supseteq$) Let $y \in f(A) \cup f(B)$. *Case 1:* $y \in f(A)$. There is $a \in A$ with $f(a) = y$; name it a $[\exists E]$. Then $a \in A \cup B$, so $y \in f(A \cup B)$, with witness a $[\exists I]$. *Case 2:* $y \in f(B)$ is the same with B in place of A . In either case $y \in f(A \cup B)$.
- (b) Both sides are subsets of X . For $x \in X$: $x \in f^{-1}(C \setminus D)$ if and only if $f(x) \in C \setminus D$ (Definition 1.6.2) if and only if $f(x) \in C$ and $f(x) \notin D$ (Definition 1.2.1) if and only if $x \in f^{-1}(C)$ and $x \notin f^{-1}(D)$ (Definition 1.6.2, and Logic and Proof, Example (Negating both sides of an iff) for the second clause) if and only if $x \in f^{-1}(C) \setminus f^{-1}(D)$. No step names or exhibits a witness, so no step carries a tag: this is the difference between (a) and (b).

Exercise. B3: The two readings of the inverse

Let $f: X \rightarrow Y$ be a bijection, with inverse function $f^{-1}: Y \rightarrow X$, and let $B \subseteq Y$. Prove that the image of B under the function f^{-1} equals the preimage of B under f . To keep the readings apart, write $g := f^{-1}$ for the inverse function, and prove $g(B) = \{x \in X : f(x) \in B\}$. The remark in “Two readings of the inverse symbol” gives the argument in brief; write it out as two inclusions, tagging each step that names or exhibits a witness with $[\exists E]$ or $[\exists I]$.

Solution. ($\subseteq$) Let $x \in g(B)$. There is $b \in B$ with $g(b) = x$; name it b $[\exists E]$. Then $x \in X$, and $f(x) = f(g(b)) = b \in B$, since $f \circ g = \text{id}_Y$. So x lies in the preimage.

($\supseteq$) Let $x \in X$ with $f(x) \in B$. Then $g(f(x)) = x$, since $g \circ f = \text{id}_X$. So $x \in g(B)$, with witness $f(x) \in B$ $[\exists I]$.

Hence $g(B)$ equals the preimage of B under f : for a bijection, the two readings of $f^{-1}(B)$ name the same set. Both equations of Definition 1.5.2 were used, one in each inclusion.

1.6.10.3 C. Going deeper

Exercise. C1: Images and differences

Let $f: X \rightarrow Y$ be a function and $A, B \subseteq X$.

- Here are the first lines of a proof that $f(A) \setminus f(B) \subseteq f(A \setminus B)$. Name the move, then complete it.
Let $y \in f(A) \setminus f(B)$. Then $y \in f(A)$, so there is $a \in A$ with $f(a) = y$; name it a $[\exists E]$. Also $y \notin f(B)$.
- Show that the reverse inclusion can fail, with $f(x) = x^2$ on $\mathbb{R}$, $A = \{-1, 1\}$, and $B = \{1\}$.
- Prove that $f(A \setminus B) = f(A) \setminus f(B)$ holds for all $A, B \subseteq X$ if and only if f is injective.

Solution.

- The move is **element-chasing**: unwind the membership of the hypothesis into its definition, and rewind into the definition of the goal. The goal is $y \in f(A \setminus B)$, which needs a witness in $A \setminus B$, and the natural candidate is the a already named. We know $a \in A$; we need $a \notin B$.
Suppose $a \in B$. Then $y = f(a) \in f(B)$, with witness a $[\exists I]$, which contradicts $y \notin f(B)$. Hence $a \notin B$ $[\neg I]$. So $a \in A \setminus B$ and $f(a) = y$, which gives $y \in f(A \setminus B)$, with witness a $[\exists I]$.
- $A \setminus B = \{-1\}$, so $f(A \setminus B) = \{1\}$. But $f(A) = \{1\} = f(B)$, so $f(A) \setminus f(B) = \emptyset$. The point 1 lies in $f(A \setminus B)$ and not in $f(A) \setminus f(B)$.
- ($\Leftarrow$) Suppose f is injective. By (a) it suffices to show $f(A \setminus B) \subseteq f(A) \setminus f(B)$. Let $y \in f(A \setminus B)$, and name $a \in A \setminus B$ with $f(a) = y$. Then $y \in f(A)$, with witness a . Suppose $y \in f(B)$, and name $b \in B$ with $f(b) = y$. Then $f(a) = f(b)$, so $a = b$ by injectivity, and $a \in B$, contradicting $a \in A \setminus B$. Hence $y \notin f(B)$, and $y \in f(A) \setminus f(B)$.

($\Rightarrow$) Suppose equality holds for all $A, B \subseteq X$. Let $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$, and take $A := \{x_1\}$, $B := \{x_2\}$. Then $f(A) = \{f(x_1)\} = f(B)$, so $f(A \setminus B) = f(A) \setminus f(B) = \emptyset$. Suppose $x_1 \neq x_2$. Then $x_1 \in A \setminus B$, so $f(x_1) \in f(A \setminus B) = \emptyset$, which is impossible. Hence $x_1 = x_2$, by RAA (Logic and Proof, Definition 2.5.1), and f is injective.

Marking scheme. (10 points)

- 3: (a) shows $a \notin B$ by deriving $y \in f(B)$ from $a \in B$, and concludes $y \in f(A \setminus B)$ with the witness a .
- 2: (b) computes both sides for the given data.
- 3: (c) ($\Leftarrow$) names two witnesses, $a \in A \setminus B$ and $b \in B$, and uses injectivity to identify them.

- 2: (c) ($\Rightarrow$) applies the hypothesis to one-point sets (or any sets that separate x_1 from x_2) and concludes $x_1 = x_2$.

Fatal errors (0 for the step): using one name for the witnesses of $y \in f(A \setminus B)$ and $y \in f(B)$ (entry (N4) of *Logic and Proof*, Chapter 5's catalogue); in ($\Rightarrow$), assuming injectivity to compute $f(A \setminus B)$.

Exercise. C2: Find the error

A student proves the following claim: *if $f: X \rightarrow Y$ is injective and $A \subseteq X$, then $f^{-1}(f(A)) \subseteq A$.*

(1) Let $x \in f^{-1}(f(A))$. (2) By the definition of preimage, $x \in X$ and $f(x) \in f(A)$. (3) Since $f(x) \in f(A)$, we have $x \in A$. (4) Since x was arbitrary, $f^{-1}(f(A)) \subseteq A$.

- Is the claim true?
- Which line fails, and which rule does it skip?
- Give a counterexample to the failing step, and repair the proof.

Solution.

- The claim is true: it is the ($\Leftarrow$) half of [Proposition 1.6.5](#) (a).
- Line (3) fails. The statement $f(x) \in f(A)$ says only that $f(a) = f(x)$ for **some** $a \in A$ ([Definition 1.6.1](#)). To use it, the proof must name that witness by $\exists E$, with a fresh name. Line (3) skips the $\exists E$ and silently takes the witness to be x itself. The name x is already in use (line (1) introduced it), so the step breaks the fresh-name condition of $\exists E$ (*Logic and Proof*, Section 4.6), the condition behind entry (N4) of the catalogue of *Logic and Proof*, Chapter 5. The tell is that injectivity, a hypothesis, is never used: an argument that ignores a necessary hypothesis cannot be right.
- A counterexample to the step " $f(x) \in f(A)$, so $x \in A$ " needs a non-injective function, since for injective ones the conclusion of the step is true. Take $f(x) = x^2$ on $\mathbb{R}$, $A = \{1\}$, and $x = -1$. Then $f(x) = 1 \in f(A)$, but $x \notin A$.

Repair: replace (3) by "Since $f(x) \in f(A)$, there is $a \in A$ with $f(a) = f(x)$; name it a . Since f is injective, $a = x$. Hence $x \in A$."

Exercise. C3: Exit check: preimages

- Let $f: X \rightarrow Y$ be a function and $(B_i)_{i \in I}$ a family of subsets of Y . Prove that $f^{-1}(\bigcap_{i \in I} B_i) = \bigcap_{i \in I} f^{-1}(B_i)$, the intersections taken in Y and in X .
- Give a function f and sets A, B with $f(A \cap B) \neq f(A) \cap f(B)$, computing both sides.

Solution.

- We prove two inclusions.
($\subseteq$) Let $x \in f^{-1}(\bigcap_{i \in I} B_i)$. By [Definition 1.6.2](#), $x \in X$ and $f(x) \in \bigcap_{i \in I} B_i$. By

Definition 1.2.4, $f(x) \in B_i$ for every $i \in I$. Let $i \in I$. Then $f(x) \in B_i$ and $x \in X$, so $x \in f^{-1}(B_i)$. Since i was arbitrary, x lies in every $f^{-1}(B_i)$, and since $x \in X$, $x \in \bigcap_{i \in I} f^{-1}(B_i)$.

($\supseteq$) Let $x \in \bigcap_{i \in I} f^{-1}(B_i)$. Then $x \in X$ and $x \in f^{-1}(B_i)$ for every $i \in I$. Let $i \in I$. Then $f(x) \in B_i$ by **Definition 1.6.2**. Since i was arbitrary, and $f(x) \in Y$, $f(x) \in \bigcap_{i \in I} B_i$. So $x \in f^{-1}(\bigcap_{i \in I} B_i)$.

If $I = \emptyset$, both sides are X , and the argument above covers this case: the “for every i ” statements hold vacuously.

- (b) Let $f: \{1, 2\} \rightarrow \{0\}$ be the constant function, $A = \{1\}$, $B = \{2\}$. Then $A \cap B = \emptyset$, so $f(A \cap B) = \emptyset$. But $f(A) = \{0\} = f(B)$, so $f(A) \cap f(B) = \{0\} \neq \emptyset$. (The squaring map with $A = [-1, 0]$, $B = [0, 1]$ is another answer: $\{0\}$ against $[0, 1]$.)

Marking scheme. (10 points)

- 1: announces two inclusions (or a chain of equivalences at a point $x \in X$).
- 3: ($\subseteq$) unwinds $x \in f^{-1}(\bigcap B_i)$ to “ $f(x) \in B_i$ for every i ”, then rewinds to $x \in f^{-1}(B_i)$ for an **arbitrary** i .
- 3: ($\supseteq$) the reverse chain, with i arbitrary.
- 1: the ambient sets are named ($x \in X$, $f(x) \in Y$) or the case $I = \emptyset$ is handled.
- 2: a correct example, with both sides computed.

A single chain of equivalences, each step justified, earns the first seven points. Fatal errors (0 for the step): fixing one i and concluding for all i (entry (N1) of *Logic and Proof*, Chapter 5’s catalogue); stating the example’s two sides without checking an element that lies in one and not the other.

1.6.10.4 R. Review

Exercise. R1

Here is a derivation of “if $A \subseteq B$, then $\mathcal{P}(A) \subseteq \mathcal{P}(B)$ ”. The steps “by def.” unwind or rewind **Definition 1.1.2** and **Definition 1.2.2**.

1		$A \subseteq B$	
2		$\forall x(x \in A \rightarrow x \in B)$	def. 1
3	S_0	$S_0 \in \mathcal{P}(A)$	
4		$S_0 \subseteq A$	def. 3
5		$\forall x(x \in S_0 \rightarrow x \in A)$	def. 4
6	x_0	$x_0 \in S_0$	
7		$x_0 \in S_0 \rightarrow x_0 \in A$	$\forall E$ 5
8		$x_0 \in A$	$\rightarrow E$ 7, 6
9		$x_0 \in A \rightarrow x_0 \in B$	$\forall E$ 2
10		$x_0 \in B$	$\rightarrow E$ 9, 8
11		$x_0 \in S_0 \rightarrow x_0 \in B$	$\rightarrow I$ 6–10
12		$\forall x(x \in S_0 \rightarrow x \in B)$	$\forall I$ 6–11
13		$S_0 \subseteq B$	def. 12
14		$S_0 \in \mathcal{P}(B)$	def. 13
15		$S_0 \in \mathcal{P}(A) \rightarrow S_0 \in \mathcal{P}(B)$	$\rightarrow I$ 3–14
16		$\forall S(S \in \mathcal{P}(A) \rightarrow S \in \mathcal{P}(B))$	$\forall I$ 3–15
17		$\mathcal{P}(A) \subseteq \mathcal{P}(B)$	def. 16
18		$A \subseteq B \rightarrow \mathcal{P}(A) \subseteq \mathcal{P}(B)$	$\rightarrow I$ 1–17

Write the proof as plain prose, the way a mathematician would. Then say which sentence of your proof carries each of the five boxes.

Solution. Proof. Suppose $A \subseteq B$. Let $S \in \mathcal{P}(A)$, so $S \subseteq A$. Let $x \in S$. Then $x \in A$, since $S \subseteq A$, and so $x \in B$, since $A \subseteq B$. Hence $S \subseteq B$, that is, $S \in \mathcal{P}(B)$. This proves $\mathcal{P}(A) \subseteq \mathcal{P}(B)$.

The boxes. “Suppose $A \subseteq B$ ” opens the $\rightarrow I$ box 1–17; its closing line 18 is the silent frame of the whole proof. “Let $S \in \mathcal{P}(A)$ ” opens two boxes at once: the $\forall I$ box for S_0 and the $\rightarrow I$ box 3–14 inside it. “Let $x \in S$ ” likewise opens the $\forall I$ box for x_0 and the $\rightarrow I$ box 6–10. “Hence $S \subseteq B$ ” closes the two inner boxes (lines 11 and 12) and rewinds the definition (line 13). “This proves $\mathcal{P}(A) \subseteq \mathcal{P}(B)$ ” closes the S boxes (lines 15, 16) and rewinds (line 17).

This is the passage from level 0 to level 3 of *Logic and Proof*, Section 5.3: each “let” carries a $\forall I$ box together with the $\rightarrow I$ box of a subset statement, the $\forall E$ and $\rightarrow E$ lines 7–10 become the two “since” clauses, the unwinding lines 2 and 5 go silent, line 4 becomes a “so” and line 14 a “that is”, and the rewinding lines 13 and 17 are the two conclusions the prose states.

Exercise. R2

Decide whether $A \setminus (B \setminus C) = (A \setminus B) \setminus C$ for all sets A, B, C , by a table of the eight possible

memberships of a point x in A , B , C . If the two sides can differ, turn a row where they differ into a counterexample.

Solution. Write P , Q , R for the statements $x \in A$, $x \in B$, $x \in C$. By [Definition 1.2.1](#), $x \in A \setminus (B \setminus C)$ is $P \wedge \neg(Q \wedge \neg R)$, and $x \in (A \setminus B) \setminus C$ is $(P \wedge \neg Q) \wedge \neg R$. The table:

P	Q	R	$P \wedge \neg(Q \wedge \neg R)$	$(P \wedge \neg Q) \wedge \neg R$
T	T	T	T	F
T	T	F	F	F
T	F	T	T	F
T	F	F	T	T
F	T	T	F	F
F	T	F	F	F
F	F	T	F	F
F	F	F	F	F

The columns differ in the first and third rows. Take the first row: put the point 1 in exactly the sets whose statement is true, all three, so $A = B = C = \{1\}$. Then $B \setminus C = \emptyset$, so $A \setminus (B \setminus C) = \{1\}$; and $A \setminus B = \emptyset$, so $(A \setminus B) \setminus C = \emptyset$. The two sides differ, so the identity is false. (The third row gives $A = C = \{1\}$, $B = \emptyset$, with the same two answers.)

Why a table suffices: each side's membership condition is a formula in P , Q , R , and a row with the two columns different gives sets directly, which are then checked by computing both sides, as above. The truth tables of *Logic and Proof*, Chapter 3 add one more thing: by Logic and Proof, Corollary 4.7.3, the first row shows that the sequent $P \wedge \neg(Q \wedge \neg R) \vdash (P \wedge \neg Q) \wedge \neg R$ is not derivable, so the propositional rules cannot carry a point of the left side into the right side.

Exercise. R3

Derive $(P \wedge \neg Q) \rightarrow R \vdash P \rightarrow (Q \vee R)$. Then read the derivation as a proof that for sets A, B, C , if $A \setminus B \subseteq C$, then $A \subseteq B \cup C$. Which step is classical?

Solution. The goal is an implication, so open a box on P with goal $Q \vee R$. The premise can fire only on $P \wedge \neg Q$, which needs $\neg Q$, and nothing gives $\neg Q$ outright. So split on $Q \vee \neg Q$, by excluded middle (Logic and Proof, Proposition 2.5.3): in the case Q the goal is immediate, and in the case $\neg Q$ the premise gives R .

1	$(P \wedge \neg Q) \rightarrow R$	
2	P	
3	$Q \vee \neg Q$	LEM
4	Q	
5	$Q \vee R$	$\vee I$ 4
6	$\neg Q$	
7	$P \wedge \neg Q$	$\wedge I$ 2, 6
8	R	$\rightarrow E$ 1, 7
9	$Q \vee R$	$\vee I$ 8
10	$Q \vee R$	$\vee E$ 3, 4–5, 6–9
11	$P \rightarrow (Q \vee R)$	$\rightarrow I$ 2–10

As sets. Suppose $A \setminus B \subseteq C$, and let $x \in A$. Take P, Q, R to be $x \in A, x \in B, x \in C$. By [Definition 1.1.2](#) and [Definition 1.2.1](#), the hypothesis at this x is $(P \wedge \neg Q) \rightarrow R$, and the goal $x \in B \cup C$ is $Q \vee R$. The derivation, used as a derived rule (Logic and Proof, Definition 2.3.1), gives $P \rightarrow (Q \vee R)$, and with P it gives $x \in B \cup C$. In prose: either $x \in B$, and then $x \in B \cup C$; or $x \notin B$, and then $x \in A \setminus B \subseteq C$, so $x \in C \subseteq B \cup C$. Since x was arbitrary, $A \subseteq B \cup C$.

The classical step is line 3, the case split “ $x \in B$ or $x \notin B$ ” by excluded middle, which is derived with RAA (Logic and Proof, Definition 2.5.1).

Exercise. R4

Show that there is no set that contains every singleton $\{x\}$ as an element.

Solution. Suppose S were such a set. Let $T := \{s \in S : s \text{ is a singleton}\}$, a set by specification ([Definition 1.1.3](#)); its elements are sets, since S may also contain numbers, and those are excluded. Consider the union $\bigcup T$ of the sets in T ([Definition 1.2.4](#), the form $\bigcup S$ for a set S whose elements are sets). For every set x , the singleton $\{x\}$ lies in S , hence in T , and $x \in \{x\}$; so $x \in \bigcup T$. Thus $\bigcup T$ is a set of which every set is an element, and [Theorem 1.1.4](#) says there is no such set. This contradiction shows that no such S exists.

The tool, and the look-alike. The tool is Russell’s theorem, [Theorem 1.1.4](#), applied not to S but to $\bigcup T$. Applied to S directly, it says only that $\{z \in S : z \notin z\} \notin S$, and that does not contradict anything: S was supposed to contain singletons, not that set. The move that makes it work is the union, which turns “contains the singleton of every set” into “contains every set”. A direct diagonal construction inside S also works, but it repeats the proof of [Theorem 1.1.4](#) instead of using it.

Chapter 2

Equivalence Relations, Orders, and Induction

Logic and Proof, Chapter 3 took one fact about the natural numbers on trust: induction. Everything else we have used about them came from school: that $7 - 3$ is a natural number, that none lies strictly between 3 and 4, that every non-empty set of them has a least member, that every integer is even or odd and not both. This chapter derives all of these from induction, extends them to the integers, and proves that the least-member property, **well-ordering**, is induction in another form. It then turns induction from a principle into a craft: which form fits a claim, where to start, when to argue from a smallest counterexample, when to prove more than was asked, and how a flawed induction fails at exactly one step. It also proves that a starting value and a rule for the next value really do define a sequence, which is what makes $\sum$, $\prod$, $n!$ and powers definitions rather than abbreviations. The first motivating problem comes from *Logic and Proof*, Chapter 2, which reduced a fraction by halving numerator and denominator for as long as both were even: *what stops the halving from going on for ever?*

The second half of the chapter is about relations. When is “the same” a legitimate way to identify different things, as a clock identifies 2 with 14, or a function identifies two points with the same value? And what does “larger” mean when not everything can be compared, as with sets under inclusion or natural numbers under divisibility? The second motivating problem: *in what sense is “ $9 + 5 = 2$ on a clock” a true statement?* The answer is an equivalence relation, the partition it cuts its set into, and a check that arithmetic on the pieces does not depend on which integer names a piece. For orders the answer is that “largest” splits into two notions, greatest and maximal, which agree only when every two elements are comparable. The last section reaches least upper bounds in a general partially ordered set (*Analysis on the Real Line*, Chapter 1’s supremum is the case of $\leq$ on $\mathbb{R}$, and its axiom says that such suprema exist) and closes with well-ordered sets, the order-theoretic name for what Section 2.1 proved about $\mathbb{N}$.

What you need. *Logic and Proof*, Chapters 1–5, and *Sets, Relations, and Cardinality*, Chapter 1. From *Logic and Proof*, Chapters 1 and 2: the $\rightarrow$ I box (Logic and Proof, Definition 1.4.1), proof by contradiction (Logic and Proof, Definition 2.5.1), and the two results whose school facts this chapter proves, the irrationality of $\sqrt{2}$ (Logic and Proof, Theorem 2.2.1) and Logic and Proof, Theorem 2.1.3. From *Logic and Proof*, Chapter 3: the principle of induction (Logic and Proof, Axiom 4.2.1) and strong induction (Logic and Proof, Theorem 4.2.2), with the worked examples that introduced them (Logic and Proof, Example (A sum of products of neighbors), Logic and Proof, Example (An inequality by induction), Logic and Proof, Example (Threes and fives)), and structural recursion and induction (Logic and Proof, Theorem 4.3.4, Logic and Proof, Theorem 4.3.1), which appear here as pointers back and in one review exercise. From *Logic and*

Proof, Chapter 4: the rules for “for every” and “there exists” (Logic and Proof, Definition 5.4.1, Logic and Proof, Definition 5.6.1), the rules for equality (Logic and Proof, Definition 5.7.1, Logic and Proof, Proposition 5.7.1), bounded quantifiers (Logic and Proof, Proposition 6.1.1), and negating quantifiers and implications (Logic and Proof, Proposition 6.2.3, Logic and Proof, Proposition 6.2.2). From *Logic and Proof*, Chapter 5: how an induction is written in prose (Logic and Proof, Example (An induction, and a legitimate symmetry)) and the catalogue of what a proof may not do (Logic and Proof, Example (What is not allowed)). From Chapter 1: sets cut out by a property (Definition 1.1.3), power sets (Definition 1.2.2), unions of families (Definition 1.2.4), relations as sets of pairs (Definition 1.3.3), functions and when two are equal (Definition 1.4.1, Proposition 1.4.1), and preimages (Definition 1.6.2). The arithmetic and order of $\mathbb{R}$ are school facts, as in the earlier chapters; *Analysis on the Real Line*, Chapter 1 states them as axioms.

Time. About 30 hours on the core track (all of A, one B and one C per section, the exit checks); about 45 hours for everything.

2.0.1 Roadmap

- **The natural numbers and well-ordering.** From induction alone, the arithmetic of $\mathbb{N}$ and $\mathbb{Z}$, the proof that every non-empty set of natural numbers has a least element, and the theorem that induction and well-ordering are equivalent, with a witness for each hypothesis it needs.
- **Induction, strong induction and the smallest counterexample.** Induction and strong induction from any integer, the smallest counterexample as a method, the division algorithm, lowest terms and prime factors as its first uses, and the ways an induction proof goes wrong, each located at one step.
- **Definition by recursion.** The theorem that a starting value and a rule define exactly one sequence, its variants from another base and with two previous values, sums, products and powers defined by it, and inductions that follow a definition’s clauses or prove more than was asked.
- **Properties of relations.** Reflexive, symmetric, antisymmetric and transitive, each failure proved by a witness, divisibility as the relation whose antisymmetry depends on the set, and a table of eight relations checked against the four properties.
- **Equivalence relations and partitions.** Relations that behave like equality, the proof that their classes partition the set and that every partition arises from exactly one of them, the integers modulo n , and why arithmetic on classes must be checked to be well defined.
- **Partial orders, maximal and greatest elements.** Relations that behave like $\leq$, greatest elements against maximal ones and when the two agree, least upper bounds in a partially ordered set, and well-ordered sets.

2.0.2 Named moves

- **Smallest counterexample.** Suppose the claim fails; well-ordering gives a least failure m ; below m the claim holds, and from that produce a failure smaller than m , or the claim at m itself.
- **Prove more (strengthen the hypothesis).** When the inductive step will not go through, ask what the induction hypothesis would need to say, and prove that stronger statement instead.
- **Recursion, then induction.** Define an object by a recursion, and prove facts about it by an induction that follows the same clauses: the base case uses the starting value, and the step uses the rule.
- **Classes from a relation, a relation from classes.** Pass between an equivalence relation

and its partition, whichever description makes the next step easier.

2.1 The Natural Numbers and Well-Ordering

Logic and Proof, Chapter 3 took one fact about the natural numbers on trust: the principle of induction (Logic and Proof, Axiom 4.2.1), which says that a property holding at 0 and passing from each n to $n + 1$ holds for every $n \in \mathbb{N}$. From it, *Logic and Proof*, Chapter 3 proved strong induction (Logic and Proof, Theorem 4.2.2), whose step may use the property at **every** smaller number. Everything else we have said about $\mathbb{N}$ was school knowledge: that $7 - 3$ is a natural number, that none lies strictly between 3 and 4, that every non-empty set of natural numbers has a least member. This section derives those facts from induction, extends them to the integers, and then proves that the least-member property, **well-ordering**, is induction in another form.

2.1.1 What we assume about the natural numbers

The principle of induction is usually used with a property written as a sentence. It has an equivalent form about sets, which is the one *Analysis on the Real Line*, Chapter 1 uses.

Remark. Let $S \subseteq \mathbb{N}$ contain 0, and suppose $n + 1 \in S$ whenever $n \in S$. Apply Logic and Proof, Axiom 4.2.1 to the property “ $n \in S$ ”: its base case is $0 \in S$, and its step is the closure of S . So every $n \in \mathbb{N}$ lies in S , that is, $\mathbb{N} \subseteq S$, and with $S \subseteq \mathbb{N}$ this gives $S = \mathbb{N}$ (Proposition 1.1.1). In words: the only subset of $\mathbb{N}$ that contains 0 and is closed under $n \mapsto n + 1$ is $\mathbb{N}$ itself. *Analysis on the Real Line*, Chapter 1 states this as “ $\mathbb{N}$ is contained in every inductive set”, where an inductive set contains 0, as here.

Here is everything the section assumes about $\mathbb{N}$. The natural numbers form a subset $\mathbb{N} \subseteq \mathbb{R}$ of the familiar real numbers such that

- (($\mathbb{N}$)0) $0 \in \mathbb{N}$;
- (NS) for every $n \in \mathbb{N}$, also $n + 1 \in \mathbb{N}$;

and Logic and Proof, Axiom 4.2.1 holds. The arithmetic and the order of the real numbers (the laws of $+$, $-$, $\cdot$ and $\leq$, trichotomy, multiplying an inequality by a positive number) are school facts in this volume and in *Logic and Proof*; *Analysis on the Real Line*, Chapter 1 states them as axioms of an ordered field. These three items are a description of $\mathbb{N}$, not new axioms: *Logic and Proof*, Chapter 3 already used all of them.

The rule for this section. From here to the end of the section, a fact about $\mathbb{N}$ is used only if it is (($\mathbb{N}$)0), (NS), Logic and Proof, Axiom 4.2.1, a statement proved earlier in the section, or Logic and Proof, Theorem 4.2.2 once the paragraph after Lemma 2.1.1 has shown that its *Logic and Proof*, Chapter 3 proof rests only on these. The real-number arithmetic is used freely. In particular “ $n \geq 0$ ” for a natural number n , which looks too obvious to mention, is not available until it is proved below, in Lemma 2.1.1 (a).

The integers and the rationals are then **defined** from $\mathbb{N}$, as subsets of $\mathbb{R}$:

$$\begin{aligned}\mathbb{Z} &:= \mathbb{N} \cup \{-n : n \in \mathbb{N}\}, \\ \mathbb{Z}_{>0} &:= \{n \in \mathbb{N} : n \neq 0\}, \\ \mathbb{Q} &:= \{m/n : m \in \mathbb{Z}, n \in \mathbb{Z}_{>0}\}.\end{aligned}$$

In words: an integer is a natural number or the negative of one; a positive integer is a natural number other than 0; a rational number is a quotient of an integer by a positive integer. These are word for word the definitions of *Analysis on the Real Line*, Section 1.5, so the two chapters speak of the same sets. Examples: $-3 \in \mathbb{Z}$ because $3 \in \mathbb{N}$; 0 lies in $\mathbb{Z}$ twice over, as $0 \in \mathbb{N}$ and as -0 , which a union does not mind; $-\frac{1}{2} = (-1)/2 \in \mathbb{Q}$. The definitions do not yet tell us that $\frac{1}{2} \notin \mathbb{Z}$, or that $\mathbb{Z}_{>0}$ is the set of natural numbers ≥ 1 . Both need the lemma below.

Two pieces of notation will save words later. For $n \in \mathbb{N}$,

$$\{1, \dots, n\} := \{k \in \mathbb{N} : 1 \leq k \leq n\}, \quad \{0, \dots, n\} := \{k \in \mathbb{N} : k \leq n\}.$$

So $\{1, \dots, 0\} = \emptyset$, since no k has $1 \leq k \leq 0$. We do not write $[n]$ for these sets: square brackets will mean an equivalence class in Section 2.5.

2.1.2 The arithmetic of the natural numbers

The facts below are the ones *Logic and Proof*, Chapter 3 and *Analysis on the Real Line*, Chapter 1 used without proof. Each of (a)–(d) is one ordinary induction, and the only work is to choose the property to induct on; (e) follows from (a) and (d).

Lemma 2.1.1. Arithmetic of the Natural Numbers

For all $m, n \in \mathbb{N}$:

- (a) $n \geq 0$; and if $n \neq 0$, then $n \geq 1$;
- (b) $m + n \in \mathbb{N}$ and $mn \in \mathbb{N}$;
- (c) if $n \neq 0$ (equivalently, by (a), $n \geq 1$), then $n - 1 \in \mathbb{N}$;
- (d) if $m \leq n$, then $n - m \in \mathbb{N}$;
- (e) there is **no** $k \in \mathbb{N}$ with $n < k < n + 1$; equivalently, if $m < n$, then $m + 1 \leq n$.

In words: (a) says that 0 is the smallest natural number and that nothing in $\mathbb{N}$ lies between 0 and 1. (b) says $\mathbb{N}$ is closed under sums and products, (c) that every natural number except 0 has a predecessor in $\mathbb{N}$, and (d) that a difference “larger minus smaller” stays in $\mathbb{N}$. (e) says the natural numbers come one unit apart, with no gaps to hide in.

Idea. Each induction needs a property that is true at 0 and survives the step, and the statement as written does not always qualify. “If $n \neq 0$, then $n \geq 1$ ” says nothing useful at 0, so for (a) we induct on the disjunction “ $n = 0$ or $n \geq 1$ ”. For (b) we fix one of the two numbers and induct on the other. For (d) we induct on m . The step from m to $m + 1$ uses the hypothesis only at the same n , with the weaker bound $m \leq n$, and then subtracts 1 using (c). So n could be fixed first, as m is in (b); the proof below puts “for every n ” inside the property instead, which is harmless.

Proof. (a) Let $P(n)$ be “ $n = 0$ or $n \geq 1$ ”. *Base case:* $P(0)$ holds by its first disjunct. *Step:* let $n \in \mathbb{N}$ with $P(n)$. If $n = 0$, then $n + 1 = 1 \geq 1$; if $n \geq 1$, then $n + 1 \geq 2 \geq 1$. In both cases $n + 1 \geq 1$, so $P(n + 1)$ holds by its second disjunct. By Logic and Proof, Axiom 4.2.1, every $n \in \mathbb{N}$ satisfies $n = 0$ or $n \geq 1$. In both cases $n \geq 0$; and if $n \neq 0$, the first disjunct is excluded, so $n \geq 1$.

(b) Fix $m \in \mathbb{N}$. *Sums:* let $P(n)$ be “ $m + n \in \mathbb{N}$ ”. Then $P(0)$ holds, since $m + 0 = m \in \mathbb{N}$. If

$m + n \in \mathbb{N}$, then $m + (n + 1) = (m + n) + 1 \in \mathbb{N}$ by (NS). By Logic and Proof, Axiom 4.2.1, $m + n \in \mathbb{N}$ for every $n \in \mathbb{N}$; as m was arbitrary, this holds for all $m, n \in \mathbb{N}$. *Products:* let $Q(n)$ be “ $mn \in \mathbb{N}$ ”. Then $m \cdot 0 = 0 \in \mathbb{N}$ by $((\mathbb{N})0)$. If $mn \in \mathbb{N}$, then $m(n + 1) = mn + m$, a sum of the two natural numbers mn and m , so it lies in $\mathbb{N}$ by the half just proved. By Logic and Proof, Axiom 4.2.1, $mn \in \mathbb{N}$ for every n .

(c) Let $P(n)$ be “ $n = 0$ or $n - 1 \in \mathbb{N}$ ”. $P(0)$ holds by its first disjunct. Let $n \in \mathbb{N}$; then $(n + 1) - 1 = n \in \mathbb{N}$, so $P(n + 1)$ holds by its second disjunct. By Logic and Proof, Axiom 4.2.1, $P(n)$ holds for every $n \in \mathbb{N}$, and for $n \neq 0$ its second disjunct is the claim.

(d) Let $P(m)$ be “for every $n \in \mathbb{N}$, if $m \leq n$, then $n - m \in \mathbb{N}$ ”. *Base case:* $n - 0 = n \in \mathbb{N}$. *Step:* let $m \in \mathbb{N}$ with $P(m)$, and let $n \in \mathbb{N}$ with $m + 1 \leq n$. Then $m < m + 1 \leq n$, so $n - m \in \mathbb{N}$ by $P(m)$. Moreover $n - m \geq 1 > 0$, so $n - m \neq 0$, and (c) gives $(n - m) - 1 \in \mathbb{N}$. Since $(n - m) - 1 = n - (m + 1)$, this is $P(m + 1)$. By Logic and Proof, Axiom 4.2.1, $P(m)$ holds for every $m \in \mathbb{N}$.

(e) Let $m < n$. By (d), $n - m \in \mathbb{N}$, and $n - m > 0$, so $n - m \neq 0$ and (a) gives $n - m \geq 1$, that is, $m + 1 \leq n$. Now suppose $k \in \mathbb{N}$ with $n < k < n + 1$. Applying what we just proved to $n < k$ gives $n + 1 \leq k$, which contradicts $k < n + 1$. So no such k exists. This proves the lemma. ■

Read the proof of (c) again. Its inductive step never mentions $P(n)$.

Example. An induction that ignores its hypothesis

Logic and Proof, Chapter 3, after its first worked induction (Logic and Proof, Example (A sum of products of neighbors)), said: “A proof by induction that never uses its induction hypothesis is either not an induction or is hiding a gap.” The proof of [Lemma 2.1.1](#) (c) never uses its induction hypothesis. Is it an induction, and is it hiding a gap?

Solution. It is an induction, and it hides no gap. Its base case $P(0)$ is proved, and its step “ $P(n) \Rightarrow P(n + 1)$ ” is a valid implication for every $n \in \mathbb{N}$: the conclusion $P(n + 1)$ is proved outright, so the implication holds whatever $P(n)$ says. Then Logic and Proof, Axiom 4.2.1 delivers $P(n)$ for every n . The hypothesis $n \in \mathbb{N}$ of the step is used, through $(n + 1) - 1 = n \in \mathbb{N}$; only $P(n)$ is not.

Nor can a case analysis replace the axiom here. A direct proof would have to say which numbers lie in $\mathbb{N}$, and the only thing that says so is induction: the step shows that every number of the form $n + 1$ with $n \in \mathbb{N}$ has the property, and it is Logic and Proof, Axiom 4.2.1 that tells us every natural number other than 0 is of this form. *Logic and Proof*, Chapter 3’s sentence is a rule of thumb, and a good one: an unused hypothesis is a reason to check the proof line by line. Here the check passes. The exception is a step whose conclusion is reached without the hypothesis, and the axiom is still what carries the property from 0 to every n .

The lemma also closes a gap in *Logic and Proof*, Chapter 3. The proof of Logic and Proof, Theorem 4.2.2 used two facts about $\mathbb{N}$ as school knowledge. In its base case it said “No natural number is less than 0, so $P(k)$ holds for every $k \in \mathbb{N}$ with $k < 0$, vacuously” and “The only

$k \in \mathbb{N}$ with $k \leq 0$ is 0”: both are (a), the second with trichotomy. In its inductive step it said “The natural numbers $k < n + 1$ are exactly those with $k \leq n$, since no natural number lies strictly between n and $n + 1$ ”, and then “A natural number $k \leq n + 1$ satisfies $k \leq n$ or $k = n + 1$ ”: both are (e), with trichotomy. The proofs of (a), (d) and (e) above use only Logic and Proof, Axiom 4.2.1, $((\mathbb{N})0)$, (NS), part (c) and the arithmetic of $\mathbb{R}$, and none of them uses strong induction. So strong induction rests on those items alone, and from now on we may cite Logic and Proof, Theorem 4.2.2 with no school fact behind it.

Quick check. Where in the proof of [Lemma 2.1.1](#) (d) is part (c) used, and could n have been fixed before the induction on m , as m was in (b)?

Solution. Part (c) is used to pass from $n - m \in \mathbb{N}$ to $n - (m + 1) = (n - m) - 1 \in \mathbb{N}$, after checking $n - m \neq 0$. The step from m to $m + 1$ is about one n with $m + 1 \leq n$, and it uses $P(m)$ at that same n , with the weaker bound $m \leq n$. So yes: fixing $n \in \mathbb{N}$ and inducting on “if $m \leq n$, then $n - m \in \mathbb{N}$ ” proves the same thing, and then n was arbitrary. Putting “for every n ” inside $P(m)$ is harmless but not needed. It would be needed only if the step used the hypothesis at a different n , such as $n - 1$.

2.1.3 The integers

Analysis on the Real Line, Chapter 1 uses that sums, differences and products of integers are integers, and that an integer greater than 3 is at least 4. Both now follow from the lemma, by sorting integers into natural numbers and their negatives.

Proposition 2.1.2. Arithmetic of the Integers

Let $m, n \in \mathbb{Z}$.

- (a) $m + n$, $m - n$ and mn are integers.
- (b) If $n \geq 0$, then $n \in \mathbb{N}$; if $n > 0$, then $n \in \mathbb{Z}_{>0}$. Hence

$$\mathbb{N} = \{n \in \mathbb{Z} : n \geq 0\}, \quad \mathbb{Z}_{>0} = \{n \in \mathbb{Z} : n > 0\} = \{n \in \mathbb{N} : n \geq 1\}.$$

- (c) (**Discreteness.**) If $m < n$, then $m + 1 \leq n$.

Idea. Every integer is a or $-a$ with $a \in \mathbb{N}$, so each claim splits into cases by signs. Only one case has content: $a + (-b) = a - b$ with $a, b \in \mathbb{N}$, which is in $\mathbb{N}$ when $b \leq a$, by [Lemma 2.1.1](#) (d), and is the negative of a natural number when $a < b$. Discreteness then follows from (b) applied to the difference $n - m$: it is a positive integer, hence a natural number with $n - m \geq 1$, by (a) of the lemma.

Proof. (a) First, $\mathbb{Z}$ is closed under $x \mapsto -x$: if $x \in \mathbb{N}$, then $-x \in \mathbb{Z}$ by the definition of $\mathbb{Z}$; if $x = -k$ with $k \in \mathbb{N}$, then $-x = k \in \mathbb{N} \subseteq \mathbb{Z}$.

Next let $a, b \in \mathbb{N}$; we show $a - b \in \mathbb{Z}$. By trichotomy there are three cases. If $a > b$ or $a = b$, then $b \leq a$, and $a - b \in \mathbb{N}$ by [Lemma 2.1.1](#) (d). If $a < b$, then $b - a \in \mathbb{N}$ by (d), and

$$a - b = -(b - a) \in \mathbb{Z}.$$

Now write $m = \pm a$ and $n = \pm b$ with $a, b \in \mathbb{N}$. The table lists the four sign patterns; in each, $a + b$ and ab lie in $\mathbb{N}$ by [Lemma 2.1.1](#) (b), and $a - b, b - a$ lie in $\mathbb{Z}$ by the previous paragraph.

m	n	$m + n$	mn
a	b	$a + b \in \mathbb{N}$	$ab \in \mathbb{N}$
a	$-b$	$a - b \in \mathbb{Z}$	$-(ab) \in \mathbb{Z}$
$-a$	b	$b - a \in \mathbb{Z}$	$-(ab) \in \mathbb{Z}$
$-a$	$-b$	$-(a + b) \in \mathbb{Z}$	$ab \in \mathbb{N}$

So $m + n$ and mn are integers. Finally $m - n = m + (-n)$, a sum of two integers, since $-n \in \mathbb{Z}$ by the first paragraph.

(b) Let $n \geq 0$. By the definition of $\mathbb{Z}$, either $n \in \mathbb{N}$, and we are done, or $n = -k$ with $k \in \mathbb{N}$. In the second case $k = -n \leq 0$, and $k \geq 0$ by [Lemma 2.1.1](#) (a), so $k = 0$ and $n = 0 \in \mathbb{N}$ by $((\mathbb{N})0)$. If moreover $n > 0$, then n is a natural number with $n \neq 0$, so $n \in \mathbb{Z}_{>0}$.

For the set equalities: $\mathbb{N} \subseteq \mathbb{Z}$, and every $n \in \mathbb{N}$ has $n \geq 0$ by (a) of the lemma; with the first sentence of (b) this gives $\mathbb{N} = \{n \in \mathbb{Z} : n \geq 0\}$. Every element of $\mathbb{Z}_{>0}$ is a natural number $n \neq 0$, so $n \geq 1 > 0$ by (a) of the lemma; with the second sentence of (b) this gives $\mathbb{Z}_{>0} = \{n \in \mathbb{Z} : n > 0\}$. Finally, for $n \in \mathbb{N}$, $n \neq 0$ implies $n \geq 1$ by (a), and $n \geq 1$ implies $n \neq 0$; so $\mathbb{Z}_{>0} = \{n \in \mathbb{N} : n \geq 1\}$.

(c) Let $m < n$. By (a), $n - m \in \mathbb{Z}$, and $n - m > 0$, so $n - m \in \mathbb{Z}_{>0}$ by (b), that is, $n - m \geq 1$. Hence $m + 1 \leq n$. This proves the proposition. ■

Discreteness is the fact behind “an integer greater than 3 is at least 4” (take $m = 3$), and behind “the integers x with $-2 \leq x \leq 2$ are $-2, -1, 0, 1, 2$ ”: an integer x with $-2 \leq x$ is -2 or at least -1 , and so on up to 2. It also settles the question left open above: $\frac{1}{2} \notin \mathbb{Z}$, since $0 < \frac{1}{2}$ would force $1 \leq \frac{1}{2}$.

2.1.4 Well-ordering

A fraction with even numerator and denominator can be halved, and halved again, and *Logic and Proof*, Chapter 2 claimed that this cannot go on for ever (the fact (L) in the proof of Logic and Proof, Theorem 2.2.1). The reason is that the absolute values of the denominators form a strictly decreasing list of positive integers, and such a list must stop. The precise form of “must stop” is a statement about least elements.

A **least element** of a set $S \subseteq \mathbb{R}$ is an element $m \in S$ such that $m \leq s$ for **every** $s \in S$. Not every set has one: $(0, 1]$ has none, since for every $x \in (0, 1]$ the number $x/2$ lies in the set and is smaller. The empty set has none either, as a least element must be an element. Reversing the inequality, a **greatest element** of S is an element $m \in S$ with $s \leq m$ for every $s \in S$; this is the maximum of *Logic and Proof*, Section 4.7.

Every non-empty set of natural numbers has a least element.

Theorem 2.1.3. Well-Ordering of the Natural Numbers

Every **non-empty** subset of $\mathbb{N}$ has a least element.

Try it first (ten minutes): suppose that $S \subseteq \mathbb{N}$ has no least element. Which property of n would you prove for every $n \in \mathbb{N}$ by strong induction, and why does it finish the proof?

Idea. A least element cannot be found by inspecting S , since we know nothing about S . So we suppose there is none and show S is empty, which contradicts “non-empty”. The property to carry is “ $n \notin S$ ”. If no number below n is in S and n were in S , then n would be a least element; so n is not in S . The hypothesis must speak about **every** number below n , which is why the induction is strong. *Before reading on: at which line is trichotomy used, and what would go wrong without it?*

Proof. Let $S \subseteq \mathbb{N}$ be non-empty, and suppose, for a contradiction, that S has no least element. For $n \in \mathbb{N}$, let $P(n)$ be “ $n \notin S$ ”.

Let $n \in \mathbb{N}$, and suppose $P(k)$ holds for every $k \in \mathbb{N}$ with $k < n$, that is, no natural number less than n lies in S . Suppose $n \in S$. Let $s \in S$. By trichotomy, $s < n$, $s = n$ or $s > n$. The first is impossible: s is a natural number less than n , so $s \notin S$ by the hypothesis. So $s \geq n$. As $s \in S$ was arbitrary, n is a least element of S , which contradicts our supposition. Hence $n \notin S$, that is, $P(n)$ holds.

By Logic and Proof, Theorem 4.2.2, $n \notin S$ for every $n \in \mathbb{N}$. Since $S \subseteq \mathbb{N}$, the set S has no elements, which contradicts S being non-empty. So S has a least element. This proves the theorem. ■

The outer frame is an RAA on the statement “ S has a least element”. The proof shows that a least element **exists**; it does not say how to find it. *Logic and Proof*, Section 5.7 explained why mathematicians have argued over proofs of this kind. Here the debate has a quiet ending: if S is given by a property that can be checked number by number, testing $0, 1, 2, \dots$ in turn finds the least element, and the theorem says the search stops. Nothing in the proof above depends on such a test.

Quick check. In the proof of [Theorem 2.1.3](#), to which property is strong induction applied, and what is the induction hypothesis at n ?

Solution. To $P(n)$: “ $n \notin S$ ”. At n , the induction hypothesis is that $k \notin S$ for every natural number $k < n$: no element of S lies below n . Together with “ $n \in S$ ” it would make n a least element, which is why $n \in S$ leads to a contradiction.

Warning. Well-ordering is a property of subsets of $\mathbb{N}$, not of every set of numbers. The set of positive rationals is non-empty and has no least element: for every positive rational q ,

the number $q/2$ is a smaller positive rational. Neither has $\mathbb{Z}$: $n - 1 < n$ is again an integer. A proof that takes “the least element” of a set must first show that the set is a non-empty subset of $\mathbb{N}$, or a non-empty set of integers with an integer lower bound, as in [Corollary 2.1.5](#) below.

Quick check. Does $\{n \in \mathbb{N} : n^2 > 50\}$ have a least element? If so, which?

Solution. Yes. The set is non-empty, since $10^2 = 100 > 50$, and it is a subset of $\mathbb{N}$, so [Theorem 2.1.3](#) applies. The least element is 8: $8^2 = 64 > 50$, while $7^2 = 49 \leq 50$, and for $n \leq 7$ we have $n^2 \leq 49$. The theorem promised a least element; finding it took a search from below.

2.1.5 Induction and well-ordering are the same principle

The proof of [Theorem 2.1.3](#) derived well-ordering from induction. The derivation also runs the other way, and the cleanest way to say so is to ask which sets of numbers satisfy which principle. For a set $X \subseteq \mathbb{R}$, consider three structural properties:

- $0 \in X$;
- **closure:** for every $n \in X$, $n + 1 \in X$;
- the **predecessor property:** for every $n \in X$ with $n \neq 0$, $n - 1 \in X$.

The set $\mathbb{N}$ has all three, by $((\mathbb{N})0)$, $(\mathbb{N}S)$ and [Lemma 2.1.1](#) (c). Say that X has the **induction property** if every $S \subseteq X$ with $0 \in S$ and $n + 1 \in S$ for every $n \in S$ equals X ; and the **least-element property** if every non-empty subset of X has a least element. The closure of S in the induction property is plain closure under $n \mapsto n + 1$, not “closure as long as one stays in X ”; the witnesses below show why that matters.

Theorem 2.1.4. Induction Is Equivalent to Well-Ordering

Let $X \subseteq \mathbb{R}$ contain 0, be closed under $n \mapsto n + 1$, and have the predecessor property. Then X has the induction property **if and only if** X has the least-element property.

Idea. $(\Rightarrow)$ A set X with the induction property and the first two properties must be $\mathbb{N}$ itself: $\mathbb{N}$ sits inside X by Logic and Proof, Axiom 4.2.1, and the induction property of X , applied to the subset $\mathbb{N}$, forces equality. Then [Theorem 2.1.3](#) applies. $(\Leftarrow)$ Take S as in the induction property and suppose $S \neq X$; the least element m of $X \setminus S$ is then a smallest failure; the number $m - 1$ just below it is not a failure, so it lies in S , and closure passes that success up to $m = (m - 1) + 1$, a contradiction. *Before reading on: in $(\Leftarrow)$, which of the three hypotheses on X gives $m - 1 \in X$, and which two are not used at all?*

Proof. $(\Rightarrow)$ Suppose X has the induction property. Apply Logic and Proof, Axiom 4.2.1 to the property “ $n \in X$ ”: its base case is $0 \in X$, and its step is the closure of X . So $\mathbb{N} \subseteq X$.

Now $\mathbb{N}$ is a subset of X that contains 0 and is closed under $n \mapsto n + 1$, by $((\mathbb{N})0)$ and (NS) ; the induction property of X gives $\mathbb{N} = X$. By [Theorem 2.1.3](#), every non-empty subset of $X = \mathbb{N}$ has a least element.

$(\Leftarrow)$ Suppose X has the least-element property. Let $S \subseteq X$ contain 0 and be closed under $n \mapsto n + 1$. Suppose $S \neq X$. Then $X \setminus S$ is non-empty: if it were empty, every element of X would lie in S , so $X \subseteq S$, and with $S \subseteq X$ this would give $S = X$ by [Proposition 1.1.1](#). So $X \setminus S$ has a least element m . We have $m \neq 0$, since $0 \in S$. So $m - 1 \in X$, by the predecessor property. Since $m - 1 < m$ and m is least in $X \setminus S$, $m - 1 \notin X \setminus S$, that is, $m - 1 \in S$. Hence $m = (m - 1) + 1 \in S$ by the closure of S , which contradicts $m \in X \setminus S$. So $S = X$, and X has the induction property. This proves the theorem. ■

Each hypothesis on X has a job, and dropping any one of them breaks the theorem. Each witness below names the direction that fails and the step of the proof it blocks.

- **Drop the predecessor property; $(\Leftarrow)$ fails.** Let $X := \mathbb{N} \cup \{n + \frac{1}{2} : n \in \mathbb{N}\}$. It contains 0 and is closed, since $(n + \frac{1}{2}) + 1 = (n + 1) + \frac{1}{2}$. It has the least-element property: for non-empty $T \subseteq X$, compare a least element of $T \cap \mathbb{N}$ with $\frac{1}{2}$ plus a least element of $\{t - \frac{1}{2} : t \in T \setminus \mathbb{N}\}$, each of the two sets, when non-empty, being a subset of $\mathbb{N}$; the smaller of the candidates is least in T (Exercise C2 asks for the details). But $S := \mathbb{N}$ contains 0, is closed, and is not X , since $\frac{1}{2} \notin \mathbb{N}$ by [Lemma 2.1.1](#) (e). The proof of $(\Leftarrow)$ breaks at “ $m - 1 \in X$ ”: here $m = \frac{1}{2}$ is least in $X \setminus S$, and $m \neq 0$, but $m - 1 = -\frac{1}{2} \notin X$.
- **Drop closure; $(\Rightarrow)$ fails.** Let $X := \{-n : n \in \mathbb{N}\} = \{0, -1, -2, \dots\}$. It contains 0, and it has the predecessor property, since $(-n) - 1 = -(n + 1)$. Its induction property holds **vacuously**: a subset $S \subseteq X$ containing 0 and closed under $n \mapsto n + 1$ would contain 1, and $1 \notin X$, since $1 = -n$ would give $n = -1 < 0$, against [Lemma 2.1.1](#) (a); so no such S exists, and every one of them (there are none) equals X . But X itself has no least element, since $-(n + 1) < -n$. The proof of $(\Rightarrow)$ breaks at “ $\mathbb{N} \subseteq X$ by Logic and Proof, Axiom 4.2.1”: the property “ $n \in X$ ” has no inductive step.
- **Drop $0 \in X$; $(\Rightarrow)$ fails.** Let $X := \{k + \frac{1}{2} : k \in \mathbb{Z}\}$. It is closed, and it has the predecessor property, since $(k + \frac{1}{2}) \pm 1 = (k \pm 1) + \frac{1}{2}$ with $k \pm 1 \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a). Its induction property holds vacuously, because $0 \notin X$, so no subset of X contains 0. (Indeed $0 = k + \frac{1}{2}$ would need $k = -\frac{1}{2} \in \mathbb{Z}$, and then $\frac{1}{2} = -k \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a), which discreteness rules out, as shown above.) And X has no least element, since $(k - 1) + \frac{1}{2} < k + \frac{1}{2}$. The proof of $(\Rightarrow)$ breaks at “ $\mathbb{N} \subseteq X$ by Logic and Proof, Axiom 4.2.1”: the property “ $n \in X$ ” has no base case, since $0 \notin X$.

In the last two witnesses the induction property is not false but vacuous. That is why the theorem uses plain closure of S . Under a “closure relative to X ” reading ($n + 1 \in S$ only when $n + 1 \in X$), the set $S = \{0\}$ would qualify in the closure witness, the induction property would fail there, and the witness would no longer separate the two properties.

So the slogan “induction and well-ordering are the same principle” means exactly [Theorem 2.1.4](#): for a set of numbers with 0, closure and predecessors, either principle implies the other. So *Logic and Proof*, Chapter 3 could have taken, alongside $((\mathbb{N})0)$ and (NS) , the predecessor property and well-ordering as its axioms, and proved induction from them by the $(\Leftarrow)$ half of [Theorem 2.1.4](#). The first witness above shows that the predecessor property cannot be left out, and with induction gone it could no longer be proved as [Lemma 2.1.1](#) (c) was. *Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* , Chapter 1 builds $\mathbb{N}$ from sets and proves the Peano axioms, induction among them, there.

2.1.6 Sets of integers

Well-ordering moves to sets of integers by a shift, provided the bound that makes the shift work is itself an integer.

Corollary 2.1.5. Well-Ordering of Sets of Integers

Let $S \subseteq \mathbb{Z}$ be non-empty. If S has a lower bound $b \in \mathbb{Z}$, then S has a least element. If S has an upper bound $b \in \mathbb{Z}$, then S has a greatest element.

Idea. Slide S to the right by $-b$, so that its lower bound lands on 0; the shifted set lies in $\mathbb{N}$, where [Theorem 2.1.3](#) applies, and sliding back preserves order. For an upper bound, reflect: $-S$ is bounded below by $-b$.

Proof. Lower bound. Let $T := \{s - b : s \in S\}$. Each $s - b$ is an integer by [Proposition 2.1.2](#) (a), and $s - b \geq 0$ since $b \leq s$; so $s - b \in \mathbb{N}$ by [Proposition 2.1.2](#) (b). Thus T is a non-empty subset of $\mathbb{N}$, and by [Theorem 2.1.3](#) it has a least element $m_0 - b$ with $m_0 \in S$. For every $s \in S$, $m_0 - b \leq s - b$, so $m_0 \leq s$. Hence m_0 is a least element of S .

Upper bound. Let $-S := \{-s : s \in S\}$. It is a non-empty set of integers (negatives of integers are integers, by the definition of $\mathbb{Z}$ and [Proposition 2.1.2](#) (a)), and $-b$ is an integer lower bound for it, since $s \leq b$ gives $-b \leq -s$. By the first half, $-S$ has a least element $-m$ with $m \in S$. For every $s \in S$, $-m \leq -s$, so $s \leq m$: m is a greatest element of S . This proves the corollary. ■

The bound must be an integer, or at least have an integer below it. Given a set of integers bounded below by the real number $-\frac{7}{2}$, we would first replace $-\frac{7}{2}$ by the integer -4 , and that step, for an arbitrary real bound, is the Archimedean property of *Analysis on the Real Line*, Chapter 1, whose integer-part lemma in Section 1.6 does exactly this. In an ordered field without the Archimedean property, a set of integers can be bounded below and have no least element (in the field $\mathbb{R}(x)$ of *Analysis on the Real Line*, Chapter 1's example of a non-Archimedean field, every $n \in \mathbb{N}$ has $n < x$, so the set $\{-n : n \in \mathbb{N}\}$ is bounded below by $-x$, and it has no least element, since $-(n+1) < -n$), so this chapter, which has no Archimedean property, proves the corollary with an integer bound, and *Analysis on the Real Line*, Chapter 1 supplies the rest.

2.1.7 Exercises

2.1.7.1 A. Check your understanding

Exercise. A1

Parts (a)–(d) are about the proof of [Theorem 2.1.3](#).

- (a) Where is trichotomy used?
- (b) Why would ordinary induction (Logic and Proof, Axiom 4.2.1) on the same property “ $n \notin S$ ” not do?

- (c) Run the proof on $S = \{4, 7, 9\}$: at which n does the argument meet an element of S , and what does it conclude there?
- (d) State the idea of the proof in two sentences.
- (e) Let $X = \mathbb{N} \cup \{n + \frac{1}{2} : n \in \mathbb{N}\}$. Which statements of [Lemma 2.1.1](#) remain true when every “ $\mathbb{N}$ ” in them is replaced by “ X ”? Decide each clause of (a), each half of (b), and (c), (d), (e) separately, with a witness for each failure.

Solution. (a) In the inductive step, to split an arbitrary $s \in S$ into $s < n$, $s = n$ or $s > n$. The first case is excluded by the hypothesis, so $s \geq n$. Without trichotomy, “ s is not less than n ” would not give “ $s \geq n$ ”, and n would not be shown to be least.

(b) Ordinary induction gives only $P(n-1)$, that is, $n-1 \notin S$, as hypothesis for $P(n)$. That does not exclude a smaller element of S , such as $n-5$, so “ $n \in S$ ” would not make n least, and no contradiction would follow. The proof needs **all** numbers below n to be outside S .

(c) $S = \{4, 7, 9\}$ has a least element, so the supposition of the proof is false, and the argument meets its contradiction. For $n = 0, 1, 2, 3$, no smaller number is in S , and indeed $n \notin S$. At $n = 4$, no natural number below 4 is in S , and $4 \in S$; the argument shows that every $s \in S$ has $s \geq 4$, so 4 is a least element of S . That contradicts the supposition “ S has no least element”, and the proof concludes that S has a least element, namely 4.

(d) If S had no least element, then no n could be in S , because the first one would be least; strong induction makes “the first one” precise. So S would be empty, against the hypothesis.

(e) Holding: the first clause of (a) (every element is ≥ 0); the sum half of (b), since a sum of two halves $(m + \frac{1}{2}) + (n + \frac{1}{2}) = m + n + 1$ is a natural number and a half plus a natural number is a half; and (d): if $x \leq y$ in X , then $y - x \in X$. For (d) there are four cases. Two natural numbers, or two halves, have a difference in $\mathbb{N}$ by [Lemma 2.1.1](#) (d). For $x = m \leq y = n + \frac{1}{2}$, we have $m \leq n$ (else $m \geq n + 1 > y$ by (e) of the lemma), so $y - x = (n - m) + \frac{1}{2} \in X$. For $x = m + \frac{1}{2} \leq y = n$, we have $m < n$, so $m + 1 \leq n$ and $y - x = (n - m - 1) + \frac{1}{2} \in X$.

Failing: the second clause of (a), since $\frac{1}{2} \neq 0$ and $\frac{1}{2} < 1$; the product half of (b), since $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \notin X$ (it is not a natural number, as $0 < \frac{1}{4} < 1$, and not a half, as $\frac{1}{4} - \frac{1}{2} < 0$); (c), since $\frac{1}{2} \neq 0$ and $\frac{1}{2} - 1 = -\frac{1}{2} \notin X$; (e), since $0 < \frac{1}{2} < 1$ with $\frac{1}{2} \in X$.

Each clause had to be decided by its own witness. The proofs of (a)–(d) were inductions over $\mathbb{N}$, and X does not have the induction property ($\mathbb{N} \subseteq X$ contains 0 and is closed), so none of those proofs carries over; some conclusions survive anyway. (c) fails because the predecessor property fails at $\frac{1}{2}$. (e) fails through the second clause of (a), not through (d), which still holds. The set $\mathbb{Z} \cup \{k + \frac{1}{2} : k \in \mathbb{Z}\}$ has the predecessor property and still has $\frac{1}{2}$ between 0 and 1, so the predecessor property alone (with 0 and closure) does not give (e); together with the least-element property it does: the set then has the induction property by [Theorem 2.1.4](#), and, as in the proof of its ($\Rightarrow$) half, it equals $\mathbb{N}$.

2.1.7.2 B. Practice

Exercise. B1

Use only [Lemma 2.1.1](#) and the arithmetic of $\mathbb{R}$.

- (a) Prove that if $m, n \in \mathbb{N}$ and $mn = 1$, then $m = n = 1$.

(b) Prove that $n^2 \geq n$ for every $n \in \mathbb{N}$.

Solution. (a) Since $0 \cdot n = 0 \neq 1$, we have $m \neq 0$, and likewise $n \neq 0$. By the second clause of Lemma 2.1.1 (a), $m \geq 1$ and $n \geq 1$. Suppose $m > 1$. Multiplying by $n > 0$ gives $mn > n \geq 1$, which contradicts $mn = 1$. So $m = 1$, and then $n = mn = 1$.

(b) Let $n \in \mathbb{N}$. By Lemma 2.1.1 (a), $n = 0$ or $n \geq 1$. If $n = 0$, then $n^2 = 0 = n$. If $n \geq 1$, multiplying $n \geq 1$ by $n > 0$ gives $n^2 \geq n$. In both cases $n^2 \geq n$.

Exercise. B2: Putting a proof in order

The lines below, (i)–(ix), include the seven lines of a proof of $(\Leftarrow)$ in Theorem 2.1.4, shuffled, and two lines that do not belong. Here X has $0 \in X$, closure and the predecessor property, and the least-element property. Write the seven lines in an order that makes a correct proof, and say why each of the other two lines fails.

- (i) So $m - 1 \in X$, by the predecessor property.
- (ii) Let m be the least element of S .
- (iii) Then $X \setminus S$ is non-empty, so it has a least element m .
- (iv) Hence $m = (m - 1) + 1 \in S$, a contradiction; so $S = X$.
- (v) Let $S \subseteq X$ contain 0 and be closed under $n \mapsto n + 1$.
- (vi) $m - 1 < m$, so $m - 1 \notin X \setminus S$, that is, $m - 1 \in S$.
- (vii) Since m is least and $m - 1 < m$, $m - 1 \in X \setminus S$.
- (viii) Suppose $S \neq X$.
- (ix) $m \neq 0$, since $0 \in S$.

Solution. The order is (v), (viii), (iii), (ix), (i), (vi), (iv). It is the only admissible one. (v) introduces S , (viii) is the supposition the rest contradicts, and (iii) introduces m from it. (ix) must sit between (iii) and (i): it needs the m of (iii), and (i) applies the predecessor property, which needs $m \neq 0$. (vi) uses the $m - 1 \in X$ of (i): only an element of X can be sorted into S or $X \setminus S$. (iv) uses (vi) and closure.

Line (ii) takes the least element of the wrong set. The set whose least element locates a failure is $X \setminus S$: its least element is the first number of X missing from S . A least element of S says nothing about where S and X differ; in the case $X = \mathbb{N}$ it is 0, which lies in S by (v) and is no failure at all. Line (vii) reads leastness backwards: “ m is least in $X \setminus S$ ” says that nothing **below** m lies in $X \setminus S$, so a smaller number of X is in S , not outside it.

Exercise. B3

Prove the second half of Corollary 2.1.5 directly from Theorem 2.1.3, without using the first half: if $S \subseteq \mathbb{Z}$ is non-empty and $b \in \mathbb{Z}$ is an upper bound of S , then S has a greatest element.

Hint: shift by $s \mapsto b - s$.

Solution. Let $T := \{b - s : s \in S\}$. For $s \in S$, $b - s$ is an integer by [Proposition 2.1.2](#) (a), and $b - s \geq 0$ because $s \leq b$; so $b - s \in \mathbb{N}$ by [Proposition 2.1.2](#) (b). Thus T is a non-empty subset of $\mathbb{N}$, and by [Theorem 2.1.3](#) it has a least element, of the form $b - s_0$ with $s_0 \in S$. For every $s \in S$, $b - s_0 \leq b - s$, so $s \leq s_0$. Hence s_0 is a greatest element of S .

2.1.7.3 C. Going deeper

Exercise. C1: A greatest element from a least upper bound

Let $S \subseteq \mathbb{N}$ be non-empty, and suppose some $M \in \mathbb{N}$ satisfies $s \leq M$ for every $s \in S$.

- (a) Prove that S has a greatest element by applying [Theorem 2.1.3](#) to the set U of upper bounds of S that lie in $\mathbb{N}$, not by citing [Corollary 2.1.5](#): show that the least element of U lies in S .
- (b) Show that the hypothesis “some $M \in \mathbb{N}$ bounds S above” cannot be dropped.

Solution (Hint 1). Consider $U := \{u \in \mathbb{N} : s \leq u \text{ for every } s \in S\}$.

Solution (Hint 2). U is non-empty because $M \in U$; let u be its least element, which exists by [Theorem 2.1.3](#).

Solution. (a) Let $U := \{u \in \mathbb{N} : s \leq u \text{ for every } s \in S\}$. Then $M \in U$, so U is a non-empty subset of $\mathbb{N}$, and [Theorem 2.1.3](#) gives a least element u of U . We show $u \in S$.

Suppose $u \notin S$. Then every $s \in S$ satisfies $s \leq u$ and $s \neq u$, so $s < u$, and $s + 1 \leq u$ by [Lemma 2.1.1](#) (e); that is, $s \leq u - 1$. Since S is non-empty, it has an element s_0 , and $0 \leq s_0 < u$ by [Lemma 2.1.1](#) (a), so $u \neq 0$. By [Lemma 2.1.1](#) (c), $u - 1 \in \mathbb{N}$. So $u - 1$ is an upper bound of S in $\mathbb{N}$, that is, $u - 1 \in U$, and $u - 1 < u$, which contradicts the choice of u . Hence $u \in S$.

Since $u \in S$ and $s \leq u$ for every $s \in S$, u is a greatest element of S .

(b) $S = \mathbb{N}$ is non-empty and has no greatest element: for every $n \in \mathbb{N}$, also $n + 1 \in \mathbb{N}$ by (NS), and $n + 1 > n$. In the proof, U is then empty, and [Theorem 2.1.3](#) has nothing to apply to.

The number u of (a) is the **least upper bound** of S among natural numbers, and the proof shows that it lies in S . Section 2.6 defines least upper bounds in any partially ordered set, and *Analysis on the Real Line*, Chapter 1 is about sets of real numbers whose least upper bound, when it exists, need **not** lie in the set.

Marking scheme. (10 points) - 2: defines the set of upper bounds in $\mathbb{N}$, shows it is non-empty, and takes its least element by [Theorem 2.1.3](#). - 2: handles $u = 0$ separately, or shows $u \neq 0$ using that S is non-empty, before using $u - 1 \in \mathbb{N}$ ([Lemma 2.1.1](#) (c)). - 3: uses the minimality of u : either shows directly that $u - 1 \notin U$, so some $s \in S$ has $s > u - 1$, or, supposing $u \notin S$, derives $u - 1 \in U$ and contradicts minimality. - 1: uses discreteness ([Lemma 2.1.1](#) (e)) to show that an element of S greater than $u - 1$ equals u , or to get

$s \leq u - 1$ in the contradiction route, and concludes that $u \in S$ is a greatest element. - 2: (b) the witness $S = \mathbb{N}$, with the reason it has no greatest element. Fatal errors (0 for the step): using $u - 1 \in \mathbb{N}$ without checking $u \neq 0$; taking a least element of U without showing U non-empty.

A second route, through the least element $M - s_0$ of $\{M - s : s \in S\} \subseteq \mathbb{N}$ (as in [Exercise \(B3\)](#)), proves the conclusion of (a) but not by the method the task names, so it earns at most 4 points for (a): 2 for showing that this set is a non-empty subset of $\mathbb{N}$ and taking its least element, and 2 for deducing $s \leq s_0$ for every $s \in S$ from $M - s_0 \leq M - s$, with $s_0 \in S$; and (b) as above.

Exercise. C2: A well-ordered set with gaps

Let $X := \mathbb{N} \cup \{n + \frac{1}{2} : n \in \mathbb{N}\}$. Prove that every non-empty subset of X has a least element.

Solution. Let $T \subseteq X$ be non-empty, and put

$$A := T \cap \mathbb{N}, \quad B := \{t - \frac{1}{2} : t \in T \setminus \mathbb{N}\}.$$

Every $t \in T \setminus \mathbb{N}$ is $n + \frac{1}{2}$ for some $n \in \mathbb{N}$, since $t \in X$; so $t - \frac{1}{2} = n \in \mathbb{N}$, and $B \subseteq \mathbb{N}$. Since $T = A \cup (T \setminus \mathbb{N})$ is non-empty, A or B is non-empty. By [Theorem 2.1.3](#), A , if non-empty, has a least element a , and B , if non-empty, has a least element b . Define

$$c := \begin{cases} a & \text{if } B = \emptyset, \\ b + \frac{1}{2} & \text{if } A = \emptyset, \\ \min\{a, b + \frac{1}{2}\} & \text{otherwise.} \end{cases}$$

In each case $c \in T$: $a \in A \subseteq T$, and $b = t - \frac{1}{2}$ for some $t \in T \setminus \mathbb{N}$, so $b + \frac{1}{2} = t \in T$.

Let $t \in T$. If $t \in \mathbb{N}$, then $t \in A$, so $A \neq \emptyset$ and $t \geq a \geq c$. If $t \notin \mathbb{N}$, then $t - \frac{1}{2} \in B$, so $B \neq \emptyset$ and $t - \frac{1}{2} \geq b$, that is, $t \geq b + \frac{1}{2} \geq c$. (In both chains, the last inequality holds in every case of the definition of c that can occur.) So c is a least element of T .

The set X has the least-element property, contains 0 and is closed under $n \mapsto n + 1$, but it does not have the induction property ($\mathbb{N} \subseteq X$ contains 0 and is closed, and $\mathbb{N} \neq X$). That is the first witness after [Theorem 2.1.4](#): without the predecessor property, well-ordering does not give induction.

Marking scheme. (10 points) - 2: splits T into its natural-number part and its half part, and shows the shifted half part lies in $\mathbb{N}$. - 2: applies [Theorem 2.1.3](#) to each non-empty part, saying that at least one part is non-empty. - 2: defines the candidate in all cases, including those where one part is empty, and checks it lies in T . - 3: checks the candidate is $\leq t$ for every $t \in T$, by cases on $t \in \mathbb{N}$ or not. - 1: concludes, with the quantifiers right: one least element for the given T . Fatal errors (0 for the step): applying [Theorem 2.1.3](#) to T itself (it need not be a subset of $\mathbb{N}$); taking a least element of an empty part.

A second route, through the least element of $\{2t : t \in T\} \subseteq \mathbb{N}$, earns the same points. It earns 4 for showing that this set lies in $\mathbb{N}$, checking both $2n$ and $2n + 1$ by [Lemma 2.1.1](#) (b), and that it is non-empty, and for taking its least element $2t_0$ with $t_0 \in T$ by [Theorem 2.1.3](#).

It earns 4 for deducing $t_0 \leq t$ from $2t_0 \leq 2t$ for every $t \in T$, and 2 for the conclusion. The same fatal errors apply.

2.2 Induction, Strong Induction and the Smallest Counterexample

Logic and Proof, Chapter 3 gave the principle of induction (Logic and Proof, Axiom 4.2.1), its strong form (Logic and Proof, Theorem 4.2.2), and the first worked proofs: a sum (Logic and Proof, Example (A sum of products of neighbors)), an inequality (Logic and Proof, Example (An inequality by induction)), and a strong induction with three base cases (Logic and Proof, Example (Threes and fives)). *Logic and Proof*, Chapter 5 showed how an induction is written in prose and which rule each phrase stands for (Logic and Proof, Example (An induction, and a legitimate symmetry)). This section is about the craft: which form of induction fits a claim, where it starts, how a smallest counterexample becomes a proof, and how an induction goes wrong, each failure located at one step. On the way it proves facts that *Logic and Proof*, Chapters 1, 2 and 5 took from school: every integer is even or odd and not both, a fraction can be reduced until numerator and denominator are not both even, and every integer $n \geq 2$ has a prime factor.

2.2.1 Induction from any integer

Logic and Proof, Chapter 3 already started inductions at numbers other than 0 (*Logic and Proof*, Section 3.2, “Starting from another number”), for a starting point $b \in \mathbb{N}$. With $\mathbb{Z}$ and Proposition 2.1.2 in hand, the starting point may be any integer, negative ones included.

To prove a statement for every integer from n_0 on, prove it at n_0 and show that it passes from each integer to the next.

Theorem 2.2.1. Induction from a Base

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

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

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

Idea. Count from the base. The integers $n \geq n_0$ are exactly $n_0 + k$ with $k \in \mathbb{N}$, and moving from n to $n + 1$ moves k to $k + 1$. So the statement “ $P(n_0 + k)$ ” about $k \in \mathbb{N}$ is ready for Logic and Proof, Axiom 4.2.1, and its base case is $P(n_0)$.

Proof. For $k \in \mathbb{N}$, let $Q(k)$ be the statement $P(n_0 + k)$. This makes sense: $n_0 + k$ is an integer by Proposition 2.1.2 (a), and $n_0 + k \geq n_0$ since $k \geq 0$ by Lemma 2.1.1 (a).

Base case. $Q(0)$ is $P(n_0)$, which holds by (i).

Inductive step. Let $k \in \mathbb{N}$, and suppose $Q(k)$, that is, $P(n_0 + k)$. The integer $n := n_0 + k$

satisfies $n \geq n_0$, so (ii) gives $P(n+1)$. Since $n+1 = n_0 + (k+1)$, this is $Q(k+1)$.

By Logic and Proof, Axiom 4.2.1, $Q(k)$ holds for every $k \in \mathbb{N}$. Now let n be an integer with $n \geq n_0$. Then $k := n - n_0$ is an integer by Proposition 2.1.2 (a), and $k \geq 0$, so $k \in \mathbb{N}$ by Proposition 2.1.2 (b). Hence $Q(k)$ holds, and $Q(k)$ is $P(n_0 + k) = P(n)$. This proves the theorem. ■

For $n_0 = 0$ the theorem is Logic and Proof, Axiom 4.2.1 itself, and for $n_0 \in \mathbb{N}$ it is *Logic and Proof*, Chapter 3's induction started at n_0 . In the inductive step, " $P(n)$ is true" is the induction hypothesis: one implication, proved once, inside a $\rightarrow I$ box whose assumption is $P(n)$ for one fixed but arbitrary integer $n \geq n_0$.

The strong form moves to a base in the same way.

Theorem 2.2.2. Strong Induction from a Base

Let n_0 be an integer, and for each integer $n \geq n_0$ let $P(n)$ be a statement. Suppose that, **for every** integer $n \geq n_0$: if $P(k)$ is true for **every** integer k with $n_0 \leq k < n$, then $P(n)$ is true. Then $P(n)$ is true for every integer $n \geq n_0$.

Proof. For $j \in \mathbb{N}$, let $Q(j)$ be $P(n_0 + j)$, as in the proof of Theorem 2.2.1. We check the hypothesis of Logic and Proof, Theorem 4.2.2 for Q . Let $j \in \mathbb{N}$, and suppose $Q(i)$ holds for every $i \in \mathbb{N}$ with $i < j$. Put $n := n_0 + j$, an integer with $n \geq n_0$. Let k be an integer with $n_0 \leq k < n$. Then $i := k - n_0$ is an integer with $i \geq 0$, so $i \in \mathbb{N}$ by Proposition 2.1.2 (b), and $i < n - n_0 = j$; so $Q(i)$, which is $P(k)$, holds. The hypothesis of the theorem at n now gives $P(n)$, that is, $Q(j)$. By Logic and Proof, Theorem 4.2.2, $Q(j)$ holds for every $j \in \mathbb{N}$, and every integer $n \geq n_0$ is $n_0 + j$ with $j = n - n_0 \in \mathbb{N}$, as in the proof of Theorem 2.2.1. This proves the theorem. ■

So Theorem 2.2.2 is a corollary of Logic and Proof, Theorem 4.2.2, by the same shift. When $n_0 \in \mathbb{N}$, it says what Logic and Proof, Example (Threes and fives) achieved with a guard: there the property was "if $n \geq 8$, then ...", stated for every $n \in \mathbb{N}$ and vacuous below 8. Here the property is stated only for $n \geq n_0$, and the hypothesis speaks only about $k \geq n_0$. The two are the same argument in two notations, and the base cases sit in the same place: they are the n for which the step reaches below n_0 .

Quick check. Every integer $n \geq 24$ is $5a + 7b$ for some integers $a, b \geq 0$. A proof by Theorem 2.2.2 writes $n = (n - 5) + 5$ in its step. How many base cases does it need, and which? Can the base 24 be lowered?

Solution. The step uses $P(n - 5)$, which the hypothesis supplies only when $n - 5 \geq 24$, that

is, $n \geq 29$. So $n = 24, 25, 26, 27, 28$ are base cases, five of them, checked directly:

$$\begin{aligned} 24 &= 5 \cdot 2 + 7 \cdot 2, & 25 &= 5 \cdot 5, & 26 &= 5 \cdot 1 + 7 \cdot 3, \\ 27 &= 5 \cdot 4 + 7 \cdot 1, & 28 &= 7 \cdot 4. \end{aligned}$$

The base cannot be lowered: 23 is not of the form. With $b = 0, 1, 2, 3$ the remainder $23 - 7b$ is 23, 16, 9, 2, none a multiple of 5, and $b \geq 4$ gives $23 - 7b < 0$.

2.2.2 The smallest counterexample

Well-ordering gives the same power as strong induction, in a form that often reads more naturally. To prove a claim about every integer from n_0 on, suppose it fails somewhere. The failures form a non-empty set of integers with the integer lower bound n_0 , so [Corollary 2.1.5](#) gives a least one. Below it the claim holds. If that is enough to produce a failure that is smaller still, or to show the claim at the least failure after all, we have a contradiction.

Algorithm. To prove “ $P(n)$ for every integer $n \geq n_0$ ” by a **smallest counterexample**:

- *Suppose it fails:* some integer $n \geq n_0$ has $P(n)$ false.
- *Take the least failure:* the set F of integers $n \geq n_0$ with $P(n)$ false is non-empty and has the integer lower bound n_0 , so it has a least element m ([Corollary 2.1.5](#); [Theorem 2.1.3](#) when $n_0 = 0$).
- *Use the claim below m :* $P(k)$ holds for every integer k with $n_0 \leq k < m$, since such a k is not in F .
- *Produce a contradiction:* a failure $k < m$, or $P(m)$ itself.

Then F is empty, and $P(n)$ holds for every integer $n \geq n_0$.

The method is strong induction with an RAA box drawn around it. The work done inside the box, “if $P(k)$ for every k with $n_0 \leq k < m$, then $P(m)$ ”, is exactly the hypothesis of [Theorem 2.2.2](#) at m ; so any smallest-counterexample proof can be rewritten as a strong induction, and the reverse rewriting works too. Which one to write is a matter of which reads better. A smallest counterexample suits claims whose failure has a natural “smaller failure”, like a fraction that can be reduced, or a number with a factor.

2.2.3 Divisibility and the division algorithm

The first uses of the method are about divisibility, and here is the vocabulary.

Definition 2.2.1. Divisibility

Let $a, b \in \mathbb{Z}$. We say that a **divides** b , written $a \mid b$, if **there exists** $k \in \mathbb{Z}$ with $b = ka$. Then a is a **divisor** of b , and b is a **multiple** of a .

So $3 \mid 12$, with $k = 4$, and $-3 \mid 12$, with $k = -4$. Every integer a divides 0, with $k = 0$; this is the degenerate case, and it shows that “ $a \mid b$ ” is a statement, not the number b/a : $0 \mid 0$ is true, while $0/0$ is not defined. Also $1 \mid b$ and $b \mid b$ for every integer b , and “ n is even” is “ $2 \mid n$ ”. If $a \mid b$ and $b \mid c$, then $a \mid c$: from $b = ka$ and $c = lb$ we get $c = (lk)a$, with $lk \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a).

Change 12 to 14: $3 \nmid 14$. The witness k would have to be $14/3$, and $4 < 14/3 < 5$, so $14/3$ is not an integer, by the discreteness clause of [Proposition 2.1.2](#). The clause that fails is “ $k \in \mathbb{Z}$ ”: the equation $14 = k \cdot 3$ has a real solution, just not an integer one.

Divisibility claims about powers are a staple of induction, because the step usually factors out one copy. For instance, $3 \mid 4^n - 1$ for every $n \in \mathbb{N}$, with 4^n in its school meaning ($4^0 = 1$, $4^{n+1} = 4^n \cdot 4$; Section 2.3 makes this a definition). Base case: $4^0 - 1 = 0 = 3 \cdot 0$. Step: if $4^n - 1 = 3k$ with $k \in \mathbb{Z}$, then

$$4^{n+1} - 1 = 4(4^n - 1) + 3 = 3(4k + 1),$$

and $4k + 1 \in \mathbb{Z}$. By Logic and Proof, Axiom 4.2.1, the claim holds for every $n \in \mathbb{N}$.

Definition 2.2.2. Prime Number

An integer p is **prime** if $p \geq 2$ and the only **positive** divisors of p are 1 and p . A **prime factor** of an integer n is a prime p with $p \mid n$.

For example, 2 is prime. Let $d > 0$ divide 2, say $2 = kd$ with $k \in \mathbb{Z}$. Then $k > 0$, so $k \geq 1$ by [Proposition 2.1.2](#) (b), and $d \leq kd = 2$; by discreteness $d \in \{1, 2\}$. The same argument shows that 3 and 7 are prime, once the candidates $d = 2$ (and $d = 2, \dots, 6$ for 7) are ruled out as $3 \nmid 14$ was. Non-examples: 9 is not prime, since $9 = 3 \cdot 3$ gives the positive divisor $3 \notin \{1, 9\}$; and 1 is not prime, though its only positive divisor is 1, because the definition demands $p \geq 2$. The clause $p \geq 2$ keeps 1 out so that factorizations into primes can be unique, a fact this chapter does not prove or use: Exercise B2 below is about existence only.

Every integer, divided by a positive integer n , leaves a remainder between 0 and $n - 1$. This is school arithmetic, and it is the tool behind “even or odd”.

Theorem 2.2.3. Division Algorithm

Let $a \in \mathbb{Z}$ and $n \in \mathbb{Z}_{>0}$. Then there are **unique** integers q and r with

$$a = qn + r \quad \text{and} \quad 0 \leq r < n.$$

The hypothesis $n > 0$ matters: for $n = 0$ no r satisfies $0 \leq r < 0$.

Try it first (ten minutes): which set of non-negative integers should have a least element, and why is that least element the remainder?

Idea. The obvious attempt subtracts n from a again and again until what is left lies in $[0, n)$, adding n instead when a is negative. “Again and again until” is a process, not yet a proof: why does it stop, and why at a number below n ? Well-ordering replaces the process by its outcome. Among all the non-negative integers $a - qn$, take the **least**. If it were n or more, subtracting one more n would leave a smaller non-negative one. For uniqueness: two remainders differ by a multiple of n , yet lie less than n apart, so the multiple is 0.

Proof. Existence. Let

$$R := \{a - qn : q \in \mathbb{Z}\} \cap \{x \in \mathbb{Z} : x \geq 0\},$$

the set of non-negative integers of the form $a - qn$ (each $a - qn$ is an integer by [Proposition 2.1.2](#) (a)). The set R is non-empty. If $a \geq 0$, take $q = 0$: then $a - 0 \cdot n = a \in R$. If $a < 0$, take $q = a$: then $a - an = a(1 - n) \geq 0$, since $a < 0$ and $1 - n \leq 0$ (as $n \geq 1$ by [Proposition 2.1.2](#) (b)). The set R is a non-empty set of integers with the integer lower bound 0, so by [Corollary 2.1.5](#) it has a least element r ; write $r = a - qn$ with $q \in \mathbb{Z}$. Then $a = qn + r$ and $r \geq 0$.

Suppose $r \geq n$. Then $r - n = a - (q + 1)n$ is an integer of the required form with $r - n \geq 0$, so $r - n \in R$; and $r - n < r$, since $n > 0$. This contradicts the choice of r . Hence $r < n$.

Uniqueness. Let $a = qn + r = q'n + r'$ with $q, q', r, r' \in \mathbb{Z}$, $0 \leq r < n$ and $0 \leq r' < n$. Suppose $r \neq r'$. By symmetry (exchanging (q, r) and (q', r') changes neither the hypotheses nor the goal) we may suppose $r' < r$. Then $r - r' = (q' - q)n$, and $k := q' - q$ is an integer with $kn = r - r' > 0$. Since $n > 0$, also $k > 0$, so $k \geq 1$ by [Proposition 2.1.2](#) (b), and

$$r - r' = kn \geq n.$$

But $r - r' \leq r < n$, since $r' \geq 0$. This contradiction shows $r = r'$. Then $qn = q'n$, and since $n \neq 0$, $q = q'$. This proves the theorem. ■

The integers q and r are the **quotient** and the **remainder** of a on division by n . For negative a the remainder is still non-negative: $-7 = 3 \cdot (-3) + 2$, with remainder 2, not -1 . Existence and uniqueness do separate jobs, and the case $n = 2$ shows both:

- **Existence** says that every integer a is $2q$ or $2q + 1$: every integer is **even or odd**.
- **Uniqueness** says that no integer is both. If $a = 2q = 2q' + 1$, then a would have remainders 0 and 1.

This is the school fact that *Logic and Proof*, Chapter 1 used in Argument A and *Logic and Proof*, Chapter 2 used in the proof of Logic and Proof, Theorem 2.1.3, now proved. The same uniqueness shows that no integer m has $2m = 1$, since $1 = 2 \cdot 0 + 1$ has remainder 1 and $2m = 2m + 0$ has remainder 0; this is the fact *Analysis on the Real Line*, Chapter 1 uses to show that $\mathbb{Z}$ is not a field. An integer $4k + 1 = 2(2k) + 1$ or $4k + 3 = 2(2k + 1) + 1$ is odd. With $n = 3$, every integer is $3k$, $3k + 1$ or $3k + 2$, for exactly one of the three forms.

Warning. “The remainder” is the r of the theorem, with $0 \leq r < n$. The equations $17 = 5 \cdot 4 - 3$ and $17 = 5 \cdot 2 + 7$ are true, but -3 and 7 are not remainders, and allowing them destroys uniqueness: 17 would have the three “remainders” -3 , 2 and 7. Every argument above that concludes from “remainder 1” that a number is not a multiple of 2 uses the range $0 \leq r < n$, through uniqueness.

2.2.4 Lowest terms and prime factors

Logic and Proof, Chapter 2 needed one more fact about fractions, the fact (L) in the proof of Logic and Proof, Theorem 2.2.1: every rational number can be written with numerator and denominator not both even. The reason given there was that halving cannot go on for ever. Here is that reason as a proof, written in the steps of the smallest-counterexample method, each step named.

Lemma 2.2.4. Lowest Terms, Up to Factors of Two

Every rational number can be written as p/q with $p \in \mathbb{Z}$, $q \in \mathbb{Z}_{>0}$, and p and q **not both** even.

Idea. A representation with both entries even can be halved. Halving makes the denominator smaller, so a least denominator cannot belong to a representation that can be halved. Well-ordering supplies the least denominator.

Proof. Suppose it fails. Fix $r \in \mathbb{Q}$, and suppose that r has no representation p/q with $p \in \mathbb{Z}$, $q \in \mathbb{Z}_{>0}$, and p and q not both even. Let

$$F := \{q \in \mathbb{Z}_{>0} : r = p/q \text{ for some } p \in \mathbb{Z}\}.$$

Under the supposition, every $q \in F$ is the denominator of a failing representation.

Take the least failure. By the definition of $\mathbb{Q}$, $r = p/q$ for some $p \in \mathbb{Z}$ and $q \in \mathbb{Z}_{>0}$, so F is non-empty; and $F \subseteq \mathbb{Z}_{>0} \subseteq \mathbb{N}$. By [Theorem 2.1.3](#), F has a least element m . Write $r = p/m$ with $p \in \mathbb{Z}$.

Use the supposition at m . The representation $r = p/m$ fails, so p and m are both even: $p = 2p'$ and $m = 2m'$ with $p', m' \in \mathbb{Z}$.

Produce a smaller failure. Since $m > 0$, $m' = m/2 > 0$, so $m' \in \mathbb{Z}_{>0}$ by [Proposition 2.1.2](#) (b). Moreover $r = 2p'/(2m') = p'/m'$, so $m' \in F$; and $m' < m' + m' = m$. So m' is an element of F smaller than its least element m , a contradiction.

Hence r has a representation of the required kind. As $r \in \mathbb{Q}$ was arbitrary, this proves the lemma. ■

This is (L), with more: the denominator can be taken positive, not merely non-zero. The proof of Logic and Proof, Theorem 2.2.1 now rests on a proved lemma. The same argument with “not both even” replaced by “no common divisor $d \geq 2$ ” puts every fraction in lowest terms in the full sense; that is Exercise C2. Two of the proof’s steps adapt those of the smallest-counterexample method above: the failures are the denominators in F , and the supposition makes every element of F one, so the third step needs only the supposition at m , not the claim below m , and the contradiction of the fourth step is the smaller failure m' . This proof is written out in full; two later exercises use the same move with less given: Exercise B2 below leaves its last step blank, and Exercise C1 of Section 2.3 leaves the last half blank.

The next result is the prime-factor fact that *Logic and Proof*, Chapter 5 used. Its proof is a strong induction from base 2; the same argument inside an RAA box is a smallest-counterexample proof, in which the least failure m gets a prime factor from its divisor a , contradicting the choice of m (compare Exercise B2).

Theorem 2.2.5. Every Integer at Least Two Has a Prime Factor

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

Idea. A prime is its own prime factor. A number that is not prime has a divisor strictly between 1 and itself; that divisor is smaller, so it has a prime factor by the strong hypothesis, and divisibility passes it on.

Proof. We use [Theorem 2.2.2](#) with $n_0 = 2$. Let $n \geq 2$ be an integer, and suppose every integer k with $2 \leq k < n$ has a prime factor.

If n is prime, then n is a prime factor of itself, since $n = 1 \cdot n$.

If n is not prime, then, as $n \geq 2$, it has a positive divisor $a \notin \{1, n\}$ ([Definition 2.2.2](#)). Write $n = ka$ with $k \in \mathbb{Z}$. Since $n > 0$ and $a > 0$, also $k > 0$, so $k \geq 1$ and $a \geq 1$ by [Proposition 2.1.2](#) (b). Hence $a \leq ka = n$. With $a \neq 1$ and $a \neq n$ this gives $1 < a < n$, and so $2 \leq a < n$ by the discreteness clause of [Proposition 2.1.2](#). By the hypothesis, a has a prime factor p , say $a = lp$ with $l \in \mathbb{Z}$. Then $n = ka = (kl)p$, with $kl \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a), so $p \mid n$.

In both cases n has a prime factor. By [Theorem 2.2.2](#), every integer $n \geq 2$ has one. This proves the theorem. ■

The hypothesis was used at a , which can be far below $n - 1$: for $n = 91 = 7 \cdot 13$ it is used at 7 or at 13. Ordinary induction, which offers only $n - 1$, would have nothing to say about 91 from 90.

2.2.5 How induction proofs go wrong

An induction proof has few moving parts: a property, a base case, a step, and a citation of the right principle. Each can fail, and each failure has a place in the proof where it can be pointed at. Two questions are kept apart throughout: is the conclusion true, and does the argument prove it?

Example. All Numbers Are Equal

Here is a “theorem” and a “proof”. Find the line that fails.

“*Theorem.*” For every integer $n \geq 1$ and all real numbers $x_1, \dots, x_n$, we have $x_1 = x_2 = \dots = x_n$.

“*Proof.*” For an integer $n \geq 1$, let $P(n)$ be the statement “for all real numbers $x_1, \dots, x_n$, all the entries of the list $x_1, \dots, x_n$ are equal”. We use [Theorem 2.2.1](#) with $n_0 = 1$.

- (1) *Base case.* A list x_1 of one number has only one entry, so all its entries are equal: $P(1)$ holds.
- (2) *Step.* Let $n \geq 1$ be an integer, suppose $P(n)$, and let $x_1, \dots, x_{n+1}$ be real numbers.
- (3) The list $x_1, \dots, x_n$ has n entries, so by $P(n)$ its entries are all equal.
- (4) The list $x_2, \dots, x_{n+1}$ has n entries, so by $P(n)$ its entries are all equal.
- (5) The entry x_2 lies in both lists, so every entry of either list equals x_2 .
- (6) Hence $x_1 = x_2 = \dots = x_{n+1}$, which is $P(n + 1)$. By [Theorem 2.2.1](#), $P(n)$ holds for

every $n \geq 1$.

Solution. Is the conclusion true? No: the list $x_1 = 0, x_2 = 1$ has unequal entries, so $P(2)$ is false.

Does the argument prove it? No, and the failing line is (5) at $n = 1$. Then the two lists of lines (3) and (4) are x_1 and x_2 , one entry each, and x_2 is not in the first. They share an entry exactly when the first list, $x_1, \dots, x_n$, contains x_2 , that is, when $n \geq 2$. In the terms of Logic and Proof, Example (What is not allowed), line (5) cites a “fact” that is false in the case at hand: entry (N9).

Everything else is correct. The base case (1) is true. For every $n \geq 2$ the step (2)–(6) is valid: if all lists of n reals had equal entries, so would all lists of $n + 1$. So the induction fails at exactly one transition, from 1 to 2, and the valid transitions from 2 on carry nothing, because $P(2)$ is false and every later $P(n)$ with it. A counterexample to the illegal step, as opposed to the conclusion: $n = 1, x_1 = 0, x_2 = 1$, where lines (3) and (4) hold and (5) does not.

statement	true?	proved by the argument?
$P(1)$	yes	yes, line (1)
$P(1) \Rightarrow P(2)$	no	no: line (5) fails
$P(n) \Rightarrow P(n + 1)$ for $n \geq 2$	yes	yes, lines (2)–(6)
$P(n)$ for every $n \geq 1$	no	no

This is the failure that *Logic and Proof*, Chapter 3’s warning about $2^n \geq n^2$ foreshadowed: a step that holds only from some point on, with a gap at the first transition. A step must be proved for **every** $n \geq n_0$, and the place to look for a gap is the smallest n , where the objects the step talks about are smallest.

The other failures are quicker to state.

The missing base case. “ $n = n + 1$ for every $n \in \mathbb{N}$.” The step is valid: if $n = n + 1$, adding 1 to both sides gives $n + 1 = n + 2$. But $P(0)$, “ $0 = 1$ ”, is false, and so is every $P(n)$. The step is a true implication between false statements, and an induction with no true base proves nothing.

The hypothesis used out of range. A strong induction from a base n_0 may use $P(k)$ only for $n_0 \leq k < n$. A step that writes n in terms of $n - 3$ needs $n - 3 \geq n_0$, and for the first few n that fails. Logic and Proof, Example (Threes and fives) handled this correctly, with three base cases; the exit check C1 below is a proof that did not.

The claim too weak to carry its own step. Sometimes the statement is true and the base case holds, but the inductive step cannot be completed, because the hypothesis gives too little. The repair is to prove **more**. Section 2.3 shows a sum bounded by 2 whose own step will not go through, and which a stronger bound carries.

Warning. The induction hypothesis is “ $P(n)$ ” for one fixed but arbitrary n : the assumption of the $\rightarrow$ I box inside the $\forall$ I box of the step (Logic and Proof, Definition 5.4.1). A step that begins “assume that $P(n)$ holds for every n ” assumes the conclusion of the theorem, and whatever follows is circular: entry (N9) of Logic and Proof, Example (What is not allowed).

The witness is a “for every n ” in the assumption that binds the induction variable n itself; it should not be there. A quantifier over another variable is part of the statement and may stay: “for every k with $n_0 \leq k < n$ ” in strong induction, or the “for every n ” inside $P(m)$ in the proof of [Lemma 2.1.1](#) (d), where the induction is on m .

2.2.6 Exercises

2.2.6.1 A. Check your understanding

Exercise. A1

- For each claim, name the form of proof that fits it most naturally (ordinary induction from 0, induction from another base, strong induction from a base with several base cases, or a smallest counterexample), and say in one clause why. Do not prove the claims. (i) $6 \mid 7^n - 1$ for every $n \in \mathbb{N}$. (ii) $2^n > n^2$ for every integer $n \geq 5$. (iii) Every integer $n \geq 12$ is $3a + 7b$ for some integers $a, b \geq 0$. (iv) There are no $m, n \in \mathbb{Z}_{>0}$ with $n^2 = 2m^2$.
- Negate “every integer $n \geq 2$ has a prime factor”, and say what a smallest counterexample to it would be.
- In [Example \(All Numbers Are Equal\)](#), what is the smallest n for which the two lists $x_1, \dots, x_n$ and $x_2, \dots, x_{n+1}$ of lines (3) and (4) share an entry?
- Find the error.* Here is a “proof” that $a^n = 1$ for every real $a \neq 0$ and every $n \in \mathbb{N}$; the powers a^n , $a^0 = 1$ and $a^{-1} = 1/a$ have their school meaning (Section 2.3 defines a^n for $n \in \mathbb{N}$ only, so negative exponents lie outside that definition). “Fix $a \neq 0$. Base: $a^0 = 1$. Step: let $n \in \mathbb{N}$ and suppose $a^k = 1$ for every $k \in \mathbb{N}$ with $k \leq n$. Then $a^{n+1} = a^n \cdot a^n / a^{n-1} = 1 \cdot 1/1 = 1$.” The conclusion is false (take $a = 2$, $n = 1$). Which transition fails, and why?

Solution. (a) (i) Ordinary induction from 0: the step factors out one copy of 7, as in the example $3 \mid 4^n - 1$ after [Definition 2.2.1](#). (ii) Induction from base 5 ([Theorem 2.2.1](#)): the claim fails at $n = 2, 3, 4$, so it cannot start at 0, and the step $2^{n+1} = 2 \cdot 2^n > 2n^2 \geq (n+1)^2$ works for $n \geq 5$. (iii) Strong induction from base 12 with three base cases $12 = 3 \cdot 4$, $13 = 3 \cdot 2 + 7$, $14 = 7 \cdot 2$: the step reaches back three, from n to $n - 3$. (iv) A smallest counterexample: in a solution (m, n) with least n , n^2 is even, so $n = 2n'$ is even (Logic and Proof, Theorem 2.1.3), and then $m^2 = 2n'^2$, so (n', m) is a solution whose second entry m is smaller than n , since $n^2 = 2m^2 > m^2$. (Strong induction would also do; the forms are interchangeable, and the answers above name the natural fit.)

(b) The negation is: there is an integer $n \geq 2$ with no prime factor. A smallest counterexample would be an integer $m \geq 2$ with no prime factor such that every integer k with $2 \leq k < m$ has one.

(c) $n = 2$. For $n = 1$ the lists are x_1 and x_2 ; for $n \geq 2$ both contain x_2 .

(d) The transition from 0 to 1 fails. At $n = 0$ the step writes $a^1 = a^0 \cdot a^0 / a^{-1}$ and uses $a^{-1} = 1$, which is “ $P(-1)$ ”; but the hypothesis covers only $k \in \mathbb{N}$ with $k \leq 0$, that is, $k = 0$, and $-1 \notin \mathbb{N}$. In fact $a^{-1} = 1/a \neq 1$ unless $a = 1$. The base case $a^0 = 1$ is true, and every transition from $n \geq 1$ on is valid, since then $n - 1 \in \mathbb{N}$ and $n - 1 \leq n$, and $a^n a^n / a^{n-1} = a^{n+1}$ for $n \geq 1$. As in [Example \(All Numbers Are Equal\)](#), the valid transitions carry nothing,

because $P(1)$ is already false.

2.2.6.2 B. Practice

Exercise. B1

Prove that $5 \mid 6^n - 1$ for every $n \in \mathbb{N}$, without adding up the powers of 6. Here 6^n has its school meaning: $6^0 = 1$, $6^{n+1} = 6^n \cdot 6$, and 6^n is an integer; Section 2.3 makes these official.

Solution. For $n \in \mathbb{N}$, let $P(n)$ be “there is $k \in \mathbb{Z}$ with $6^n - 1 = 5k$ ”. *Base case:* $6^0 - 1 = 0 = 5 \cdot 0$. *Step:* let $n \in \mathbb{N}$ and suppose $6^n - 1 = 5k$ with $k \in \mathbb{Z}$. Then

$$6^{n+1} - 1 = 6 \cdot 6^n - 1 = 6(6^n - 1) + 5 = 5(6k + 1),$$

and $6k + 1 \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a). So $P(n+1)$ holds. By Logic and Proof, Axiom 4.2.1, $5 \mid 6^n - 1$ for every $n \in \mathbb{N}$.

Exercise. B2: Products of primes

An integer is a **product of primes** if it equals $p_1 p_2 \cdots p_k$ for some $k \geq 1$ and primes $p_1, \dots, p_k$ (a single prime counts, with $k = 1$). Here $p_1 p_2 \cdots p_k$ has its school meaning, and multiplying two such products gives the product of the joined list; Section 2.3 defines finite products ([Definition 2.3.1](#)). Here is a proof that every integer $n \geq 2$ is a product of primes, with its last step blank. *Name the move, then complete it.*

Suppose not. Then the set F of integers $n \geq 2$ that are not products of primes is non-empty and has the integer lower bound 2, so it has a least element m by [Corollary 2.1.5](#). Since a prime is a product of primes, m is not prime; as $m \geq 2$, it has a positive divisor $a \notin \{1, m\}$, and $m = ab$ with $b \in \mathbb{Z}$. As in the proof of [Theorem 2.2.5](#), $b > 0$ and $2 \leq a < m$. Since $a < m = ab$ and $a > 0$, we get $b > 1$, so $b \geq 2$ by discreteness; and since $a \geq 2$, $b < 2b \leq ab = m$. So $2 \leq a < m$ and $2 \leq b < m$, and as m is the least element of F , neither a nor b lies in F : both are products of primes.

...

Solution. The move is the **smallest counterexample**, and the blank is its last step, *produce a contradiction*. Write $a = p_1 \cdots p_k$ and $b = p'_1 \cdots p'_l$ with primes p_i, p'_j . Then

$$m = ab = p_1 \cdots p_k p'_1 \cdots p'_l$$

is a product of primes, which contradicts $m \in F$. Hence F is empty, and every integer $n \geq 2$ is a product of primes.

Exercise. B3

- (a) Find the quotient and remainder of [Theorem 2.2.3](#) for $a = 17$, $n = 5$, and for $a = -17$, $n = 5$.

- (b) Prove that the square of an odd integer leaves remainder 1 on division by 8.
 (c) Prove that if $2 \mid n$ and $3 \mid n$, then $6 \mid n$, for every integer n .

Hint for (c): Theorem 2.2.3, dividing by 3.

Solution. (a) $17 = 5 \cdot 3 + 2$, so $q = 3$, $r = 2$. And $-17 = 5 \cdot (-4) + 3$, so $q = -4$, $r = 3$; in both cases $0 \leq r < 5$, so by uniqueness these are the quotient and remainder. (Not $-17 = 5 \cdot (-3) - 2$: a remainder is never negative.)

(b) Let $x = 2k + 1$ with $k \in \mathbb{Z}$. Then $x^2 = 4k^2 + 4k + 1 = 4k(k + 1) + 1$. By Theorem 2.2.3 with $n = 2$, $k = 2j$ or $k = 2j + 1$ with $j \in \mathbb{Z}$. In the first case $k(k + 1) = 2 \cdot j(2j + 1)$; in the second, $k(k + 1) = (2j + 1)(2j + 2) = 2(2j + 1)(j + 1)$. Either way $k(k + 1) = 2t$ with $t \in \mathbb{Z}$, so $x^2 = 8t + 1$. Since $0 \leq 1 < 8$, the uniqueness in Theorem 2.2.3 says that the remainder of x^2 on division by 8 is 1.

(c) Let $n = 2a$ with $a \in \mathbb{Z}$, and divide a by 3: $a = 3q + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < 3$, so $r \in \{0, 1, 2\}$ by discreteness. Then $n = 6q + 2r$. Since $3 \mid n$, say $n = 3s$ with $s \in \mathbb{Z}$, we get $2r = 3(s - 2q)$, so $3 \mid 2r$, where $2r \in \{0, 2, 4\}$. But $2 = 3 \cdot 0 + 2$ and $4 = 3 \cdot 1 + 1$ have non-zero remainders on division by 3, while a multiple $3t = 3t + 0$ has remainder 0; by uniqueness, neither 2 nor 4 is a multiple of 3. So $2r = 0$, $r = 0$, and $n = 6q$: $6 \mid n$.

2.2.6.3 C. Going deeper

Exercise. C1: Exit check: a flawed induction

Here is a claim and a “proof”.

Claim. Every integer $n \geq 18$ is $4a + 7b$ for some integers $a, b \geq 0$.

- (1) For an integer $n \geq 18$, let $P(n)$ be “ $n = 4a + 7b$ for some integers $a, b \geq 0$ ”. We use strong induction from base 18 (Theorem 2.2.2).
- (2) *Base case.* $18 = 4 \cdot 1 + 7 \cdot 2$, so $P(18)$ holds.
- (3) *Step.* Let $n \geq 18$ be an integer, and suppose $P(k)$ holds for every integer k with $18 \leq k \leq n$.
- (4) Write $n + 1 = (n - 3) + 4$.
- (5) Since $n - 3 < n$, the induction hypothesis gives $n - 3 = 4a + 7b$ with integers $a, b \geq 0$.
- (6) Then $n + 1 = 4(a + 1) + 7b$, with $a + 1 \geq 0$ and $b \geq 0$, so $P(n + 1)$ holds.
- (7) By Theorem 2.2.2, $P(n)$ holds for every integer $n \geq 18$.

Answer each question. (i) Is the claim true? (ii) Does the argument prove it? (iii) Which line fails, and for which n ? (iv) Which rule does that line break? (v) Give a counterexample to the failing step, not merely to some conclusion. (vi) Repair the proof.

Solution. (i) The claim is **true**. Directly,

$$18 = 4 \cdot 1 + 7 \cdot 2, \quad 19 = 4 \cdot 3 + 7, \quad 20 = 4 \cdot 5, \quad 21 = 7 \cdot 3,$$

and from these four the repaired induction in (vi) reaches every larger integer.

(ii) The argument does **not** prove it, although its conclusion is true.

(iii) Line (5) fails for $n = 18, 19$ and 20 . There $n - 3$ is $15, 16, 17$, all below 18 , and the induction hypothesis says nothing about them. For $n \geq 21$, $18 \leq n - 3 < n$, and line (5) is correct. Lines (1)–(4), (6) and (7) are correct. Line (7) needs one remark, because the step is written from n to $n + 1$ and not in the theorem’s own form. At $N = 18$ the theorem’s hypothesis has a vacuous “if” part, so it asks for $P(18)$ outright, which is line (2). At an integer $N \geq 19$, put $n = N - 1 \geq 18$. By the discreteness clause of [Proposition 2.1.2](#), the integers k with $18 \leq k < N$ are exactly those with $18 \leq k \leq n$, so lines (3)–(6), were they valid for every $n \geq 18$, would prove the hypothesis at N .

(iv) Line (5) applies $\forall E$ to the hypothesis “for every integer k with $18 \leq k \leq n$, $P(k)$ ” at $k = n - 3$. That is a bounded quantifier, and its elimination needs the line “ $18 \leq n - 3$ ” (Logic and Proof, Proposition 6.1.1); for $n \leq 20$ that line is false. The proof checked only half of the bound, $n - 3 < n$.

(v) Run the same argument with base $14 = 7 \cdot 2$: it “proves” that every integer $n \geq 14$ is $4a + 7b$. That conclusion is false at 17 : with $b = 0, 1, 2$ the remainder $17 - 7b$ is $17, 10, 3$, none a multiple of 4 , and $b \geq 3$ makes it negative. The step that breaks is line (5) at $n = 16$: it takes $P(13)$ from the hypothesis, but $13 < 14$ is outside it, and 13 is not of the form (13 and 6 are not multiples of 4). At $n = 16$ the induction hypothesis is true ($14 = 7 \cdot 2$, $15 = 4 \cdot 2 + 7$, $16 = 4 \cdot 4$), and the step concludes the false $P(17)$. So line (5) is an illegal step, not merely an unlucky one.

(vi) Take four base cases, $18, 19, 20, 21$, as in (i), and run the step only for $n \geq 21$, where $n - 3 \geq 18$. In the form of [Theorem 2.2.2](#): for $n \in \{18, 19, 20, 21\}$, $P(n)$ holds directly; for $n \geq 22$, $n - 4 \geq 18$ and $n - 4 < n$, so the hypothesis gives $n - 4 = 4a + 7b$ and $n = 4(a + 1) + 7b$.

Marking scheme. (10 points) - 2: judges the conclusion **true**, with the small cases $18, 19, 20, 21$ written as $4a + 7b$. - 3: identifies the failing line exactly: the appeal to the hypothesis for $n - 3$, for $n = 18, 19, 20$. - 2: states the reason as a hypothesis used outside its range ($\forall E$ at a k with $k < 18$). - 2: gives a counterexample **to the step**: the same step carrying a true statement to a false one (for example, base 14 and the failure at 17), not merely a false conclusion. - 1: gives a correct repair (four base cases, or any equivalent). Fatal errors (0 for the item): calling the conclusion false; blaming the base case.

Exercise. C2: Lowest terms

Prove that every rational number can be written as p/q with $p \in \mathbb{Z}$, $q \in \mathbb{Z}_{>0}$, and no common divisor $d \geq 2$ of p and q : there is no integer $d \geq 2$ with $d \mid p$ and $d \mid q$.

Solution (Hint 1). Follow the proof of [Lemma 2.2.4](#), step by step.

Solution (Hint 2). Suppose $r \in \mathbb{Q}$ has no such representation, and let m be the least element of $F := \{q \in \mathbb{Z}_{>0} : r = p/q \text{ for some } p \in \mathbb{Z}\}$.

Solution. Fix $r \in \mathbb{Q}$, and suppose that r has no representation of the required kind. Let F be as in Hint 2. By the definition of $\mathbb{Q}$, F is non-empty, and $F \subseteq \mathbb{Z}_{>0} \subseteq \mathbb{N}$, so by [Theorem 2.1.3](#) it has a least element m . Write $r = p/m$ with $p \in \mathbb{Z}$.

By the supposition, this representation fails: there is an integer $d \geq 2$ with $d \mid p$ and $d \mid m$, say $p = dp'$ and $m = dm'$ with $p', m' \in \mathbb{Z}$. Since $m > 0$ and $d > 0$, $m' = m/d > 0$, so $m' \in \mathbb{Z}_{>0}$ by [Proposition 2.1.2](#) (b). Then $r = dp'/(dm') = p'/m'$, so $m' \in F$. And $m = dm' \geq 2m' > m'$, since $d \geq 2$ and $m' > 0$. So m' is an element of F smaller than its least element, a contradiction. Hence r has a representation p/q with no common divisor $d \geq 2$.

The statement contains [Lemma 2.2.4](#), since $d = 2$ is one of the excluded common divisors. The only change in the proof is the size of the step down: $m' = m/d$ instead of $m/2$, and $d \geq 2$ is what makes it a step **down**.

Marking scheme. (10 points) - 2: sets up the smallest counterexample: fixes r , supposes no good representation, and defines the set of denominators. - 2: shows that set is a non-empty subset of $\mathbb{N}$ and takes its least element by [Theorem 2.1.3](#). - 3: extracts the common divisor $d \geq 2$ and writes the reduced representation p'/m' , checking $m' \in \mathbb{Z}_{>0}$. - 2: shows $m' < m$, using $d \geq 2$ and $m' > 0$. - 1: concludes, with the contradiction to the choice of m . A strong induction on the denominator earns the same points step for step. Fatal errors (0 for the step): taking a least element of a set not shown to lie in $\mathbb{N}$; dividing by d without $d \neq 0$.

Exercise. C3: Smallest, not largest

- (a) Prove that a set $A \subseteq \mathbb{R}$ has at most one least element.
- (b) Deduce that a claim “ $P(n)$ for every $n \in \mathbb{N}$ ” that is false has exactly one smallest counterexample.
- (c) Show that a false claim about $\mathbb{N}$ need not have a **largest** counterexample, by examining the claim “ $n < 5$ for every $n \in \mathbb{N}$ ”.

Solution. (a) Let m and m' be least elements of A . Since $m' \in A$ and m is least, $m \leq m'$; since $m \in A$ and m' is least, $m' \leq m$. Hence $m = m'$.

(b) The smallest counterexamples are the least elements of the set $F := \{n \in \mathbb{N} : P(n) \text{ is false}\}$. The claim is false, so F is non-empty, and [Theorem 2.1.3](#) gives a least element. By (a), there is only one.

(c) The claim is false, and its counterexamples form $F = \{n \in \mathbb{N} : n \geq 5\}$, which is non-empty. It has no greatest element: for every $n \in F$, also $n + 1 \in \mathbb{N}$ by (NS), and $n + 1 \geq 5$, so $n + 1 \in F$ and $n + 1 > n$. So the claim has a smallest counterexample, 5, and no largest one. Well-ordering has no upward twin in $\mathbb{N}$: a non-empty subset of $\mathbb{N}$ has a greatest element when it is bounded above ([Exercise \(C1: A greatest element from a least upper bound\)](#)), and not in general.

2.3 Definition by Recursion

Since school we have defined powers by two clauses: $u^0 = 1$, and $u^{n+1} = u^n u$. Sums $a_1 + \cdots + a_n$, factorials and many sequences are defined the same way: a starting value, and a rule that produces the next value from the last. *Logic and Proof*, Chapter 3 used induction to prove facts about such objects, such as $3^n \geq 2n + 1$ in *Logic and Proof*, Example (An inequality by induction). But induction proves facts about a function that **already exists**, and here the

function is what we want to obtain. This section proves that two such clauses always define exactly one function on $\mathbb{N}$, and then uses the theorem to define sums, products and powers properly and to prove their first laws.

2.3.1 Definitions that define something

Logic and Proof, Chapter 3 met the same question for formulas. A recursive definition on formulas gives a value for each atom and a rule for each way of building a compound formula, and *Logic and Proof*, Theorem 4.3.4 showed that such clauses define exactly one function, because unique readability leaves no choice of clause. On $\mathbb{N}$ the clauses are simpler: one starting value at 0, and one rule for passing from n to $n + 1$. What we want to prove is this.

A starting value and a rule for the next value determine exactly one sequence.

Theorem 2.3.1. Recursion Theorem

Let X be a set, let $a \in X$, and let $g: \mathbb{N} \times X \rightarrow X$ be a function. Then there is **exactly one** function $f: \mathbb{N} \rightarrow X$ such that

$$f(0) = a \quad \text{and} \quad f(n + 1) = g(n, f(n)) \quad \text{for every } n \in \mathbb{N}.$$

In words: a is the starting value, and g is the rule that turns the stage n and the current value $f(n)$ into the next value. The rule may look at n as well as at $f(n)$: the factorial needs this, since $(n + 1)! = (n + 1) \cdot n!$ is $g(n, x) = (n + 1)x$ at $x = n!$. The hypothesis that g maps into X is what lets the rule be applied again at the next stage. The conclusion has two halves: **existence** (some f obeys the two equations) and **uniqueness** (no two different functions do).

Idea. Uniqueness is an ordinary induction: two solutions agree at 0, and if they agree at n , the second equation makes them agree at $n + 1$. Existence is harder, because induction cannot produce f ; it can only prove statements. So we prove by induction a statement about finite pieces of f . The roadmap:

- ① **Uniqueness**, by induction on n .
- ② **Claim:** for every $n \in \mathbb{N}$ there is exactly one function on $\{0, \dots, n\}$ obeying the two equations wherever they make sense (a **partial solution**). Its graph is the graph of the previous one with one pair added.
- ③ **Glue:** collect the graphs of all partial solutions in one set $\mathcal{S}$ and take the union. The partial solutions agree where they overlap, so the union is the graph of a function on all of $\mathbb{N}$, and that function obeys the equations.

Before reading on: self-explain the Claim's inductive step.

Proof. For $n \in \mathbb{N}$, the set $\{0, \dots, n\} = \{k \in \mathbb{N} : k \leq n\}$ of Section 2.1 contains 0, since $0 \leq n$ by Lemma 2.1.1 (a). Call a function $h: \{0, \dots, n\} \rightarrow X$ a **partial solution up to n** if $h(0) = a$ and $h(k + 1) = g(k, h(k))$ for every $k \in \mathbb{N}$ with $k + 1 \leq n$. (For such k , both k and $k + 1$ lie in $\{0, \dots, n\}$, so the equation makes sense.)

Step 1 (uniqueness). Let $f, f': \mathbb{N} \rightarrow X$ both satisfy the two equations, and let $P(n)$ be

“ $f(n) = f'(n)$ ”. $P(0)$ holds, since $f(0) = a = f'(0)$. If $P(n)$ holds, then

$$f(n+1) = g(n, f(n)) = g(n, f'(n)) = f'(n+1),$$

so $P(n+1)$ holds. By Logic and Proof, Axiom 4.2.1, $f(n) = f'(n)$ for every $n \in \mathbb{N}$. The two functions have the same domain and the same codomain, so $f = f'$ by [Proposition 1.4.1](#).

Step 2.

Claim. For every $n \in \mathbb{N}$ there is exactly one partial solution up to n .

Proof of Claim. We use Logic and Proof, Axiom 4.2.1 on n .

Base case. By [Lemma 2.1.1](#) (a), every $k \in \mathbb{N}$ has $k \geq 0$, so $k \leq 0$ holds only for $k = 0$, and $\{0, \dots, 0\} = \{0\}$. The graph $\{(0, a)\}$ with codomain X is a function on $\{0\}$, since $a \in X$: the input 0 has the one output a . It is a partial solution up to 0: its value at 0 is a , and the second condition is vacuous, because $k+1 \geq 1 > 0$ for every $k \in \mathbb{N}$. Any partial solution up to 0 has value a at 0, so it equals this one by [Proposition 1.4.1](#).

Inductive step. Let $n \in \mathbb{N}$, and suppose there is exactly one partial solution up to n ; call it f_n , with graph G_n . Let

$$G_{n+1} := G_n \cup \{(n+1, g(n, f_n(n)))\}.$$

First, $\{0, \dots, n+1\} = \{0, \dots, n\} \cup \{n+1\}$: a natural number $k \leq n+1$ is either $n+1$ or less than $n+1$, and then $k \leq n$, since no natural number lies strictly between n and $n+1$ ([Lemma 2.1.1](#) (e)). Also $n+1 \notin \{0, \dots, n\}$, since $n+1 > n$.

We check the clauses of [Definition 1.4.1](#) for G_{n+1} on $\{0, \dots, n+1\}$ with codomain X . Every pair of G_{n+1} lies in $\{0, \dots, n+1\} \times X$, because g takes values in X . (Fn1): an input $k \leq n$ has the output $f_n(k)$ in G_n , and the input $n+1$ has the output $g(n, f_n(n))$. (Fn2): the pairs with first coordinate $k \leq n$ all lie in G_n , since the added pair has first coordinate $n+1 \neq k$, and G_n satisfies (Fn2); the only pair with first coordinate $n+1$ is the added one, since the first coordinates in G_n are at most n . So $f_{n+1} := (G_{n+1}, X)$ is a function on $\{0, \dots, n+1\}$, and it agrees with f_n on $\{0, \dots, n\}$.

It is a partial solution up to $n+1$. Its value at 0 is $f_n(0) = a$. Let $k \in \mathbb{N}$ with $k+1 \leq n+1$, so $k \leq n$. If $k < n$, then $k+1 \leq n$ by [Lemma 2.1.1](#) (e), and $f_{n+1}(k+1) = f_n(k+1) = g(k, f_n(k)) = g(k, f_{n+1}(k))$, because f_n is a partial solution. If $k = n$, then $f_{n+1}(n+1) = g(n, f_n(n)) = g(n, f_{n+1}(n))$ by the choice of the added pair.

It is the only one. Let h be a partial solution up to $n+1$. Its restriction $h|_{\{0, \dots, n\}}$ (Section 1.4) is a partial solution up to n , because the conditions on it are among the conditions on h . By the induction hypothesis it is f_n . So $h(k) = f_n(k) = f_{n+1}(k)$ for $k \leq n$, and

$$h(n+1) = g(n, h(n)) = g(n, f_n(n)) = f_{n+1}(n+1).$$

By [Proposition 1.4.1](#), $h = f_{n+1}$. This proves the claim. ■

For each $n \in \mathbb{N}$, let f_n be the partial solution up to n , and G_n its graph. If $m \leq n$, then $\{0, \dots, m\} \subseteq \{0, \dots, n\}$, and the restriction of f_n to $\{0, \dots, m\}$ is a partial solution up to m , so by the claim it is f_m . In terms of graphs,

$$G_m = G_n \cap (\{0, \dots, m\} \times X) \subseteq G_n \quad (m \leq n).$$

Step 3 (existence). Let

$$\mathcal{S} := \{G \in \mathcal{P}(\mathbb{N} \times X) : G \text{ is the graph of a partial solution up to some } n \in \mathbb{N}\},$$

a set by [Definition 1.1.3](#) and [Definition 1.2.2](#). By the claim, the elements of $\mathcal{S}$ are exactly the graphs G_n . Let $F := \bigcup \mathcal{S}$ ([Definition 1.2.4](#)), so a pair lies in F if and only if it lies in some G_n ; in particular $F \subseteq \mathbb{N} \times X$. We check the clauses of [Definition 1.4.1](#) for F .

(Fn1): for $n \in \mathbb{N}$, the pair $(n, f_n(n))$ lies in $G_n \in \mathcal{S}$, hence in F .

(Fn2): let $(k, y) \in F$ and $(k, y') \in F$, say $(k, y) \in G_m$ and $(k, y') \in G_n$. If $m \leq n$, then $G_m \subseteq G_n$, so both pairs lie in G_n , and $y = y'$ because G_n satisfies (Fn2). If $n < m$, the same argument with m and n exchanged gives $y = y'$.

So $f := (F, X)$ is a function $\mathbb{N} \rightarrow X$. For $m \geq n$, the pair $(n, f_m(n))$ lies in $G_m \subseteq F$, so $f(n) = f_m(n)$. Hence $f(0) = f_0(0) = a$, and for every $n \in \mathbb{N}$,

$$f(n+1) = f_{n+1}(n+1) = g(n, f_{n+1}(n)) = g(n, f(n)),$$

where the middle equation holds because f_{n+1} is a partial solution. This proves existence, and with Step 1, the theorem. ■

The starting value was used in both halves: Step 1 needs $f(0) = f'(0)$ as its base case, and Step 2 needs it to fix the first partial solution. The set $\mathcal{S}$ is cut out of a power set by a condition, so no family $n \mapsto f_n$ had to be formed before the union was taken. No choice is made anywhere: at every stage the claim leaves exactly one candidate. From now on, a definition of the form “ $f(0) := a$, $f(n+1) := g(n, f(n))$ ” is legitimate whenever $a \in X$ and g is a function $\mathbb{N} \times X \rightarrow X$, and we say that f is **defined by recursion**.

Remark. Because the rule may depend on n , the theorem also covers constructions in which the step uses outside data indexed by n . *Analysis on the Real Line*, Chapter 1 builds closed intervals I_{n+1} from I_n and the $(n+1)$ -st number of a given list in exactly this way, with no choice principle involved. The values may be sets as well as numbers. With $X = \mathcal{P}(\mathbb{N})$, $A_0 := \mathbb{N}$ and $A_{n+1} := A_n \setminus \{n\}$ define a sequence of sets of the kind Section 1.2 promised, and an induction that follows the same two clauses shows $A_n = \{k \in \mathbb{N} : k \geq n\}$: the base is [Lemma 2.1.1](#) (a), and the step removes n from $\{k : k \geq n\}$, leaving $\{k : k > n\} = \{k : k \geq n+1\}$ by [Lemma 2.1.1](#) (e).

2.3.2 Starting elsewhere, looking two steps back

Two variants come up constantly. Sequences are often indexed from 1, as in Section 1.4, and some recursions use the two previous values. Both reduce to the theorem.

Corollary 2.3.2. Recursion from a Base

Let n_0 be an integer, $D := \{n \in \mathbb{Z} : n \geq n_0\}$, X a set, $a \in X$, and $g: D \times X \rightarrow X$ a function. Then there is exactly one function $f: D \rightarrow X$ with $f(n_0) = a$ and $f(n+1) = g(n, f(n))$ for every $n \in D$.

Idea. Rename the index: $k = n - n_0$ turns D into $\mathbb{N}$ and the recursion on D into one on $\mathbb{N}$, where [Theorem 2.3.1](#) applies.

Proof. For $k \in \mathbb{N}$, the number $n_0 + k$ is an integer ([Proposition 2.1.2](#) (a)) and $n_0 + k \geq n_0$ ([Lemma 2.1.1](#) (a)), so it lies in D . Conversely, for $n \in D$, $n - n_0$ is an integer ([Proposition 2.1.2](#) (a)) and ≥ 0 , so $n - n_0 \in \mathbb{N}$ by [Proposition 2.1.2](#) (b). Define $g': \mathbb{N} \times X \rightarrow X$ by $g'(k, x) := g(n_0 + k, x)$. By [Theorem 2.3.1](#) there is exactly one $f': \mathbb{N} \rightarrow X$ with $f'(0) = a$ and $f'(k+1) = g(n_0 + k, f'(k))$ for every $k \in \mathbb{N}$.

Existence. Let $t: D \rightarrow \mathbb{N}$, $t(n) := n - n_0$, and $f := f' \circ t$ ([Definition 1.4.2](#)). Then $f(n_0) = f'(0) = a$, and for $n \in D$, with $k = n - n_0$,

$$f(n+1) = f'(k+1) = g(n_0 + k, f'(k)) = g(n, f(n)).$$

Uniqueness. If $\tilde{f}: D \rightarrow X$ also satisfies both equations, then $k \mapsto \tilde{f}(n_0 + k)$ satisfies the equations that define f' , so it is f' by the uniqueness in [Theorem 2.3.1](#). Hence $\tilde{f}(n) = f'(n - n_0) = f(n)$ for every $n \in D$, and $\tilde{f} = f$ by [Proposition 1.4.1](#). ■

The renaming is the move that gave induction from a base ([Theorem 2.2.1](#)). The corollary covers every sequence $(a_n)_{n \geq 1}$ given by a first term and a rule, and *Analysis on the Real Line*, Chapter 1's intervals, which are listed from 1.

Corollary 2.3.3. Two-Step Recursion

Let X be a set, $a, b \in X$, and $h: \mathbb{N} \times (X \times X) \rightarrow X$ a function ([Definition 1.3.2](#)); write $h(n, x, y)$ for $h(n, (x, y))$. Then there is exactly one function $f: \mathbb{N} \rightarrow X$ with

$$f(0) = a, \quad f(1) = b, \quad f(n+2) = h(n, f(n), f(n+1)) \quad \text{for every } n \in \mathbb{N}.$$

Idea. A rule that looks two steps back is a rule that looks one step back at **pairs**. Carry the pair $(f(n), f(n+1))$ along: the next pair is $(f(n+1), f(n+2))$, and both of its entries are computed from the current pair.

Proof. Define $G: \mathbb{N} \times (X \times X) \rightarrow X \times X$ by $G(n, (x, y)) := (y, h(n, x, y))$. By [Theorem 2.3.1](#) with the set $X \times X$ and the starting value (a, b) , there is exactly one $p: \mathbb{N} \rightarrow X \times X$ with $p(0) = (a, b)$ and $p(n+1) = G(n, p(n))$. Write $p(n) = (p_1(n), p_2(n))$. The equation $p(n+1) = (p_2(n), h(n, p_1(n), p_2(n)))$ gives, by the characteristic property of pairs ([Definition 1.3.1](#)),

$$p_1(n+1) = p_2(n), \quad p_2(n+1) = h(n, p_1(n), p_2(n)).$$

Let $f(n) := p_1(n)$. Then $f(0) = a$, $f(1) = p_2(0) = b$, and

$$f(n+2) = p_2(n+1) = h(n, p_1(n), p_2(n)) = h(n, f(n), f(n+1)).$$

For uniqueness, let f' satisfy the three equations, and let $p'(n) := (f'(n), f'(n+1))$. Then

$p'(0) = (a, b)$ and

$$p'(n+1) = (f'(n+1), h(n, f'(n), f'(n+1))) = G(n, p'(n)),$$

so $p' = p$ by the uniqueness in [Theorem 2.3.1](#), and $f'(n) = p_1(n) = f(n)$ for every n . By [Proposition 1.4.1](#), $f' = f$. ■

2.3.3 Sums, products and powers

The symbols $\sum$ and $\prod$ are recursive definitions, and so are powers. Writing them as instances of [Theorem 2.3.1](#) settles what they mean at the edges: the empty sum, the empty product and u^0 .

Definition 2.3.1. Finite Sums and Products

Let $(a_k)_{k \geq 1}$ be a sequence of real numbers. The **sums** $\sum_{k=1}^n a_k$ and **products** $\prod_{k=1}^n a_k$, for $n \in \mathbb{N}$, are defined by recursion:

$$\begin{aligned} \sum_{k=1}^0 a_k &:= 0, & \sum_{k=1}^{n+1} a_k &:= \sum_{k=1}^n a_k + a_{n+1}, \\ \prod_{k=1}^0 a_k &:= 1, & \prod_{k=1}^{n+1} a_k &:= \left(\prod_{k=1}^n a_k \right) a_{n+1}. \end{aligned}$$

The **factorial** of $n \in \mathbb{N}$ is $n! := \prod_{k=1}^n k$. For integers $m \leq n+1$ and reals a_k given for every integer $k \geq m$,

$$\sum_{k=m}^n a_k := \sum_{i=1}^{n-m+1} a_{m-1+i}, \quad \prod_{k=m}^n a_k := \prod_{i=1}^{n-m+1} a_{m-1+i}.$$

In words: a sum starts at 0 and adds one term at each stage; a product starts at 1 and multiplies by one term at each stage. The sum is [Theorem 2.3.1](#) with $X = \mathbb{R}$, $a = 0$ and $g(n, x) = x + a_{n+1}$, a function $\mathbb{N} \times \mathbb{R} \rightarrow \mathbb{R}$ because $n+1 \geq 1$ is an index of the sequence ([Lemma 2.1.1](#) (a), (NS)); the product is $a = 1$, $g(n, x) = x a_{n+1}$. The letter k is a dummy: $\sum_{k=1}^n a_k$ and $\sum_{j=1}^n a_j$ are the same number. In the general form, $n-m+1$ is an integer ≥ 0 , hence in $\mathbb{N}$ ([Proposition 2.1.2](#) (b)), and the terms a_{m-1+i} , for $i \geq 1$, run through $a_m, a_{m+1}, \dots$; when $m = 1$ it gives the first form back.

The edge cases are the point of the definition. The **empty sum** is 0 and the **empty product** is 1: $\sum_{k=m}^{m-1} a_k = 0$ and $0! = 1$. These are the only values that make the recursive clause true at the first stage for every sequence, since $\sum_{k=1}^1 a_k$ should be a_1 and $\prod_{k=1}^1 a_k$ should be a_1 . So $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$.

Three consequences are used constantly, and each is a short check from the definition.

- **Finite lists.** $\sum_{k=1}^n a_k$ depends only on $a_1, \dots, a_n$. Precisely: if two sequences agree at $1, \dots, N$, their sums up to n agree for every $n \leq N$. Induct on n with the property “if $n \leq N$, the two sums up to n are equal”. The base is $0 = 0$. In the step, $n+1 \leq N$ gives $n \leq N$, so the sums up to n agree, and the added terms a_{n+1} agree because $1 \leq n+1 \leq N$. So a finite list $a_1, \dots, a_N$ is summed by extending it in any way, say by $a_k := 0$ for $k > N$,

and the result does not depend on the extension. The same holds for products, and for the general bounds with $a_m, \dots, a_n$.

- **Shifting the index.** For every integer p , $\sum_{k=m}^n a_k = \sum_{k=m+p}^{n+p} a_{k-p}$, because both sides unfold to $\sum_{i=1}^{n-m+1} a_{m-1+i}$. *Analysis on the Real Line*, Chapter 1's estimate of $y^k - x^k$ uses the shift $p = 1$.
- **Splitting off the last term.** For $m \leq n+1$, $\sum_{k=m}^{n+1} a_k = \sum_{k=m}^n a_k + a_{n+1}$. Unfold the left side to $\sum_{i=1}^{n-m+2} a_{m-1+i}$ and apply the recursive clause; its last term is $a_{m-1+(n-m+2)} = a_{n+1}$.

Definition 2.3.2. Powers

Let $u \in \mathbb{R}$. The **powers** u^n , for $n \in \mathbb{N}$, are defined by recursion:

$$u^0 := 1, \quad u^{n+1} := u^n u.$$

This is [Theorem 2.3.1](#) with $X = \mathbb{R}$, $a = 1$ and $g(n, x) = xu$. So $u^1 = u^0 u = u$, and $u^2 = uu$. The definition gives $0^0 = 1$, in line with the empty product: $u^n = \prod_{k=1}^n u$, since both sides obey the same two clauses and so agree by the uniqueness in [Theorem 2.3.1](#). Only exponents in $\mathbb{N}$ are defined here; roots $a^{1/n}$ come in *Analysis on the Real Line*, Chapter 1.

A sequence that approaches $\sqrt{2}$. Let $X := (0, \infty)$ and $g(n, x) := \frac{1}{2}(x + 2/x)$. Before writing a recursion, check that g maps X into X : for $x > 0$, $2/x$ is defined and positive, so $x + 2/x > 0$, and half of it is positive. By [Theorem 2.3.1](#), $a_0 := 2$ and $a_{n+1} := \frac{1}{2}(a_n + 2/a_n)$ define a sequence in $(0, \infty)$. Because the check was made on g , no separate induction is needed to show $a_n > 0$ before dividing by it. The first values are

$$a_0 = 2, \quad a_1 = \frac{1}{2}(2 + 1) = \frac{3}{2}, \quad a_2 = \frac{1}{2}\left(\frac{3}{2} + \frac{4}{3}\right) = \frac{17}{12},$$

and $(17/12)^2 = 289/144$, which exceeds 2 by $1/144$. That the sequence converges to $\sqrt{2}$ is a question for *Analysis on the Real Line*, Chapter 2, where recursively defined sequences become objects of study.

A two-step example. By [Corollary 2.3.3](#), $a_0 = 2$, $a_1 = 3$ and $a_{n+2} = 3a_{n+1} - 2a_n$ define a unique sequence of reals ($X = \mathbb{R}$, $h(n, x, y) = 3y - 2x$), and its first values are 2, 3, 5, 9, 17. These are $2^n + 1$. To prove it, induct on the **pair** statement $Q(n)$: “ $a_n = 2^n + 1$ and $a_{n+1} = 2^{n+1} + 1$ ”. $Q(0)$ is $2 = 1 + 1$ and $3 = 2 + 1$. If $Q(n)$ holds, its second half is the first half of $Q(n+1)$, and

$$a_{n+2} = 3(2^{n+1} + 1) - 2(2^n + 1) = 2 \cdot 2^{n+1} + 1 = 2^{n+2} + 1,$$

where $2 \cdot 2^n = 2^n \cdot 2 = 2^{n+1}$ and $2 \cdot 2^{n+1} = 2^{n+1} \cdot 2 = 2^{n+2}$ by [Definition 2.3.2](#) (and commutativity). So Logic and Proof, Axiom 4.2.1 gives $Q(n)$ for every n , and in particular $a_n = 2^n + 1$. The induction carries a pair because the recursion does: the proof follows the clauses of the definition.

Warning. A recursion needs its starting values, and its rule must stay inside X . “ $a_{n+2} = a_{n+1} + a_n$ ” alone defines nothing: $0, 1, 1, 2, \dots$ and $2, 1, 3, 4, \dots$ both obey it, and [Corollary 2.3.3](#) asks for a_0 and a_1 . And “ $a_0 = 2$, $a_{n+1} = 1/(a_n - 1)$ ” is not a definition by recursion: $a_1 = 1$, and then a_2 would be $1/0$. The rule $x \mapsto 1/(x - 1)$ is not a function from any set X containing 2 into X , since X would have to contain its value 1, where the rule is undefined.

Quick check. Compute $\prod_{k=1}^4 (1 + \frac{1}{k})$, and say why the answer is so simple.

Solution. Each factor is $1 + \frac{1}{k} = \frac{k+1}{k}$, so the product is

$$\frac{2}{1} \cdot \frac{3}{2} \cdot \frac{4}{3} \cdot \frac{5}{4} = 5.$$

Each numerator cancels the next denominator: the product **telescopes**. In general $\prod_{k=1}^n \frac{k+1}{k} = n+1$, by induction on n : the empty product is $1 = 0+1$, and $(n+1) \cdot \frac{n+2}{n+1} = n+2$.

2.3.4 Recursion, then induction

A function defined by recursion is studied by induction, and the induction follows the clauses of the definition: the base case uses the starting value, the step uses the rule. The laws of powers are the first example.

Proposition 2.3.4. Laws of Powers

Let $u, v \in \mathbb{R}$ and $m, n \in \mathbb{N}$.

- (a) $u^{m+n} = u^m u^n$.
- (b) $(u^m)^n = u^{mn}$.
- (c) $(uv)^n = u^n v^n$.
- (d) If $u \in \mathbb{Z}$, then $u^n \in \mathbb{Z}$; if $u \in \mathbb{N}$, then $u^n \in \mathbb{N}$.
- (e) If $u > 0$, then $u^n > 0$. If $u > 1$ and $m < n$, then $u^m < u^n$.

Idea. Each part is an induction on n from 0, with m fixed. The base case is $u^0 = 1$ in some form, and the step peels one factor off with $u^{n+1} = u^n u$ and applies the induction hypothesis to u^n . For the second half of (e), write $n = m + j$ with $j \geq 1$ and use (a).

Proof.

- (a) Fix m and induct on n . For $n = 0$: $u^{m+0} = u^m = u^m \cdot 1 = u^m u^0$. Suppose $u^{m+n} = u^m u^n$. Since $m+n \in \mathbb{N}$ ([Lemma 2.1.1](#) (b)), [Definition 2.3.2](#) applies at $m+n$, and

$$u^{m+(n+1)} = u^{(m+n)+1} = u^{m+n} u = u^m u^n u = u^m u^{n+1},$$

where the third equation is the induction hypothesis and the last is [Definition 2.3.2](#). By Logic and Proof, Axiom 4.2.1, (a) holds for every n .

- (b) Fix m . For $n = 0$: $(u^m)^0 = 1 = u^0 = u^{m \cdot 0}$. If $(u^m)^n = u^{mn}$, then $(u^m)^{n+1} = (u^m)^n u^m = u^{mn} u^m = u^{mn+m} = u^{m(n+1)}$, using (a) with the exponents $mn, m \in \mathbb{N}$ ([Lemma 2.1.1](#) (b)).
- (c) $(uv)^0 = 1 = 1 \cdot 1 = u^0 v^0$. If $(uv)^n = u^n v^n$, then $(uv)^{n+1} = (uv)^n uv = u^n v^n uv = (u^n u)(v^n v) = u^{n+1} v^{n+1}$.

- (d) $u^0 = 1 = 0 + 1$ lies in $\mathbb{N}$ by $((\mathbb{N})0)$ and (NS) , and $\mathbb{N} \subseteq \mathbb{Z}$. If $u^n \in \mathbb{Z}$ and $u \in \mathbb{Z}$, then $u^{n+1} = u^n u \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a). For $u \in \mathbb{N}$ the same induction works with [Lemma 2.1.1](#) (b) in place of [Proposition 2.1.2](#).
- (e) Let $u > 0$. Then $u^0 = 1 > 0$, and if $u^n > 0$, then $u^{n+1} = u^n u > 0$ as a product of positive numbers. Now let $u > 1$. First, $u^j > 1$ for every integer $j \geq 1$ (such j lies in $\mathbb{N}$ by [Proposition 2.1.2](#) (b), so u^j is defined), by [Theorem 2.2.1](#) with base 1: $u^1 = u > 1$, and if $u^j > 1$, then $u^{j+1} = u^j u > 1 \cdot u > 1$, multiplying $u^j > 1$ by $u > 0$. Let $m < n$. Then $j := n - m \in \mathbb{N}$ by [Lemma 2.1.1](#) (d), and $j \neq 0$, so $j \geq 1$ by [Lemma 2.1.1](#) (a). By (a), $u^n = u^{m+j} = u^m u^j > u^m$, multiplying $u^j > 1$ by $u^m > 0$.

■

Some instances that earlier chapters used on trust: $(t^2)^2 = t^4$ is (b) with $m = n = 2$; $2^{s+t} = 2^s 2^t$ is (a); $n \mapsto 2^n$ is strictly increasing on $\mathbb{N}$ by (e); and an integer power of an integer is an integer by (d). For $(-1)^n$: by (b), $(-1)^{2k} = ((-1)^2)^k = 1^k = 1$, where $1^k = 1$ by the induction $1^{k+1} = 1^k \cdot 1$; and $(-1)^{2k+1} = (-1)^{2k}(-1) = -1$. Every $n \in \mathbb{N}$ is $2k+r$ with $k \in \mathbb{Z}$ and $r \in \{0, 1\}$ ([Theorem 2.2.3](#) with divisor 2, and [Proposition 2.1.2](#) (c) for $r \leq 1$), and $k \geq 0$: if $k < 0$, then $k \leq -1$ ([Proposition 2.1.2](#) (c)), so $n = 2k+r \leq -2+1 < 0$, contradicting $n \geq 0$ ([Lemma 2.1.1](#) (a)). So $k \in \mathbb{N}$ by [Proposition 2.1.2](#) (b), n is $2k$ or $2k+1$, and $(-1)^n$ is 1 for even n and -1 for odd n .

The next two formulas are proved the same way; each is a recursion (the sum) followed by an induction whose step splits off the last term.

Proposition 2.3.5. Sum of the First Natural Numbers

For every $n \in \mathbb{N}$, $\sum_{k=1}^n k = \frac{n(n+1)}{2}$.

Proof. For $n = 0$, both sides are 0: the left by the empty sum, the right since $0 \cdot 1 = 0$. If the formula holds for n , then by [Definition 2.3.1](#)

$$\sum_{k=1}^{n+1} k = \frac{n(n+1)}{2} + (n+1) = \frac{(n+1)(n+2)}{2},$$

which is the formula for $n+1$. By Logic and Proof, Axiom 4.2.1 it holds for every $n \in \mathbb{N}$.

■

Proposition 2.3.6. Geometric Sum

Let $r \in \mathbb{R}$ with $r \neq 1$. For every $n \in \mathbb{N}$,

$$\sum_{k=1}^n r^{k-1} = \frac{1-r^n}{1-r}.$$

Proof. The terms are defined: for $k \geq 1$, $k-1 \in \mathbb{N}$ by [Lemma 2.1.1](#) (c), so r^{k-1} is given by [Definition 2.3.2](#), and the term added at stage $n+1$ is $r^{(n+1)-1} = r^n$. The right side is

defined because $1 - r \neq 0$. For $n = 0$, both sides are 0, since $r^0 = 1$. If the formula holds for n , then, by [Definition 2.3.1](#) and the induction hypothesis,

$$\begin{aligned}\sum_{k=1}^{n+1} r^{k-1} &= \frac{1 - r^n}{1 - r} + r^n = \frac{1 - r^n + r^n - r^{n+1}}{1 - r} \\ &= \frac{1 - r^{n+1}}{1 - r},\end{aligned}$$

where $r^n(1 - r) = r^n - r^{n+1}$ by [Definition 2.3.2](#). By Logic and Proof, Axiom 4.2.1 the formula holds for every $n \in \mathbb{N}$. ■

For $r = 1$ the sum is n , since every term is 1; that is the case the hypothesis excludes, where the right side would divide by 0.

The next inequality is the most used estimate of a power in analysis. Its proof shows a hypothesis doing exactly one job.

Theorem 2.3.7. Bernoulli's Inequality

Let $x \in \mathbb{R}$ with $x \geq -1$. Then $(1 + x)^n \geq 1 + nx$ for every $n \in \mathbb{N}$.

Idea. Induct on n . To pass from n to $n + 1$, multiply the induction hypothesis by $1 + x$; the product $(1 + nx)(1 + x) = 1 + (n + 1)x + nx^2$ is at least $1 + (n + 1)x$, because the extra term nx^2 is non-negative. Multiplying an inequality by $1 + x$ keeps its direction only if $1 + x \geq 0$, and that is what $x \geq -1$ says.

Proof. For $n = 0$: $(1 + x)^0 = 1 \geq 1 = 1 + 0 \cdot x$. Suppose $(1 + x)^n \geq 1 + nx$. Since $x \geq -1$, we have $1 + x \geq 0$, and multiplying the induction hypothesis by $1 + x$ gives

$$\begin{aligned}(1 + x)^{n+1} &= (1 + x)^n(1 + x) \geq (1 + nx)(1 + x) \\ &= 1 + (n + 1)x + nx^2 \geq 1 + (n + 1)x,\end{aligned}$$

where the first equation is [Definition 2.3.2](#) and the last step uses $nx^2 \geq 0$, true because $n \geq 0$ ([Lemma 2.1.1](#) (a)) and $x^2 \geq 0$. By Logic and Proof, Axiom 4.2.1 the inequality holds for every $n \in \mathbb{N}$. ■

The hypothesis $x \geq -1$ appears in the proof once, in the sentence “since $x \geq -1$, we have $1 + x \geq 0$ ”, and it cannot simply be dropped: for $x = -4$ and $n = 3$, $(1 + x)^3 = (-3)^3 = -27$, while $1 + 3x = -11$, and $-27 < -11$. Exercise C2 below shows that the statement can survive at some $x < -1$ even though the proof does not.

The last result of this subsection is not about a recursive definition, but its induction runs over lists, and it is the fact *Analysis on the Real Line*, Chapter 1 needs about finite sets. A **maximum** of a set $A \subseteq \mathbb{R}$ is an element $M \in A$ with $a \leq M$ for every $a \in A$; a **minimum** is

an element $m \in A$ with $m \leq a$ for every $a \in A$. For reals $x_1, \dots, x_n$, we write $\{x_1, \dots, x_n\} := \{x_k : k \in \{1, \dots, n\}\}$.

Proposition 2.3.8. Finite Sets Have a Maximum

For every integer $n \geq 1$ and all reals $x_1, \dots, x_n$, the set $\{x_1, \dots, x_n\}$ has a maximum and a minimum.

Idea. Add the elements one at a time. The maximum of the longer list is the old maximum or the new element, whichever is larger, and comparing the two is all the step needs.

Proof. Let $P(n)$ be “for all reals $x_1, \dots, x_n$, the set $\{x_1, \dots, x_n\}$ has a maximum”. We prove $P(n)$ for every $n \geq 1$ by Theorem 2.2.1 with base 1.

Base case. $\{x_1\}$ has the maximum x_1 .

Inductive step. Let $n \geq 1$, suppose $P(n)$, and let $x_1, \dots, x_{n+1}$ be reals. Since $\{1, \dots, n+1\} = \{1, \dots, n\} \cup \{n+1\}$ (Lemma 2.1.1 (e), as in the proof of Theorem 2.3.1), the set $B := \{x_1, \dots, x_{n+1}\}$ is $A \cup \{x_{n+1}\}$, where $A := \{x_1, \dots, x_n\}$. By $P(n)$, applied to the list $x_1, \dots, x_n$, the set A has a maximum M . By trichotomy, $x_{n+1} \leq M$ or $M < x_{n+1}$.

Case 1: $x_{n+1} \leq M$. Then $M \in A \subseteq B$, and every element of B is an element $a \in A$, with $a \leq M$, or is $x_{n+1} \leq M$. So M is a maximum of B .

Case 2: $M < x_{n+1}$. Then $x_{n+1} \in B$, and every $a \in A$ satisfies $a \leq M < x_{n+1}$. So x_{n+1} is a maximum of B .

In both cases $P(n+1)$ holds. By Theorem 2.2.1, $P(n)$ holds for every $n \geq 1$.

For the minimum, apply the result to $-x_1, \dots, -x_n$: if $-x_j$ is a maximum of $\{-x_1, \dots, -x_n\}$, then $-x_k \leq -x_j$, that is $x_j \leq x_k$, for every k , so x_j is a minimum of $\{x_1, \dots, x_n\}$. ■

The base is 1, not 0: for $n = 0$ the set is $\emptyset$, which has neither a maximum nor a minimum, since both must be elements. The proof used only order properties of $\leq$: any two elements are comparable (trichotomy), $\leq$ is reflexive and transitive, and $x \mapsto -x$ reverses $\leq$. All of these hold in any ordered field, so the proof works word for word there; that is the setting in which *Analysis on the Real Line*, Chapter 1 uses it. Passing from a finite **set** to a list $x_1, \dots, x_n$ that enumerates it is what “finite” will mean in Chapter 3.

2.3.5 When the claim is too weak to carry its step

Sometimes the statement we want is true, but it cannot be proved by induction as it stands: its inductive step does not go through. The cure looks paradoxical. We prove **more**.

Example. Proving more to prove anything

Prove that $\sum_{k=1}^n \frac{1}{k^2} < 2$ for every integer $n \geq 1$.

Solution. The direct attempt fails. Let $S_n := \sum_{k=1}^n 1/k^2$, so $S_{n+1} = S_n + 1/(n+1)^2$ by Definition 2.3.1. The base $S_1 = 1 < 2$ holds. For the step, all we may assume is $S_n < 2$, and that is too little: it allows $S_n = 2 - \frac{1}{2(n+1)^2}$, and then $S_n + \frac{1}{(n+1)^2} > 2$. The values of S_n never come so close to 2, but the induction hypothesis cannot know that.

Prove more. We prove $P(n)$: “ $S_n \leq 2 - \frac{1}{n}$ ” for every $n \geq 1$, by Theorem 2.2.1 with base 1. It implies the claim, since $2 - \frac{1}{n} < 2$.

Base case. $S_1 = 1 = 2 - \frac{1}{1}$.

Inductive step. Let $n \geq 1$ and suppose $S_n \leq 2 - \frac{1}{n}$. Then

$$S_{n+1} = S_n + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n} + \frac{1}{(n+1)^2}.$$

It remains to show $-\frac{1}{n} + \frac{1}{(n+1)^2} \leq -\frac{1}{n+1}$, that is, $\frac{1}{(n+1)^2} \leq \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}$. Both denominators are positive, and $n(n+1) \leq (n+1)^2$ because $n \leq n+1$ and $n+1 > 0$; so the inequality holds. Hence $S_{n+1} \leq 2 - \frac{1}{n+1}$, which is $P(n+1)$.

By Theorem 2.2.1, $S_n \leq 2 - \frac{1}{n} < 2$ for every $n \geq 1$.

The stronger statement carries its own step because its hypothesis says **how far** below 2 the sum is, and that margin pays for the next term. When a step fails, ask what the hypothesis would need to say, and prove that instead. Exercise C3 below uses it twice over, proving two identities together because neither carries its own one-step induction.

2.3.6 Exercises

2.3.6.1 A. Check your understanding

Exercise. A1

These questions are about the proof of Theorem 2.3.1.

- Why does uniqueness need only an induction, while existence needs the claim about partial solutions?
- In Step 3, which property of the graphs G_n makes $\bigcup \mathcal{S}$ satisfy (Fn2), and which part of the proof supplies it?
- Run the construction for $X = \mathbb{R}$, $a = 1$ and $g(n, x) = (n+1)x$: write the graphs G_0 , G_1 , G_2 . Which familiar function is f ?
- State the idea of the proof in two sentences.
- Write $\sum_{k=1}^n a_k$ and $n!$ as instances of Theorem 2.3.1: name X , a and g for each.

Solution.

- Uniqueness is a statement about two functions that are **given**: it compares their values, and comparing values at $n+1$ through the values at n is an induction. Existence asks for a function that is not given. Induction proves statements and does not build objects, so the proof proves a statement (the claim) whose content is that the finite pieces exist and are unique, and then builds f from them as a union.

- (b) The graphs agree where they overlap: $G_m \subseteq G_n$ for $m \leq n$, more precisely $G_m = G_n \cap (\{0, \dots, m\} \times X)$. This comes from the **uniqueness** half of the claim: the restriction of f_n to $\{0, \dots, m\}$ is a partial solution up to m , so it is f_m . With it, two pairs $(k, y) \in G_m$ and $(k, y') \in G_n$ lie in the larger graph, whose (Fn2) gives $y = y'$. So it verifies (Fn2) of [Definition 1.4.1](#) for $\bigcup \mathcal{S}$.
- (c) $G_0 = \{(0, 1)\}$. Then $g(0, 1) = 1 \cdot 1 = 1$, so $G_1 = \{(0, 1), (1, 1)\}$. Then $g(1, 1) = 2 \cdot 1 = 2$, so $G_2 = \{(0, 1), (1, 1), (2, 2)\}$. The function f is $n \mapsto n!$: by [Definition 2.3.1](#), $0! = 1$ and $(n+1)! = n!(n+1) = g(n, n!)$, so $n \mapsto n!$ satisfies both equations, and by the uniqueness in [Theorem 2.3.1](#) it is f .
- (d) For each n , induction shows that there is exactly one function on $\{0, \dots, n\}$ obeying the equations, and these pieces extend one another. Their union is the required function, and an induction on n shows there is no other.
- (e) The sum: $X = \mathbb{R}$, $a = 0$, $g(n, x) = x + a_{n+1}$. The factorial: $X = \mathbb{R}$ (or $\mathbb{N}$), $a = 1$, $g(n, x) = (n+1)x$, as in (c).

2.3.6.2 B. Practice

Exercise. B1: Sum of cubes

Prove that $\sum_{k=1}^n k^3 = \left(\frac{n(n+1)}{2}\right)^2$ for every $n \in \mathbb{N}$.

Solution. For $n = 0$, both sides are 0 (the empty sum, and $0^2 = 0$). Suppose the formula holds for n . By [Definition 2.3.1](#) and the induction hypothesis,

$$\begin{aligned} \sum_{k=1}^{n+1} k^3 &= \frac{n^2(n+1)^2}{4} + (n+1)^3 = (n+1)^2 \frac{n^2 + 4n + 4}{4} \\ &= \frac{(n+1)^2(n+2)^2}{4} = \left(\frac{(n+1)(n+2)}{2}\right)^2, \end{aligned}$$

which is the formula for $n+1$. By Logic and Proof, Axiom 4.2.1 the formula holds for every $n \in \mathbb{N}$. (Compare [Proposition 2.3.5](#): the sum of the first n cubes is the square of the sum of the first n numbers.)

Exercise. B2: Starting at the right place

Logic and Proof, Chapter 3 observed, in the warning after Logic and Proof, Example (An inequality by induction), that “ $2^n \geq n^2$ ” fails at $n = 3$ and holds again from $n = 4$ on. Prove that $2^n \geq n^2$ for every integer $n \geq 4$, using [Theorem 2.2.1](#). Say why the base must be 4 and not 3.

Solution. Let $P(n)$ be “ $2^n \geq n^2$ ”; we use [Theorem 2.2.1](#) with $n_0 = 4$.

Base case. $2^4 = 16 \geq 16 = 4^2$.

Inductive step. Let $n \geq 4$ and suppose $2^n \geq n^2$. Then $2^{n+1} = 2 \cdot 2^n \geq 2n^2$ by [Definition 2.3.2](#) and the hypothesis. Moreover $2n^2 \geq (n+1)^2$, because $2n^2 - (n+1)^2 = n^2 - 2n - 1 = (n-1)^2 - 2$,

and $(n-1)^2 \geq 9 > 2$ for $n \geq 4$. Hence $2^{n+1} \geq (n+1)^2$.

By [Theorem 2.2.1](#), $2^n \geq n^2$ for every $n \geq 4$.

The base cannot be 3, because $P(3)$ is false: $2^3 = 8 < 9 = 3^2$. The step is valid from $n = 3$ on, since $(n-1)^2 - 2 \geq 2$ for $n \geq 3$, but it is useless at 3, where its hypothesis is false. At $n = 2$ the step fails $((2-1)^2 - 2 = -1)$, which is why the true $P(2)$ does not lead to $P(3)$.

Exercise. B3: A closed form

The sequence $a_0 := 0$, $a_{n+1} := 2a_n + 1$ is defined by [Theorem 2.3.1](#). Compute its first five values, guess a formula for a_n , and prove it.

Solution. The values are 0, 1, 3, 7, 15: one less than 1, 2, 4, 8, 16. We prove $a_n = 2^n - 1$ for every $n \in \mathbb{N}$. For $n = 0$: $a_0 = 0 = 2^0 - 1$. If $a_n = 2^n - 1$, then

$$a_{n+1} = 2(2^n - 1) + 1 = 2^{n+1} - 1,$$

using $2 \cdot 2^n = 2^{n+1}$ ([Definition 2.3.2](#)). By Logic and Proof, Axiom 4.2.1, $a_n = 2^n - 1$ for every $n \in \mathbb{N}$. The induction follows the two clauses of the definition: the base case is the starting value, the step is the rule.

2.3.6.3 C. Going deeper

Exercise. C1: Powers of two

Prove that every $n \in \mathbb{Z}_{>0}$ can be written as $2^k q$ with $k \in \mathbb{N}$ and q odd (that is, $q = 2j + 1$ for an integer j). Below, the proof is started; the second half is blank. *Name the move, then complete it.*

Suppose not. Then the set F of $n \in \mathbb{Z}_{>0}$ that cannot be written in this form is non-empty, and $F \subseteq \mathbb{N}$; let m be its least element. Then $m \geq 1$, and m is not odd, since otherwise $m = 2^0 m$ would be of the required form. _____

Why is the statement made for $n \geq 1$ only?

Solution. The move is the **smallest counterexample** of Section 2.2: suppose the claim fails, take the least failure by well-ordering ([Theorem 2.1.3](#)), and produce a smaller failure.

Completion. By [Theorem 2.2.3](#) with divisor 2, every integer is even or odd; m is not odd, so $m = 2m'$ for an integer m' . Since $m > 0$, $m' = m/2 > 0$, so $m' \in \mathbb{Z}_{>0}$ ([Proposition 2.1.2](#) (b)), and $m' < 2m' = m$. As m is the least element of F , $m' \notin F$, so $m' = 2^k q$ with $k \in \mathbb{N}$ and q odd. Then

$$m = 2 \cdot 2^k q = 2^{k+1} q$$

by [Definition 2.3.2](#), and $k+1 \in \mathbb{N}$ by (NS). So m has the required form, contradicting $m \in F$. Hence $F = \emptyset$, and every $n \in \mathbb{Z}_{>0}$ is $2^k q$ with q odd.

The restriction to $n \geq 1$ is needed: $0 = 2^k q$ forces $q = 0$, since $2^k > 0$ ([Proposition 2.3.4](#) (e)), and $0 = 2 \cdot 0$ is even, not odd ([Theorem 2.2.3](#): not both). So 0 is a genuine counterexample,

and in the proof it is excluded by taking $m \in \mathbb{Z}_{>0}$, which is what makes $m' > 0$.

Marking scheme. (10 points) - 2: names the move (smallest counterexample) and says that the least element m exists by well-ordering ([Theorem 2.1.3](#)). - 2: from “ m is not odd” concludes $m = 2m'$ with $m' \in \mathbb{Z}$, citing even-or-odd. - 3: shows $m' \in \mathbb{Z}_{>0}$ and $m' < m$, so $m' \notin F$. - 2: writes $m = 2^{k+1}q$ with $k+1 \in \mathbb{N}$ and reaches the contradiction. - 1: explains why $n = 0$ is excluded. Fatal errors (0 for the step): concluding $m' \notin F$ (or applying the induction hypothesis to m') without having shown both $m' < m$ and $m' \geq 1$; missing either one is fatal. A reader who restarts by [Theorem 2.2.2](#) with base 1 earns the points for the other steps (the induction hypothesis at m' replaces “ $m' \notin F$ ”), but not the 2 points for naming the move.

Exercise. C2: Bernoulli beyond its hypothesis

- (a) Show that $(1+x)^n \geq 1+nx$ holds for $x = -2$ and every $n \in \mathbb{N}$.
- (b) Say which line of the proof of [Theorem 2.3.7](#) fails at $x = -2$, and why the statement survives anyway.
- (c) At $x = -4$, find the first n where the inequality fails, and the first n where the step “ $P(n) \Rightarrow P(n+1)$ ” fails.

Solution.

- (a) At $x = -2$ the claim is $(-1)^n \geq 1 - 2n$. By the remark after [Proposition 2.3.4](#), $(-1)^n$ is 1 or -1 . For $n = 0$ the claim reads $1 \geq 1$. For $n \geq 1$, $1 - 2n \leq -1 \leq (-1)^n$.
- (b) The line “multiplying the induction hypothesis by $1+x$ ” fails. At $x = -2$ the multiplier is $1+x = -1 < 0$, and multiplying $(1+x)^n \geq 1+nx$ by a negative number **reverses** it: we get $(1+x)^{n+1} \leq (1+nx)(1+x)$, the wrong direction. The rest of the proof ($nx^2 \geq 0$) still holds. The statement survives because (a) proves it by a different argument, which uses the special values $(-1)^n = \pm 1$. A proof that fails does not make its conclusion false.
- (c) At $x = -4$: $n = 0$ gives $1 \geq 1$; $n = 1$ gives $-3 \geq -3$; $n = 2$ gives $9 \geq -7$; $n = 3$ gives $-27 \geq -11$, false. So the first failure is $n = 3$. The step $P(n) \Rightarrow P(n+1)$ holds for $n = 0$ and $n = 1$ (both sides are true) and fails at $n = 2$, where $P(2)$ is true and $P(3)$ false.

Exercise. C3: Two identities at once

Use [Corollary 2.3.3](#) to define the sequence $F_0 := 0$, $F_1 := 1$, $F_{n+2} := F_{n+1} + F_n$; name X , a , b and h . Then prove, for every $n \in \mathbb{N}$, the statement $P(n)$:

$$F_{2n+1} = F_n^2 + F_{n+1}^2 \quad \text{and} \quad F_{2n+2} = F_{n+1}(F_n + F_{n+2}).$$

Check $P(0)$, $P(1)$ and $P(2)$ numerically first.

Solution (Hint 1). Neither identity carries its own one-step inductive step: passing from

F_{2n+1} to F_{2n+3} needs F_{2n+2} , which is the other identity. Prove both together, and use $F_n + F_{n+1} = F_{n+2}$ to merge terms.

Solution (Hint 2). In the step, start with $F_{2n+3} = F_{2n+2} + F_{2n+1}$ and substitute both halves of $P(n)$.

Solution. The definition. [Corollary 2.3.3](#) with $X = \mathbb{R}$, $a = 0$, $b = 1$ and $h(n, x, y) = x + y$ gives exactly one sequence with $F_0 = 0$, $F_1 = 1$ and $F_{n+2} = F_n + F_{n+1}$. Its first values are 0, 1, 1, 2, 3, 5, 8.

Small cases. $P(0)$: $F_1 = 1 = 0^2 + 1^2$ and $F_2 = 1 = 1 \cdot (0 + 1)$. $P(1)$: $F_3 = 2 = 1^2 + 1^2$ and $F_4 = 3 = 1 \cdot (1 + 2)$. $P(2)$: $F_5 = 5 = 1^2 + 2^2$ and $F_6 = 8 = 2 \cdot (1 + 3)$.

Proof. We use Logic and Proof, Axiom 4.2.1. $P(0)$ was checked. Suppose $P(n)$. Using the recursion and then both halves of $P(n)$,

$$\begin{aligned} F_{2n+3} &= F_{2n+2} + F_{2n+1} \\ &= F_{n+1}F_n + F_{n+1}F_{n+2} + F_n^2 + F_{n+1}^2 \\ &= F_n(F_n + F_{n+1}) + F_{n+1}F_{n+2} + F_n^2 \\ &= F_{n+2}(F_n + F_{n+1}) + F_{n+1}^2 = F_{n+2}^2 + F_{n+1}^2, \end{aligned}$$

where the last two lines use $F_n + F_{n+1} = F_{n+2}$ twice. This is the first half of $P(n + 1)$. For the second half,

$$\begin{aligned} F_{2n+4} &= F_{2n+3} + F_{2n+2} \\ &= F_{n+1}^2 + F_{n+2}^2 + F_{n+1}F_n + F_{n+1}F_{n+2} \\ &= F_{n+1}(F_{n+1} + F_n) + F_{n+2}^2 + F_{n+1}F_{n+2} \\ &= F_{n+2}(2F_{n+1} + F_{n+2}) = F_{n+2}(F_{n+1} + F_{n+3}), \end{aligned}$$

using the first half of $P(n + 1)$ just proved, the second half of $P(n)$, $F_{n+1} + F_n = F_{n+2}$, and finally $F_{n+1} + F_{n+2} = F_{n+3}$. So $P(n + 1)$ holds, and by Logic and Proof, Axiom 4.2.1 $P(n)$ holds for every $n \in \mathbb{N}$.

The second identity is written with F_{n+2} rather than in the form $F_{2n} = F_n(F_{n-1} + F_{n+1})$, so that both identities make sense on all of $\mathbb{N}$ and no F_{-1} is needed. The two identities hold together, and together they carry a one-step induction that neither carries alone: this is “prove more” ([Example \(Proving more to prove anything\)](#)) in the form of a pair, as in the two-step example after [Definition 2.3.2](#). (A two-step induction can prove the first identity alone; see the marking scheme.)

Marking scheme. (10 points) - 2: defines the sequence by [Corollary 2.3.3](#), naming X , a , b and h . - 1: checks the base case $P(0)$, both halves. - 3: proves the first identity at $n + 1$ (or for all n): $F_{2n+3} = F_{n+1}^2 + F_{n+2}^2$. - 3: proves the second identity at $n + 1$ (or for all n): $F_{2n+4} = F_{n+2}(F_{n+1} + F_{n+3})$. - 1: concludes by induction (one-step or two-step), with a hypothesis that covers every value the step uses. Alternative route: 3 for proving $F_{2n+1} = F_n^2 + F_{n+1}^2$ alone by a two-step induction (via $F_{k+2} = 3F_k - F_{k-2}$ for $k \geq 2$ and the matching recursion for $F_n^2 + F_{n+1}^2$, with both bases $n = 0, 1$ checked), and 3 for deriving the second identity as $F_{2n+2} = F_{2n+3} - F_{2n+1} = F_{n+2}^2 - F_n^2 = F_{n+1}(F_n + F_{n+2})$. Fatal errors (0

for the step): inducting on one identity alone while using the other at n unproved; using the identities at $n + 1$ to prove themselves.

2.4 Properties of Relations

Chapter 1 made relations into sets: a relation on a set X is a subset $R \subseteq X \times X$, and $x R y$ means $(x, y) \in R$ (Definition 1.3.3). It named the properties that make a relation an order or an equivalence and left them to this chapter. Equality, $\leq$, inclusion $\subseteq$, divisibility $|$, and “leaves the same remainder on division by 3” are all relations, and they behave alike in some ways and differently in others. This section names four properties that sort them, proves the ones that divisibility has, and collects witnesses for every failure. Sections 2.5 and 2.6 then study the two combinations that matter most.

2.4.1 Four properties

Start from equality. Every x equals itself; if $x = y$, then $y = x$; and if $x = y$ and $y = z$, then $x = z$. The relation $\leq$ on $\mathbb{R}$ keeps the first and third of these and loses the second, but it has a substitute: $x \leq y$ and $y \leq x$ together force $x = y$. Each of these is a property a relation may or may not have.

Each property is one rule for what the relation must contain, given what it already contains.

Definition 2.4.1. Properties of a Relation

Let X be a set and R a relation on X . Then R is

- **reflexive on X** if $x R x$ for every $x \in X$:

$$\forall x \in X \quad x R x;$$

- **symmetric** if $x R y$ implies $y R x$, for all $x, y \in X$:

$$\forall x \in X \quad \forall y \in X \quad (x R y \rightarrow y R x);$$

- **antisymmetric** if $x R y$ and $y R x$ together imply $x = y$, for all $x, y \in X$:

$$\forall x \in X \quad \forall y \in X \quad ((x R y \wedge y R x) \rightarrow x = y);$$

- **transitive** if $x R y$ and $y R z$ together imply $x R z$, for all $x, y, z \in X$:

$$\forall x \in X \quad \forall y \in X \quad \forall z \in X \quad ((x R y \wedge y R z) \rightarrow x R z).$$

In words: reflexive says every element is related to itself. Symmetric says the relation can be read backwards: whenever it relates x to y , it also relates y to x . Antisymmetric says it can be read both ways only between an element and itself. Transitive says two steps can always be replaced by one.

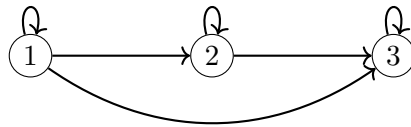
Three of the four are implications, so a relation has them unless some related pair (or triple) makes the conclusion false: for symmetry and transitivity a pair is missing from R , and for antisymmetry two different elements are related both ways. Negating with Logic and Proof,

Proposition 6.2.3 and Logic and Proof, Proposition 6.2.2 turns each failure into a **witness**:

- not reflexive: $\exists x \in X \neg(x R x)$,
- not symmetric: $\exists x \in X \exists y \in X (x R y \wedge \neg(y R x))$,
- not antisymmetric: $\exists x \in X \exists y \in X ((x R y \wedge y R x) \wedge x \neq y)$,
- not transitive: $\exists x \in X \exists y \in X \exists z \in X ((x R y \wedge y R z) \wedge \neg(x R z))$.

So “not symmetric” is shown by **one** pair related in one direction only, and “not transitive” by one triple. A single witness settles a failure; a success needs an argument for all elements.

Reading a relation as a picture. Draw the elements of X as points and an arrow from x to y whenever $x R y$. The relation $\leq$ on $\{1, 2, 3\}$ is $\{(1, 1), (2, 2), (3, 3), (1, 2), (1, 3), (2, 3)\}$:



Reflexive means a loop at every point. Symmetric means every arrow has a return arrow. Antisymmetric means no two **different** points are joined in both directions. Transitive means every path of two arrows has a direct arrow from its start to its end. In the picture there are loops everywhere, no return arrows, and the two-arrow path $1 \rightarrow 2 \rightarrow 3$ has its shortcut $1 \rightarrow 3$; so $\leq$ on $\{1, 2, 3\}$ is reflexive, antisymmetric and transitive, and not symmetric, with witness the pair $(1, 2)$. A path may also return to its start: $x \rightarrow y \rightarrow x$. Transitivity then demands the loop at x . That case is the source of the warning below.

Why “reflexive on X ”. The other three properties mention elements only when they are related, so they depend on the set R alone. Reflexivity depends on X too. The relation $\{(0, 0)\}$ is reflexive on $\{0\}$, and the same set of pairs, viewed as a relation on $\{0, 1\}$, is not reflexive on $\{0, 1\}$, with witness 1. That is why the definition names the set.

2.4.2 Divisibility

Divisibility shows that the same rule can give an antisymmetric relation on one set and not on a larger one, because the larger set brings in more pairs: $|$ on $\mathbb{Z}$ contains both $(2, -2)$ and $(-2, 2)$, which $|$ on $\mathbb{N}$ does not. Recall [Definition 2.2.1](#): for integers a, b , $a \mid b$ means $b = ka$ for some $k \in \mathbb{Z}$.

Proposition 2.4.1. Divisibility on the Natural Numbers

The relation $|$ is reflexive, antisymmetric and transitive on $\mathbb{N}$. On $\mathbb{Z}$ it is reflexive and transitive, and **not** antisymmetric.

Idea. Reflexivity and transitivity are one line each from [Definition 2.2.1](#). For antisymmetry, $b = ka$ and $a = lb$ give $a = lka$. If $a \neq 0$, this forces $kl = 1$, and for **positive** integers that means $k = l = 1$. On $\mathbb{N}$ the multipliers are positive, so $b = a$; on $\mathbb{Z}$ they may both be -1 .

Proof. Reflexive. For every integer a , $a = 1 \cdot a$, so $a \mid a$.

Transitive. Let $a \mid b$ and $b \mid c$, say $b = ka$ and $c = lb$ with $k, l \in \mathbb{Z}$. Then $c = (lk)a$, and $lk \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a), so $a \mid c$.

These two arguments work for all integers, so on $\mathbb{N}$ and on $\mathbb{Z}$.

Antisymmetric on $\mathbb{N}$. Let $a, b \in \mathbb{N}$ with $a \mid b$ and $b \mid a$, say $b = ka$ and $a = lb$ with $k, l \in \mathbb{Z}$.

Case 1: $a = 0$ or $b = 0$. If $a = 0$, then $b = k \cdot 0 = 0$; if $b = 0$, then $a = l \cdot 0 = 0$. Either way $a = b$.

Case 2: $a \neq 0$ and $b \neq 0$. By [Lemma 2.1.1](#) (a), $a \geq 1$ and $b \geq 1$, so both are positive. Since $ka = b > 0$ and $a > 0$, we have $k > 0$ (if $k \leq 0$, then $ka \leq 0$); likewise $l > 0$. By [Proposition 2.1.2](#) (b), $k, l \in \mathbb{Z}_{>0} = \{n \in \mathbb{N} : n \geq 1\}$, so $k \geq 1$ and $l \geq 1$. Now $a = lb = lka$, and $a \neq 0$, so $kl = 1$. If $k > 1$, then multiplying by $l > 0$ gives $kl > l \geq 1$, contradicting $kl = 1$. So $k \leq 1$, hence $k = 1$, and $b = a$.

Not antisymmetric on $\mathbb{Z}$. $-2 = (-1) \cdot 2$ and $2 = (-1) \cdot (-2)$, so $2 \mid -2$ and $-2 \mid 2$, but $2 \neq -2$. ■

The failure on $\mathbb{Z}$ is a non-example by minimal change: enlarging the set from $\mathbb{N}$ to $\mathbb{Z}$ keeps reflexivity and transitivity, and antisymmetry fails exactly because an integer and its negative divide each other. Two facts about $(\mathbb{N}, \mid)$ will matter in Section 2.6: every $a \in \mathbb{N}$ divides 0, since $0 = 0 \cdot a$, and 1 divides every a , since $a = a \cdot 1$. In the divisibility relation, 0 sits above everything and 1 below everything.

2.4.3 A table of examples

Here are eight relations checked against the four properties. Each “no” comes with its witness, which is how a failure is proved; each “yes” is a one-line argument given after the table.

Example. Eight relations, four properties

Decide which of the four properties each relation has: $=$, $\leq$ and $<$ on $\mathbb{R}$; $\subseteq$ on $\mathcal{P}(X)$ for a set X with at least one element; $\mid$ on $\mathbb{N}$ and on $\mathbb{Z}$; “ $x - y \in \mathbb{Z}$ ” on $\mathbb{R}$; and the empty relation $\emptyset$ on a set X with at least one element.

Solution.

Relation	reflexive	symmetric	antisymmetric	transitive
$=$ on $\mathbb{R}$	yes	yes	yes	yes
$\leq$ on $\mathbb{R}$	yes	no: 0, 1	yes	yes
$<$ on $\mathbb{R}$	no: 0	no: 0, 1	yes	yes
$\subseteq$ on $\mathcal{P}(X)$	yes	no: $\emptyset, X$	yes	yes
$\mid$ on $\mathbb{N}$	yes	no: 1, 2	yes	yes
$\mid$ on $\mathbb{Z}$	yes	no: 1, 2	no: 2, -2	yes
$x - y \in \mathbb{Z}$ on $\mathbb{R}$	yes	yes	no: 0, 1	yes
$\emptyset$ on X	no: any x	yes	yes	yes

The witnesses. $0 \leq 1$ and not $1 \leq 0$; $0 < 1$ and not $1 < 0$; not $0 < 0$. $\emptyset \subseteq X$, since the empty set is a subset of every set (Section 1.1, after Proposition 1.1.3), and $X \not\subseteq \emptyset$, since X has an element and $\emptyset$ has none. $1 \mid 2$, and $2 \nmid 1$: $1 = 2k$ would give $0 < k < 1$, and no integer lies there (Proposition 2.1.2 (c)). The pair $2, -2$ is from Proposition 2.4.1. For $x - y \in \mathbb{Z}$: $0 - 1 = -1$ and $1 - 0 = 1$ are integers, and $0 \neq 1$. For $\emptyset$: no x has $(x, x) \in \emptyset$.

The “yes” entries. Equality is reflexive by =I (Logic and Proof, Definition 5.7.1), symmetric and transitive by Logic and Proof, Proposition 5.7.1, and antisymmetric because its conclusion $x = y$ is one of its hypotheses. For $\leq$ and $<$, antisymmetry and transitivity are order facts about $\mathbb{R}$, and $\leq$ is reflexive because $x = x$ gives $x \leq x$; for $<$, antisymmetry holds **vacuously**, since $x < y$ and $y < x$ never hold together. For $\subseteq$, reflexivity is immediate from Definition 1.1.2, transitivity is Proposition 1.1.2, and antisymmetry is Proposition 1.1.1. For $\mid$, see Proposition 2.4.1. For $x - y \in \mathbb{Z}$: $x - x = 0 \in \mathbb{Z}$; $y - x = -(x - y)$; and $x - z = (x - y) + (y - z)$; these are integers by Proposition 2.1.2 (a). The empty relation is symmetric, antisymmetric and transitive vacuously: each property is an implication whose hypothesis requires a pair in $\emptyset$.

Three degenerate cases are worth a sentence each, because each shows a property holding for a reason other than the one we expect.

- The **empty relation** on a non-empty X has every property except reflexivity, all vacuously. It is the standard witness that the three implication properties do not imply reflexivity.
- The **full relation** $X \times X$ relates everything to everything. It is reflexive, symmetric and transitive; it is antisymmetric only if X has at most one element.
- If $X = \emptyset$, then $\mathcal{P}(X) = \{\emptyset\}$, and $\subseteq$ on it is $\{(\emptyset, \emptyset)\}$: with only one element, it is **symmetric**. The non-emptiness of X in the table is what the witness $\emptyset, X$ needs.

Quick check. On the set $\{2, 3, 4, \dots\}$ of integers ≥ 2 , let $x C y$ mean “ x and y have a common prime factor”, that is, some prime p divides both. Which of the four properties does C have?

Solution. Reflexive: every integer $x \geq 2$ has a prime factor p (Theorem 2.2.5), and p divides x and x . Symmetric: the condition “ $p \mid x$ and $p \mid y$ ” does not depend on the order of x and y . Not antisymmetric: $2 C 4$ and $4 C 2$ (both divisible by 2), and $2 \neq 4$. Not transitive: $2 C 6$ (common factor 2) and $6 C 3$ (common factor 3), but 2 and 3 have no common prime factor, since a prime dividing 2 is a positive divisor of 2 that is at least 2, hence is 2 (the example after Definition 2.2.2), and $2 \nmid 3$: $3 = 2k$ would give $1 < k < 2$, and no integer lies there (Proposition 2.1.2 (c)). Sharing a factor with a middle element does not pass along, because the shared factor may change.

Warning. A false theorem: “symmetric and transitive imply reflexive.” The “proof”: let $x \in X$, and take y with $x R y$; by symmetry $y R x$, and by transitivity $x R x$. The conclusion is false: the empty relation on $\{1\}$ is symmetric and transitive (vacuously) and not reflexive, with witness 1. The illegal step, “take y with $x R y$ ”, is an $\exists E$ (Logic and Proof, Definition 5.6.1) on $\exists y \in X x R y$, which was never derived and need not be true: (N9) of Logic and Proof, Example (What is not allowed) (Exercise C2 gives the correct version).

Remark. “Antisymmetric” is not “not symmetric”. Equality is both symmetric and antisymmetric. And $\{(1, 2), (2, 1), (2, 3)\}$ on $\{1, 2, 3\}$ is neither: it is not symmetric, since $2 R 3$ and not $3 R 2$; and it is not antisymmetric, since $1 R 2$, $2 R 1$ and $1 \neq 2$.

Two combinations of the properties organize the rest of the chapter. A relation that is reflexive, symmetric and transitive behaves like equality; Section 2.5 calls it an **equivalence relation** and shows that it cuts X into classes. A relation that is reflexive, antisymmetric and transitive behaves like $\leq$; Section 2.6 calls it a **partial order**; inclusion on $\mathcal{P}(X)$ and divisibility on $\mathbb{N}$ are examples that are not comparisons of sizes.

2.4.4 Exercises

2.4.4.1 A. Check your understanding

Exercise. A1

- (a) State in symbols what it means for a relation R on X to be antisymmetric, and write its negation in symbols.
- (b) True or false, with a witness: every antisymmetric relation fails to be symmetric.
- (c) Which of the four properties does the relation “ $|x - y| \leq 1$ ” have on $\mathbb{R}$?
- (d) Let $X = \{p, q\}$ with $p \neq q$. How many relations on X are reflexive on X ?

Solution.

- (a) $\forall x \in X \forall y \in X ((x R y \wedge y R x) \rightarrow x = y)$. Its negation, by Logic and Proof, Proposition 6.2.3 and Logic and Proof, Proposition 6.2.2: $\exists x \in X \exists y \in X ((x R y \wedge y R x) \wedge x \neq y)$.
- (b) False. Equality on any set is antisymmetric and symmetric. So is the empty relation (vacuously).
- (c) Reflexive: $|x - x| = 0 \leq 1$. Symmetric: $|y - x| = |x - y|$. Not antisymmetric: $|0 - 1| \leq 1$ both ways and $0 \neq 1$. Not transitive: $|0 - 1| \leq 1$ and $|1 - 2| \leq 1$, but $|0 - 2| = 2 > 1$. “Close to” does not pass along a chain, because small differences add up.
- (d) Four. A reflexive relation must contain (p, p) and (q, q) . The other two pairs of $X \times X$, (p, q) and (q, p) , may each be in or out, independently, which gives $2 \cdot 2 = 4$ relations: $\{(p, p), (q, q)\}$ together with one of $\emptyset$, $\{(p, q)\}$, $\{(q, p)\}$, $\{(p, q), (q, p)\}$.

2.4.4.2 B. Practice

Exercise. B1: Five relations

For each relation, determine which of the four properties it has. Justify your answer, with a witness for each failure.

- (a) On $\mathbb{N} \times \mathbb{N}$: $(a, b) E (c, d)$ if and only if $a + d = b + c$.
- (b) On $\mathbb{R}$: $x S y$ if and only if $x^2 = y^2$.
- (c) On the set $\mathbb{R}^{\mathbb{R}}$ of functions $\mathbb{R} \rightarrow \mathbb{R}$: $f L g$ if and only if $f(x) \leq g(x)$ for every $x \in \mathbb{R}$.

- (d) On $\mathbb{R}$: $x P y$ if and only if $xy > 0$.
 (e) On $\mathbb{Z}$: $x N y$ if and only if $y = x + 1$.

Solution.

- (a) Reflexive: $a + b = b + a$. Symmetric: $a + d = b + c$ gives $c + b = d + a$. Transitive: from $a + d = b + c$ and $c + f = d + e$, adding gives $a + d + c + f = b + c + d + e$, and canceling $c + d$ gives $a + f = b + e$. Not antisymmetric: $(0, 0) E (1, 1)$ both ways, and $(0, 0) \neq (1, 1)$. (The pairs (a, b) and (c, d) are related exactly when $a - b = c - d$; *Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* , Chapter 2 builds $\mathbb{Z}$ from the classes of this relation.)
- (b) Reflexive, symmetric and transitive, because “ $x^2 = y^2$ ” is equality of the values x^2 and y^2 , and equality has those three properties. Not antisymmetric: $1^2 = (-1)^2$ and $1 \neq -1$.
- (c) Reflexive: $f(x) \leq f(x)$. Transitive: $f(x) \leq g(x) \leq h(x)$ at each x . Antisymmetric: if $f L g$ and $g L f$, then $f(x) = g(x)$ for every x , and $f = g$ by [Proposition 1.4.1](#) (same domain and codomain). Not symmetric: the constant functions 0 and 1 satisfy $0 L 1$ and not $1 L 0$. In addition, two functions need not be comparable: for $f(x) = x$ and the constant 0, neither $f L 0$ (fails at $x = 1$) nor $0 L f$ (fails at $x = -1$).
- (d) Symmetric: $xy = yx$. Transitive: if $xy > 0$ and $yz > 0$, then $(xy)(yz) = xzy^2 > 0$, and $y^2 > 0$ (since $y \neq 0$, as $xy \neq 0$), so $xz > 0$. **Not reflexive:** $0 \cdot 0 = 0$, so 0 is not related to itself. Not antisymmetric: $1 P 2$ and $2 P 1$ (both products are $2 > 0$), and $1 \neq 2$. This is the planted case: P is symmetric and transitive, yet not reflexive on $\mathbb{R}$, exactly at the one element related to nothing. On $\mathbb{R} \setminus \{0\}$ it is reflexive. Compare the warning before these exercises.
- (e) Not reflexive: $0 \neq 0 + 1$. Not symmetric: $0 N 1$ and not $1 N 0$ ($0 \neq 2$). Not transitive: $0 N 1$ and $1 N 2$, and not $0 N 2$. Antisymmetric, vacuously: $x N y$ and $y N x$ would give $y = x + 1$ and $x = y + 1 = x + 2$, which is impossible, so no pair satisfies the hypothesis.

Exercise. B2: Example generation

Give an example of a relation on $\{1, 2, 3\}$ that is reflexive and transitive, but neither symmetric nor antisymmetric. Check each property. Then list the properties that every correct answer must have.

Solution. One answer:

$$R = \{(1, 1), (2, 2), (3, 3), (1, 2), (2, 1), (1, 3), (2, 3)\}.$$

Reflexive: the three loops are in R . Not symmetric: $1 R 3$ and not $3 R 1$. Not antisymmetric: $1 R 2$, $2 R 1$ and $1 \neq 2$. Transitive: we check every two-step path $x R y R z$ with $x \neq y$ and $y \neq z$ (a step along a loop changes nothing). They are $1 \rightarrow 2 \rightarrow 1$, $1 \rightarrow 2 \rightarrow 3$, $2 \rightarrow 1 \rightarrow 2$ and $2 \rightarrow 1 \rightarrow 3$, and their shortcuts $(1, 1)$, $(1, 3)$, $(2, 2)$, $(2, 3)$ are all in R . Element 3 has no arrow out except its loop, so no other path exists.

Every correct answer must: - contain the three pairs $(1, 1)$, $(2, 2)$, $(3, 3)$; - contain two different elements related **both** ways (to fail antisymmetry); - contain a pair related **one** way only (to fail symmetry); - contain the shortcut of every two-arrow path.

The second and fourth items together force more: if $a R b$, $b R a$ and $b R c$, transitivity puts $a R c$ in. First, there is only one two-way pair: two of them, say $\{a, b\}$ and $\{b, c\}$, give $a R c$ and $c R a$ by transitivity, so R is the full relation, which is symmetric. Call the two-way pair $\{a, b\}$ and the third element c . Every arrow between a and b is two-way, so a one-way arrow must have c at one end, and one such arrow is never enough: transitivity relates c to the other element of the pair too, in the same direction. For example, $\{(1, 1), (2, 2), (3, 3), (1, 2), (2, 1), (3, 1)\}$ fails transitivity, since $3 R 1 R 2$ and not $3 R 2$. If $c R a$, then $c R a R b$ gives $c R b$, and neither $a R c$ nor $b R c$ can hold, since that would make a second two-way pair; so both arrows point out of c . Likewise $a R c$ forces $b R c$, and both arrows point into c . Conversely, each of these six relations is a correct answer. Those with c above the pair are the answer above with the elements renamed. Those with c below are obtained from them by reversing every arrow, which keeps the loops, the two-way pair and the one-way arrows, and keeps transitivity: a path $z \rightarrow y \rightarrow x$ of the reversed relation is the reverse of a path $x \rightarrow y \rightarrow z$, and the reverse of its shortcut (x, z) is (z, x) . So $\{1, 2, 3\}$ carries exactly six correct answers: one for each choice of the two-way pair (three choices) and of whether the third element lies above the pair or below it (two choices); the answer above has 3 above the pair $\{1, 2\}$.

Exercise. B3: Intersections and unions

Let R and S be transitive relations on a set X . Prove that $R \cap S$ is transitive. Show by an example on $\{1, 2, 3\}$ that $R \cup S$ need not be transitive.

Solution. Let $x, y, z \in X$ with $(x, y) \in R \cap S$ and $(y, z) \in R \cap S$. Then $(x, y), (y, z) \in R$, so $(x, z) \in R$ because R is transitive; likewise $(x, z) \in S$. Hence $(x, z) \in R \cap S$.

For the union, let $R = \{(1, 2)\}$ and $S = \{(2, 3)\}$. Each is transitive, vacuously: in R a two-step path would need a pair starting at 2, and there is none; in S one would need a pair ending at 2. But $R \cup S = \{(1, 2), (2, 3)\}$ contains $(1, 2)$ and $(2, 3)$ and not $(1, 3)$, so it is not transitive. The intersection keeps both relations' shortcuts; the union creates paths that neither relation contains, so neither supplies their shortcuts.

2.4.4.3 C. Going deeper

Exercise. C1: Properties as equations between sets

Let R be a relation on X , let R^{-1} be its inverse relation (Definition 1.3.3), and let $\Delta_X := \{(x, x) : x \in X\}$ be the diagonal.

- Prove that R is symmetric if and only if $R = R^{-1}$, and that R is antisymmetric if and only if $R \cap R^{-1} \subseteq \Delta_X$.
- Deduce that a relation that is both symmetric and antisymmetric satisfies $R \subseteq \Delta_X$. Is every subset of Δ_X both symmetric and antisymmetric?

Solution. Recall that $(x, y) \in R^{-1}$ if and only if $(y, x) \in R$.

- Symmetric.* ($\Rightarrow$) Let R be symmetric. If $(x, y) \in R$, then $(y, x) \in R$, that is, $(x, y) \in R^{-1}$; so $R \subseteq R^{-1}$. If $(x, y) \in R^{-1}$, then $(y, x) \in R$, so $(x, y) \in R$ by symmetry;

so $R^{-1} \subseteq R$. By [Proposition 1.1.1](#), $R = R^{-1}$. ($\Leftarrow$) If $R = R^{-1}$ and $x R y$, then $(x, y) \in R^{-1}$, so $(y, x) \in R$.

Antisymmetric. A pair (x, y) lies in $R \cap R^{-1}$ exactly when $x R y$ and $y R x$, and it lies in Δ_X exactly when $x = y$ (by the characteristic property of pairs, [Definition 1.3.1](#)). So “ $R \cap R^{-1} \subseteq \Delta_X$ ” says: for all $x, y \in X$, if $x R y$ and $y R x$, then $x = y$. That is antisymmetry, word for word.

- (b) If R is symmetric and antisymmetric, then by (a), $R = R \cap R = R \cap R^{-1} \subseteq \Delta_X$. Conversely, every $R \subseteq \Delta_X$ is symmetric (each of its pairs is (x, x) , which is its own reverse) and antisymmetric ($x R y$ already forces $x = y$). So the relations with both properties are exactly the subsets of the diagonal; equality, Δ_X itself, is the largest of them, and $\emptyset$ the smallest.

Exercise. C2: What symmetry and transitivity do give

Let R be a symmetric and transitive relation on X , and let $D := \{x \in X : \exists y \in X \ x R y\}$ be its domain. Prove that $R \subseteq D \times D$ and that R , as a relation on D , is reflexive on D , symmetric and transitive.

Solution. $R \subseteq D \times D$. Let $(x, y) \in R$. Then $x \in D$, with witness y . By symmetry $y R x$, so $y \in D$, with witness x . Hence $(x, y) \in D \times D$.

Reflexive on D . Let $x \in D$. By the definition of D there is $y \in X$ with $x R y$; fix one (an $\exists E$, legitimate now, because $x \in D$ is exactly the statement $\exists y \in X \ x R y$). By symmetry $y R x$, and by transitivity, from $x R y$ and $y R x$, $x R x$.

Symmetric and transitive on D . These are implications about related elements, and every related pair lies in $D \times D$; so they hold on D because they hold on X .

This is the correct version of the false theorem in the warning: its argument proves reflexivity exactly on the set of elements for which “take y with $x R y$ ” is available. Elements outside D are related to nothing, and there reflexivity fails. In the language of Section 2.5, R is an equivalence relation on D .

Marking scheme. (10 points) - 2: shows $R \subseteq D \times D$, using symmetry to put the second coordinate in D . - 4: for reflexivity on D , obtains y with $x R y$ **from** $x \in D$, and says why this step is now legitimate. - 2: applies symmetry and then transitivity, in that order, to get $x R x$. - 2: explains why symmetry and transitivity restrict to D . Fatal errors (0 for the step): taking y with $x R y$ for an arbitrary $x \in X$ rather than $x \in D$ (the error of the warning); applying transitivity to $x R y$ and $x R y$.

2.5 Equivalence Relations and Partitions

Section 2.4 sorted relations by four properties: reflexive, symmetric, antisymmetric and transitive ([Definition 2.4.1](#)). This section studies the relations that have the first, second and fourth. They are the relations that say two things are “the same” in some respect, and each of them cuts its set into disjoint pieces of mutually “same” elements. We prove that the pieces always form a partition, and that every partition arises from exactly one such relation. Then we build the integers modulo n , and meet the check that every definition on pieces needs: that it does

not depend on which element names the piece. Section 1.4 promised that check for functions defined on classes; here it is.

2.5.1 Arithmetic on a clock

On a twelve-hour clock, five hours after nine o'clock it is two o'clock. People say “ $9 + 5 = 2$ on a clock”, and nobody is confused. Read as a statement about integers, it is false: $9 + 5 = 14$. So what is it a statement about?

The clock does not distinguish 2 from 14, or from 26 or -10 : all of them put the hand in the same place. Two integers name the same hour when they differ by a multiple of 12. The sentence “ $9 + 5 = 2$ ” is then about hours, and an hour is not one integer but a whole collection of them, $\{\dots, -10, 2, 14, 26, \dots\}$. The addition adds collections: take any integer from the first, any from the second, add them, and see which collection the sum lands in.

Two things have to be checked before this is mathematics. First, “differ by a multiple of 12” must behave like “is the same as”: it must cut the integers into separate collections, and no integer may land in two of them. Second, the answer must not depend on the choice: if we take 21 for “nine” and -7 for “five”, the sum 14 must still name two o'clock. It does, since $21 + (-7) = 14$. The first check is the theory of equivalence relations and partitions. The second is called **well-definedness**, and it comes at the end of the section.

2.5.2 Equivalence relations

Equality has three properties that every argument with “=” uses silently: $x = x$; if $x = y$, then $y = x$; and if $x = y$ and $y = z$, then $x = z$ (Logic and Proof, Proposition 5.7.1). They are exactly what lets us replace a thing by an equal thing, as often as we like, in either direction. “Same hour on the clock”, “same remainder on division by 3” and “same absolute value” all have the same three properties, although none of them is equality. They deserve a common name.

An equivalence relation declares things the same in a way that behaves like equality.

Definition 2.5.1. Equivalence Relation

Let X be a set. An **equivalence relation on X** is a relation $\sim$ on X (Definition 1.3.3) that is reflexive, symmetric and transitive on X (Definition 2.4.1):

(Eq1) **for every** $x \in X$, $x \sim x$;

(Eq2) **for all** $x, y \in X$, if $x \sim y$, then $y \sim x$;

(Eq3) **for all** $x, y, z \in X$, if $x \sim y$ and $y \sim z$, then $x \sim z$.

When $x \sim y$, we say that x is **equivalent** to y (under $\sim$).

In words: (Eq1) says everything is the same as itself. (Eq2) says “the same” has no direction: if x counts as the same as y , then y counts as the same as x . (Eq3) says sameness passes along a chain: two things that are each the same as a third are linked through it. In symbols, (Eq1)–(Eq3) are

$$\begin{aligned} \forall x \in X \quad x \sim x, \quad \forall x, y \in X \quad (x \sim y \rightarrow y \sim x), \\ \forall x, y, z \in X \quad ((x \sim y \wedge y \sim z) \rightarrow x \sim z). \end{aligned}$$

The symbol $\sim$ is a letter for a relation, like R in Section 2.4; it is written between its arguments because it is read like $=$. Nothing needs checking for well-definedness: the definition names a property of a set of pairs.

Example. Four equivalence relations

Check (Eq1)–(Eq3) for each relation.

- (a) Equality on any set X : $x \sim y$ if and only if $x = y$.
- (b) The relation $X \times X$ on a set X , which relates every x to every y .
- (c) “Same absolute value” on $\mathbb{R}$: $x \sim y$ if and only if $|x| = |y|$.
- (d) The empty relation $\emptyset$ on the empty set.

Solution.

- (a) (Eq1)–(Eq3) are the three properties of equality recalled above: $x = x$ is =I, and symmetry and transitivity are Logic and Proof, Proposition 5.7.1.
- (b) Every pair (x, y) with $x, y \in X$ lies in $X \times X$, so each of (Eq1)–(Eq3) has a conclusion of the form $(u, v) \in X \times X$ with $u, v \in X$, and that conclusion is true. This is the coarsest sameness: everything is the same as everything.
- (c) Each of (Eq1)–(Eq3) for $\sim$ is the same property for equality, applied to the numbers $|x|, |y|, |z|$. For instance, (Eq3): if $|x| = |y|$ and $|y| = |z|$, then $|x| = |z|$ by transitivity of $=$. Here $3 \sim -3$, and $3 \not\sim 2$.
- (d) Each of (Eq1)–(Eq3) begins “for every element of X ”, and $X = \emptyset$ has none, so all three hold vacuously. The empty relation **is** an equivalence relation on $\emptyset$.

Part (d) is the degenerate case, and it is worth keeping: it says that the definition makes no demand on an empty set, so the theorems below hold for $X = \emptyset$ without a separate case. On a **non-empty** set the empty relation is not an equivalence relation: (Eq1) fails at any element x , since $(x, x) \notin \emptyset$. The same set of pairs passes on one set and fails on another, because (Eq1) speaks about every element of X .

Part (c) is an instance of a pattern that returns later in this section, with the fibers of a function: “ $x \sim y$ when some quantity attached to x equals the same quantity attached to y ”. Any such relation inherits (Eq1)–(Eq3) from equality.

A non-example by minimal change. On $\mathbb{R}$, let $x \sim y$ mean $x - y \in \mathbb{Z}$, and let $x R y$ mean $x - y \in \mathbb{N}$. The first is an equivalence relation, as an example later in this section checks. The second changes one letter. It is still reflexive, since $x - x = 0 \in \mathbb{N}$ by $((\mathbb{N})0)$, and still transitive: if $x - y \in \mathbb{N}$ and $y - z \in \mathbb{N}$, then $x - z = (x - y) + (y - z) \in \mathbb{N}$ by Lemma 2.1.1 (b). But it is not symmetric: $1 R 0$, since $1 - 0 = 1 \in \mathbb{N}$, while $0 - 1 = -1$, and $-1 \notin \mathbb{N}$ because every natural number is ≥ 0 (Lemma 2.1.1 (a)). The clause that fails is (Eq2). The relation R is the order $y \leq x$ restricted to pairs at integer distance; it has a direction, and sameness has none.

Why these three clauses. Each clause has a job, and each failure has a witness.

- Drop (Eq1), and an element may be the same as nothing at all. On $\{1\}$, the empty relation is symmetric and transitive (Section 2.4’s warning), and 1 is related to nothing; it would sit in no piece of the cutting.
- Drop (Eq2), and the pieces overlap. The relation $\leq$ on $\{1, 2\}$ is reflexive and transitive; the set of elements below 2 is $\{1, 2\}$, the set below 1 is $\{1\}$, and the two share 1 without being equal.
- Drop (Eq3), and sameness leaks. “ $|x - y| \leq 1$ ” on $\mathbb{R}$ is reflexive and symmetric, and 0 is related to 1, 1 to 2, 2 to 3, and so on, while 0 is not related to 2. Every integer is linked

to every other through a chain of steps of length 1, though few are related directly.

The name comes from “equal value”: equivalent elements are equal in whatever respect the relation measures.

Warning. An equivalence relation is an equivalence relation **on a named set**. The relation $\{(1, 1), (2, 2)\}$ is equality on $\{1, 2\}$, an equivalence relation; on $\{1, 2, 3\}$ it is the same set of pairs and not an equivalence relation, since $(3, 3)$ is missing and (Eq1) fails at 3. Before checking (Eq1), name the set.

The payoff is the object the clock was after: the collection of everything that counts as the same as a given element.

2.5.3 Equivalence classes

An equivalence class is everything that counts as the same as one element.

Definition 2.5.2. Equivalence Class

Let $\sim$ be an equivalence relation on a set X , and let $a \in X$. The **equivalence class** of a is

$$[a] := \{x \in X : x \sim a\}.$$

When the relation must be named, we write $[a]_{\sim}$. Any element of a class is called a **representative** of it.

The class is a subset of X , by specification (Definition 1.1.3). Its definition fixes the order of the letters as $x \sim a$; by (Eq2) the order does not matter, and $x \in [a]$ holds exactly when $a \sim x$.

For the four relations of Example (Four equivalence relations): under equality, $[a] = \{a\}$; under $X \times X$, $[a] = X$ for every $a \in X$; under “same absolute value”, $[3] = \{-3, 3\}$ and $[0] = \{0\}$; and on the empty set there are no elements and so no classes.

Quick check. On $\{1, 2, \dots, 30\}$, let $x \sim y$ mean that x and y have the same last digit. List the classes. How many are there, and how many elements has each?

Solution. The classes are $\{1, 11, 21\}$, $\{2, 12, 22\}$, $\dots$, $\{9, 19, 29\}$ and $\{10, 20, 30\}$: ten classes, three elements each. Each class is the class of any of its elements; for instance $[11] = [21] = [1]$.

The quick check shows the features the theorem below proves in general: every class is non-empty, no number is in two classes, and every number is in one. A family of subsets with these three features has a name.

Definition 2.5.3. Partition

Let X be a set. A **partition** of X is a set $\mathcal{P}$ of subsets of X , called **blocks**, such that

- (P1) every block is **non-empty**: $P \neq \emptyset$ for every $P \in \mathcal{P}$;
 (P2) the blocks are **pairwise disjoint**: for all $P, Q \in \mathcal{P}$, if $P \neq Q$, then $P \cap Q = \emptyset$;
 (P3) the blocks cover X : $\bigcup \mathcal{P} = X$.

Here $\bigcup \mathcal{P}$ is the union of the set of sets $\mathcal{P}$ (Definition 1.2.4): the set of points lying in some block. Since every block is a subset of X , the union is always a subset of X , and (P3) says that every $x \in X$ lies in **some** block. With (P2), it lies in **exactly one**: if $x \in P$ and $x \in Q$, then $P \cap Q \neq \emptyset$, so $P = Q$.

Examples: $\{\{1, 2\}, \{3\}\}$ is a partition of $\{1, 2, 3\}$. For a non-empty X , $\{X\}$ is a partition with one block, and $\{\{x\} : x \in X\}$ is one with a block for each point. The even and the odd integers form a partition of $\mathbb{Z}$ with two blocks, because every integer is even or odd and not both (Theorem 2.2.3 with $n = 2$). For $X = \emptyset$, the empty set of blocks $\mathcal{P} = \emptyset$ is a partition: (P1) and (P2) are vacuous, and $\bigcup \emptyset = \emptyset$.

Non-examples by minimal change. $\{\{1, 2\}, \{2, 3\}\}$ covers $\{1, 2, 3\}$ and its sets are non-empty, but 2 lies in both, so (P2) fails. $\{\{1, 2\}, \{3\}, \emptyset\}$ satisfies (P2) and (P3), and fails (P1). Without (P1), “the block containing x ” would still be unique, but a partition could carry useless empty blocks, and the correspondence with equivalence relations below would break, since no class is empty.

A partition is a set of subsets of X . So is the set of all classes of an equivalence relation, and it gets its own name.

Definition 2.5.4. Quotient Set

Let $\sim$ be an equivalence relation on a set X . The **quotient set** of X by $\sim$ is the set of all equivalence classes,

$$X/\sim := \{C \in \mathcal{P}(X) : C = [a] \text{ for some } a \in X\}.$$

The **quotient map** is $q: X \rightarrow X/\sim$, $q(a) := [a]$.

The quotient set is a set by specification inside $\mathcal{P}(X)$ (Definition 1.1.3, Definition 1.2.2), and it is often written $\{[a] : a \in X\}$. The quotient map is a function by Definition 1.4.1: each $a \in X$ has the value $[a]$, which lies in $X/\sim$, and only that value. It is **surjective**: an element of $X/\sim$ is $[a]$ for some $a \in X$, and then it is $q(a)$ (Definition 1.4.4). The quotient map forgets everything about a except which class it lies in, and that is the point of it. Read “ $X/\sim$ ” as “ X with equivalent elements identified”.

2.5.4 The classes form a partition

Here is the theorem the clock needs. It also says exactly when two names give the same class.

Theorem 2.5.1. Equivalence Classes Partition the Set

Let $\sim$ be an equivalence relation on a set X .

- (a) For every $a \in X$, $a \in [a]$; in particular every class is non-empty.
- (b) For all $a, b \in X$, $[a] = [b]$ **if and only if** $a \sim b$.
- (c) For all $a, b \in X$, if $[a] \cap [b] \neq \emptyset$, then $[a] = [b]$.

(d) The quotient set $X/\sim$ is a partition of X .

Idea. Each property of $\sim$ does one job. Reflexivity puts a in its own class, which gives (a) and the covering. For (b), a class is determined by any one of its elements: if $a \sim b$, then anything the same as a is the same as b , by transitivity, and the reverse needs symmetry first. For (c), a common element c of two classes has $c \sim a$ and $c \sim b$; symmetry turns the first around, transitivity joins the two, and (b) finishes. Then (d) is (a), (c) and the covering, read against the three clauses of [Definition 2.5.3](#).

Before reading on: which property of $\sim$ does each line use?

Proof. (a) Let $a \in X$. By (Eq1), $a \sim a$, so $a \in [a]$ by [Definition 2.5.2](#). Hence $[a] \neq \emptyset$.

(b) $(\Rightarrow)$ Suppose $[a] = [b]$. By (a), $a \in [a] = [b]$, so $a \sim b$.

$(\Leftarrow)$ Suppose $a \sim b$. $(\subseteq)$ Let $x \in [a]$, so $x \sim a$. With $a \sim b$, (Eq3) gives $x \sim b$, so $x \in [b]$.

$(\supseteq)$ Let $x \in [b]$, so $x \sim b$. By (Eq2) applied to $a \sim b$, $b \sim a$; with $x \sim b$, (Eq3) gives $x \sim a$, so $x \in [a]$. Hence $[a] = [b]$ by [Proposition 1.1.1](#).

(c) Suppose $[a] \cap [b] \neq \emptyset$, and let $c \in [a] \cap [b]$. Then $c \sim a$ and $c \sim b$. By (Eq2), $a \sim c$; with $c \sim b$, (Eq3) gives $a \sim b$. By (b), $[a] = [b]$.

(d) We check the clauses of [Definition 2.5.3](#) for $\mathcal{P} := X/\sim$, a set of subsets of X by [Definition 2.5.4](#).

(P1) Every element of $X/\sim$ is $[a]$ for some $a \in X$, and is non-empty by (a).

(P2) Let $C, D \in X/\sim$ with $C \neq D$, say $C = [a]$ and $D = [b]$. If $C \cap D \neq \emptyset$, then $C = D$ by (c), which is false. Hence $C \cap D = \emptyset$.

(P3) Every class is a subset of X , so $\bigcup X/\sim \subseteq X$. Conversely, let $x \in X$. Then $x \in [x]$ by (a), and $[x] \in X/\sim$, so $x \in \bigcup X/\sim$. Hence $\bigcup X/\sim = X$ by [Proposition 1.1.1](#).

This proves the theorem. ■

Part (b) is the rule we will use most. It turns a statement about classes, $[a] = [b]$, into a statement about elements, $a \sim b$, and back. On the clock: $[14] = [2]$ because $14 - 2 = 12$ is a multiple of 12. By (b), the quotient map q of [Definition 2.5.4](#) has $q(a) = q(b)$ exactly when $a \sim b$. So q is injective only when every class has a single element: two different elements x, y of one class $[c]$ have $[x] = [c] = [y]$ by (b), so $q(x) = q(y)$. Part (c) says that two classes are either equal or disjoint; there is no third possibility. That is what the relations of “Why these three clauses” lacked: for $\leq$, the sets $[2] = \{1, 2\}$ and $[1] = \{1\}$ given by the formula of [Definition 2.5.2](#) overlap and differ, and with $a = 2$, $b = 1$, $c = 1$ the line of (c) that fails is the one that uses (Eq2) (Exercise A1(c)).

Warning. A class has many names. In the quick check, $[1]$, $[11]$ and $[21]$ are one set, and nothing about that set prefers 1. So any argument or definition that starts from “ $[a]$ ” and then uses a itself (its size, its sign, its last digit) must be shown not to depend on which representative was chosen. “The smallest element of $[a]$ ” makes sense for the last-digit classes;

“ $[a+1]$ ” as a function of the class $[a]$ needs a proof, which the section supplies for congruences under “Operations on classes”.

2.5.5

Every partition comes from one equivalence relation

The theorem turns an equivalence relation into a partition. The reverse is just as natural: given the blocks, call two points “the same” when they lie in a common block.

Theorem 2.5.2. Partitions and Equivalence Relations Correspond

Let X be a set.

- (a) Let $\mathcal{P}$ be a partition of X . For $x, y \in X$, let $x \sim_{\mathcal{P}} y$ mean that there is a block $P \in \mathcal{P}$ with $x \in P$ and $y \in P$. Then $\sim_{\mathcal{P}}$ is an equivalence relation on X , and its classes are the blocks: $X/\sim_{\mathcal{P}} = \mathcal{P}$.
- (b) Let $\sim$ be an equivalence relation on X . Then $\sim_{X/\sim}$, the relation of (a) for the partition $X/\sim$, is $\sim$ itself.
- (c) Hence every partition of X is the quotient set of **exactly one** equivalence relation on X , namely $\sim_{\mathcal{P}}$.

Idea. For (a), (Eq1) and (Eq2) are immediate from (P3) and the symmetry of “lie in a common block”. The real work is (Eq3): x, y lie in one block and y, z in another, and we need one block holding x and z . The two blocks share y , so (P2) makes them equal. The classes are the blocks because each point lies in exactly one block. Then (b) is [Theorem 2.5.1](#) (b) in disguise, and (c) puts (a) and (b) together.

Proof. (a) (Eq1) Let $x \in X$. By (P3), $x \in \bigcup \mathcal{P}$, so $x \in P$ for some $P \in \mathcal{P}$; this block contains x and x , so $x \sim_{\mathcal{P}} x$. (Eq2) If a block contains x and y , it contains y and x . (Eq3) Let $x \sim_{\mathcal{P}} y$ and $y \sim_{\mathcal{P}} z$. Take a block P with $x, y \in P$, and a block Q with $y, z \in Q$. Then $y \in P \cap Q$, so $P \cap Q \neq \emptyset$, and $P = Q$ by (P2). Hence $x, z \in P$, and $x \sim_{\mathcal{P}} z$.

Now the classes. Let $x \in X$, and let P be a block with $x \in P$ (one exists by (P3)). We show $[x] = P$, the class taken under $\sim_{\mathcal{P}}$. ($\subseteq$) Let $y \in [x]$, so $y \sim_{\mathcal{P}} x$: some block Q contains y and x . Then $x \in P \cap Q$, so $Q = P$ by (P2), and $y \in P$. ($\supseteq$) Let $y \in P$. Then P contains y and x , so $y \sim_{\mathcal{P}} x$, and $y \in [x]$.

So every class is a block: $[x]$ is the block containing x . Conversely, every block P is non-empty by (P1); for any $x \in P$, the argument just given shows $P = [x]$. Hence $X/\sim_{\mathcal{P}} = \mathcal{P}$, each inclusion being one of the last two sentences.

(b) Both relations are subsets of $X \times X$, so we show that they contain the same pairs ([Definition 1.1.1](#)). Let $x, y \in X$. If $x \sim y$, then the class $[y] \in X/\sim$ contains y by [Theorem 2.5.1](#) (a) and, by [Definition 2.5.2](#), x ; so $x \sim_{X/\sim} y$. Conversely, if $x \sim_{X/\sim} y$, some class $[a]$ contains x and y , so $x \sim a$ and $y \sim a$. By (Eq2), $a \sim y$, and by (Eq3), $x \sim y$.

(c) By (a), $\mathcal{P}$ is the quotient set of $\sim_{\mathcal{P}}$. Suppose $\mathcal{P} = X/\sim$ for an equivalence relation $\sim$. Then, by (b), $\sim$ is $\sim_{X/\sim}$, the relation of (a) for the partition $X/\sim = \mathcal{P}$; that is, $\sim$ is $\sim_{\mathcal{P}}$. So

there is exactly one such relation. This proves the theorem. ■

So the two passages undo each other. Going from $\sim$ to $X/\sim$ and back gives $\sim$ again, by (b); going from $\mathcal{P}$ to $\sim_{\mathcal{P}}$ and back gives $\mathcal{P}$ again, by (a). An equivalence relation and a partition are two descriptions of one structure: the relation says which pairs are the same, and the partition lists the piles. This is the move *classes from a relation, a relation from classes*, and we choose whichever description makes the next step easier.

The transitivity step in (a) is where (P2) does its real work (the classes part uses it again, the same way), and it is the step that a hasty proof skips. If the blocks were allowed to overlap, as in the cover $\{\{1, 2\}, \{2, 3\}\}$ of $\{1, 2, 3\}$, then “lie in a common set” would relate 1 to 2 and 2 to 3 but not 1 to 3.

2.5.6 Congruence modulo n

Back to the clock, with any number of hours. The divisibility notation $d \mid m$, “there is $k \in \mathbb{Z}$ with $m = kd$ ”, is [Definition 2.2.1](#).

Definition 2.5.5. Congruence Modulo n

Let $n \in \mathbb{Z}_{>0}$. Integers a and b are **congruent modulo n** , written $a \equiv b \pmod{n}$, if $n \mid a - b$, that is, if $a - b = kn$ for some $k \in \mathbb{Z}$.

So $14 \equiv 2 \pmod{12}$, since $14 - 2 = 1 \cdot 12$; $-10 \equiv 2 \pmod{12}$, since $-10 - 2 = (-1) \cdot 12$; and $7 \not\equiv 2 \pmod{12}$, since 5 is not a multiple of 12 (the multiples $12k$ are 0 for $k = 0$, at least 12 for $k \geq 1$ and at most -12 for $k \leq -1$, by discreteness, [Proposition 2.1.2](#) (c)). The symbol $\equiv$ here always carries its modulus, so it does not clash with the semantic equivalence $\varphi \equiv \psi$ of *Logic and Proof*, Chapter 3.

Remark. Why $n \geq 1$? The condition makes sense for $n = 0$, but it says nothing new: $0 \mid a - b$ means $a - b = k \cdot 0 = 0$, so “congruence modulo 0” is equality on $\mathbb{Z}$. Its classes would be the singletons $\{a\}$, a copy of $\mathbb{Z}$ and not $\mathbb{Z}$ itself, and the count of classes proved below would fail. Excluding $n = 0$ keeps that count exact.

Proposition 2.5.3. Congruence Is an Equivalence Relation

For every $n \in \mathbb{Z}_{>0}$, congruence modulo n is an equivalence relation on $\mathbb{Z}$.

Proof. (Eq1) $a - a = 0 = 0 \cdot n$, and $0 \in \mathbb{Z}$. (Eq2) If $a - b = kn$ with $k \in \mathbb{Z}$, then $b - a = (-k)n$, and $-k = 0 - k \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a). (Eq3) If $a - b = kn$ and $b - c = ln$ with $k, l \in \mathbb{Z}$, then $a - c = (a - b) + (b - c) = (k + l)n$, and $k + l \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a). ■

The quotient set is written $\mathbb{Z}/n$, “the integers modulo n ”. Its elements are the **residue classes** $[a] = \{x \in \mathbb{Z} : x \equiv a \pmod{n}\} = \{a + kn : k \in \mathbb{Z}\}$; the second form holds because $x \equiv a$ means

$x - a = kn$, that is, $x = a + kn$. When two moduli are in play we write $[a]_n$ for the class of a in $\mathbb{Z}/n$. For example, $[2]_{12} = \{\dots, -22, -10, 2, 14, 26, \dots\}$, the integers that name two o'clock.

How many classes are there? The clock has twelve positions, and the division algorithm says why: each integer has exactly one remainder among $0, 1, \dots, 11$.

Proposition 2.5.4. The Integers Modulo n

Let $n \in \mathbb{Z}_{>0}$. For every $a \in \mathbb{Z}$ there is **exactly one** $r \in \{0, \dots, n-1\}$ with $[a] = [r]$ in $\mathbb{Z}/n$, namely the remainder of a on division by n . Hence

$$\mathbb{Z}/n = \{[0], [1], \dots, [n-1]\},$$

and the n classes listed are pairwise different.

Here $n-1 \in \mathbb{N}$ by Lemma 2.1.1 (c), since $n \neq 0$, so $\{0, \dots, n-1\}$ is the set of Section 2.1.

Proof. Let $a \in \mathbb{Z}$. By Theorem 2.2.3 there are $q, r \in \mathbb{Z}$ with $a = qn + r$ and $0 \leq r < n$. Since $r \geq 0$, $r \in \mathbb{N}$ (Proposition 2.1.2 (b)); since $r < n$, $r+1 \leq n$ by discreteness (Proposition 2.1.2 (c)), so $r \leq n-1$. Hence $r \in \{0, \dots, n-1\}$. Also $a - r = qn$, so $a \equiv r \pmod{n}$, and $[a] = [r]$ by Theorem 2.5.1 (b).

For uniqueness, let $r, s \in \{0, \dots, n-1\}$ with $[a] = [r]$ and $[a] = [s]$. Then $[r] = [s]$, so $r \equiv s \pmod{n}$ by Theorem 2.5.1 (b): $r - s = kn$ for some $k \in \mathbb{Z}$. So the integer r can be written both as $r = 0 \cdot n + r$ and as $r = kn + s$, and both remainders r and s satisfy $0 \leq r, s < n$. By the uniqueness in Theorem 2.2.3, $s = r$.

Finally, every element of $\mathbb{Z}/n$ is $[a]$ for some $a \in \mathbb{Z}$, hence equal to one of $[0], \dots, [n-1]$; and if $[r] = [s]$ with $r, s \in \{0, \dots, n-1\}$, the uniqueness just proved, applied to $a = r$, gives $r = s$. This proves the proposition. ■

So $\mathbb{Z}/12$ has exactly the twelve classes $[0], \dots, [11]$, one for each position of the hand. (The clock face writes 12 where we write $[0]$, and $[12] = [0]$.) The uniqueness also gives a test that is often quicker than the definition: $a \equiv b \pmod{n}$ **if and only if** a and b leave the same remainder on division by n . Indeed, if $a \equiv b$, then $[a] = [b]$, and the one r with $[a] = [r]$ is also the one with $[b] = [r]$; conversely, if both leave the remainder r , then $[a] = [r] = [b]$. With $n = 2$, the two classes of $\mathbb{Z}/2$ are the even integers $[0]$ and the odd integers $[1]$: the partition of $\mathbb{Z}$ into evens and odds is the quotient set of congruence modulo 2.

2.5.7 Two more sources of equivalence relations

Congruence is one way to manufacture an equivalence relation. Here is the most general way, and a second example on the real line that later chapters return to.

Example. The fibers of a function

Let $f: X \rightarrow Y$ be a function, and for $x, y \in X$ let $x \sim_f y$ mean $f(x) = f(y)$. Show that $\sim_f$ is an equivalence relation on X , and describe its classes. Then show that every equivalence relation $\sim$ on X is of this form, for a suitable function.

Solution. (Eq1)–(Eq3) for $\sim_f$ are the properties of equality applied to the values $f(x), f(y), f(z)$, exactly as in [Example \(Four equivalence relations\)](#) (c): $f(x) = f(x)$; $f(x) = f(y)$ gives $f(y) = f(x)$; and $f(x) = f(y), f(y) = f(z)$ give $f(x) = f(z)$ (Logic and Proof, Proposition 5.7.1).

The class of $a \in X$ is

$$[a] = \{x \in X : f(x) = f(a)\} = f^{-1}(\{f(a)\}),$$

the preimage of the one-point set $\{f(a)\}$ ([Definition 1.6.2](#)). Such a preimage $f^{-1}(\{y\})$ is called the **fiber** of f over y . So the classes of $\sim_f$ are the fibers over the points y of the range $f(X)$. A fiber over a point outside the range is empty, and is not a class: (P1) is what excludes it.

For the converse, let $\sim$ be an equivalence relation on X , and take for f the quotient map $q: X \rightarrow X/\sim$ of [Definition 2.5.4](#). Then, for $x, y \in X$,

$$x \sim_q y \iff q(x) = q(y) \iff [x] = [y] \iff x \sim y,$$

where the last step is [Theorem 2.5.1](#) (b). So $\sim_q$ and $\sim$ contain the same pairs, and are equal ([Definition 1.1.1](#)).

Two instances. For $f(x) = |x|$ on $\mathbb{R}$, $\sim_f$ is “same absolute value”, and the fiber over 3 is the class $\{-3, 3\}$. For the map $\mathbb{Z} \rightarrow \{0, 1, 2\}$ sending each integer to its remainder on division by 3 (a function by the uniqueness in [Theorem 2.2.3](#)), $\sim_f$ is congruence modulo 3, by the remainder test above. The example says that “the same in some respect” always means “equal value of some function”: an equivalence relation is the same thing as a function, up to the names of its values.

Example. The reals modulo the integers

On $\mathbb{R}$, let $x \sim y$ mean $x - y \in \mathbb{Z}$. Show that $\sim$ is an equivalence relation. Find the classes of 0, of $\frac{1}{2}$ and of $\sqrt{2}$, and show that these three classes are different.

Solution. An equivalence relation. (Eq1) $x - x = 0 \in \mathbb{Z}$. (Eq2) If $x - y \in \mathbb{Z}$, then $y - x = 0 - (x - y) \in \mathbb{Z}$, by [Proposition 2.1.2](#) (a). (Eq3) If $x - y \in \mathbb{Z}$ and $y - z \in \mathbb{Z}$, then $x - z = (x - y) + (y - z) \in \mathbb{Z}$, by [Proposition 2.1.2](#) (a).

The classes. $[0] = \{x \in \mathbb{R} : x - 0 \in \mathbb{Z}\} = \mathbb{Z}$. Next, $x \in [\frac{1}{2}]$ means $x - \frac{1}{2} = k$ for some $k \in \mathbb{Z}$, so

$$[\frac{1}{2}] = \{k + \frac{1}{2} : k \in \mathbb{Z}\} = \{\dots, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \dots\},$$

and in the same way $[\sqrt{2}] = \{\sqrt{2} + k : k \in \mathbb{Z}\} = \{\dots, \sqrt{2} - 1, \sqrt{2}, \sqrt{2} + 1, \dots\}$.

Three different classes. By [Theorem 2.5.1](#) (b), two classes $[x]$ and $[y]$ are equal exactly when $x - y \in \mathbb{Z}$. A real number t with $0 < t < 1$ is not an integer: if it were, discreteness ([Proposition 2.1.2](#) (c)) applied to $0 < t$ would give $1 \leq t$. Now $\frac{1}{2} - 0 = \frac{1}{2}$ lies strictly between 0 and 1, so $[\frac{1}{2}] \neq [0]$. Since $\sqrt{2} \geq 0$ and $(\sqrt{2})^2 = 2$, Logic and Proof, Theorem 2.4.5 rules out $\sqrt{2} \leq 1$ (it would give $2 \leq 1$) and $2 \leq \sqrt{2}$ (it would give $4 \leq 2$), so $1 < \sqrt{2} < 2$, and $0 < \sqrt{2} - 1 < 1$; then $\sqrt{2} - 1 \notin \mathbb{Z}$, and therefore $\sqrt{2} - 0 \notin \mathbb{Z}$ (else $\sqrt{2} - 1$ would be a difference of integers), so $[\sqrt{2}] \neq [0]$. Finally, in the same way $\sqrt{2} \leq \frac{1}{2}$ would give $2 \leq \frac{1}{4}$, and $\frac{3}{2} \leq \sqrt{2}$

would give $\frac{9}{4} \leq 2$, so $\frac{1}{2} < \sqrt{2} < \frac{3}{2}$, and $0 < \sqrt{2} - \frac{1}{2} < 1$, and $[\sqrt{2}] \neq [\frac{1}{2}]$.

Picture each class as one point of a circle of circumference 1: the real line is wound around it, and two numbers land on the same point when they differ by an integer. Each class has exactly one representative in $[0, 1)$: the class of $\sqrt{2}$ meets $[0, 1)$ in $\sqrt{2} - 1$, for instance. Proving this for every class needs the integer part of a real number, which *Analysis on the Real Line*, Chapter 1 constructs; we do not use it here. With $\mathbb{Q}$ in place of $\mathbb{Z}$, the relation “ $x - y \in \mathbb{Q}$ ” is again an equivalence relation on $\mathbb{R}$, and choosing one point from each of its classes is how *Measure and Integration*, Chapter 1 builds a set of real numbers that has no length.

2.5.8 Operations on classes

Now the second check of the clock. To add hours we want to define

$$[a] + [b] := [a + b] \quad \text{in } \mathbb{Z}/n.$$

The right-hand side is computed from the representatives a and b . If another pair of representatives of the same two classes gave a different class, the “definition” would assign two answers to one input, and [Definition 1.4.1](#) forbids that: a function gives each input **exactly one** value. So the definition is legitimate only if the answer does not depend on the representatives. That property is called **well-definedness**, and it was met in Section 1.4, where “write a rational number as p/q , then send it to its numerator p ” failed to be a function ([Example \(Relations that are not functions\)](#)).

Proposition 2.5.5. Congruence Respects Sums and Products

Let $n \in \mathbb{Z}_{>0}$ and $a, a', b, b' \in \mathbb{Z}$. If $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$, then

$$a + b \equiv a' + b' \pmod{n} \quad \text{and} \quad ab \equiv a'b' \pmod{n}.$$

Hence the operations $[a] + [b] := [a + b]$ and $[a][b] := [ab]$ on $\mathbb{Z}/n$ are **well defined**.

Idea. The sum is direct. For the product, the difference $ab - a'b'$ changes both factors at once; change one at a time, by adding and subtracting ab' .

Proof. Write $a - a' = kn$ and $b - b' = ln$ with $k, l \in \mathbb{Z}$. Then

$$(a + b) - (a' + b') = (a - a') + (b - b') = (k + l)n,$$

and $k + l \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a). For the product,

$$\begin{aligned} ab - a'b' &= a(b - b') + (a - a')b' \\ &= (al + kb')n, \end{aligned}$$

and $al + kb' \in \mathbb{Z}$ by [Proposition 2.1.2](#) (a). This proves both congruences.

For well-definedness, suppose $[a] = [a']$ and $[b] = [b']$. By [Theorem 2.5.1](#) (b), $a \equiv a'$ and $b \equiv b'$, so $a + b \equiv a' + b'$ and $ab \equiv a'b'$ by what was just proved, and $[a + b] = [a' + b']$ and $[ab] = [a'b']$, again by [Theorem 2.5.1](#) (b). So the class assigned to the pair of classes $([a], [b])$ is the same whichever representatives are used.

In the language of [Definition 1.4.1](#): the addition of $\mathbb{Z}/n$ is the function $\mathbb{Z}/n \times \mathbb{Z}/n \rightarrow \mathbb{Z}/n$ whose graph is the set of all pairs $(([a], [b]), [a + b])$ with $a, b \in \mathbb{Z}$. Clause (Fn1) holds because every element of $\mathbb{Z}/n$ is a class $[a]$; clause (Fn2) is the proposition. The clock sentence of the opening is now a true statement: in $\mathbb{Z}/12$,

$$[9] + [5] = [14] = [2],$$

the first equality by the definition of $+$, the second because $14 - 2 = 12$. And $[21] + [-7] = [14] = [2]$ as well, as the proposition guarantees.

Well-definedness can fail, and the next example shows a failure next to a success, with the minimal change between them.

Example. A rule on classes that is not a function

Decide whether each rule defines a function: (a) $F: \mathbb{Z}/6 \rightarrow \mathbb{Z}/4$, $F([a]_6) := [a]_4$; (b) $G: \mathbb{Z}/6 \rightarrow \mathbb{Z}/3$, $G([a]_6) := [a]_3$.

Solution.

- (a) No. The class $[0]_6$ is also $[6]_6$, since $6 - 0 = 1 \cdot 6$. Read with the representative 0, the rule gives $[0]_4$; read with 6, it gives $[6]_4 = [2]_4$, since $6 - 2 = 1 \cdot 4$. And $[0]_4 \neq [2]_4$, since $2 - 0 = 2$ is not a multiple of 4 ($0 < 2 < 4$, and the multiples of 4 are 0, at least 4, or at most -4). So one input receives two values, and (Fn2) of [Definition 1.4.1](#) fails.
- (b) Yes. If $[a]_6 = [a']_6$, then $a - a' = 6k$ for some $k \in \mathbb{Z}$ ([Theorem 2.5.1](#) (b)), so $a - a' = 3(2k)$ with $2k \in \mathbb{Z}$ ([Proposition 2.1.2](#) (a)), and $[a]_3 = [a']_3$. The value does not depend on the representative. What made it work is $3 \mid 6$: a multiple of 6 is a multiple of 3, while a multiple of 6, such as 6 itself, need not be a multiple of 4.

The general lesson: a rule “ $[a] \mapsto$ something computed from a ” defines a function on the classes exactly when equivalent representatives give the same output. The check always has the same shape: take $a \sim a'$, and show that the outputs agree.

2.5.9 Exercises

2.5.9.1 A. Check your understanding

Exercise. A1

Parts (a)–(d) are about [Theorem 2.5.1](#) and its proof.

- (a) Which property of $\sim$ is used to show that every class is non-empty, and which properties are used to show that two overlapping classes are equal?
- (b) Run the proof of part (c) for congruence modulo 3, with $a = 4$, $b = 10$ and the common element $c = 7$: write each line with its numbers.
- (c) On $\{1, 2, 3\}$, let $x \sim y$ mean $x \leq y$, and define $[a]$ by the formula of [Definition 2.5.2](#). Show that $[2]$ and $[1]$ overlap without being equal, and name the line of the proof of part (c) that fails, with $a = 2$, $b = 1$.

- (d) Give the idea of the proof of part (c) in two sentences.
 (e) How many classes has the relation $x^2 = y^2$ on $\{-2, -1, 0, 1, 2\}$? List them.

Solution.

- (a) Non-emptiness uses only reflexivity (Eq1): $a \sim a$ puts a in $[a]$. Overlapping classes are equal by symmetry (Eq2), which turns $c \sim a$ into $a \sim c$, and transitivity (Eq3), which joins $a \sim c$ and $c \sim b$; the final step, part (b), uses (Eq2) and (Eq3) again.
 (b) Since $7 - 4 = 3$, $7 \equiv 4 \pmod{3}$, so $7 \in [4]$; since $7 - 10 = -3$, $7 \equiv 10 \pmod{3}$, so $7 \in [10]$. By (Eq2), $4 \equiv 7$ (indeed $4 - 7 = -3$). By (Eq3), $4 \equiv 10$ (indeed $4 - 10 = -6 = (-2) \cdot 3$). By part (b), $[4] = [10]$.
 (c) $[2] = \{x : x \leq 2\} = \{1, 2\}$ and $[1] = \{x : x \leq 1\} = \{1\}$. They share 1 and differ. In the proof of (c), take $c = 1$: $c \leq 2$ and $c \leq 1$ hold. The next line, “by (Eq2), $a \sim c$ ”, would need $2 \leq 1$, which is false: $\leq$ is not symmetric, and that line fails.
 (d) A common element of two classes is equivalent to both representatives, so symmetry and transitivity make the two representatives equivalent. Equivalent representatives have the same class, because each class is determined by any one of its elements.
 (e) Three: $\{0\}$, $\{-1, 1\}$ and $\{-2, 2\}$. This is the relation $\sim_f$ of [Example \(The fibers of a function\)](#) for $f(x) = x^2$, whose values on the set are 0, 1, 4, one class for each value.

2.5.9.2 B. Practice

Exercise. B1: Fractions as classes

- (a) On $\mathbb{Z} \times \mathbb{Z}_{>0}$, let $(a, b) \sim (c, d)$ mean $ad = bc$. Prove that $\sim$ is an equivalence relation, and describe the class of $(1, 2)$.
 (b) On $\mathbb{Z} \times \mathbb{N}$, let $(a, b) R (c, d)$ again mean $ad = bc$. Determine which of (Eq1)–(Eq3) hold. Justify your answer.

Solution.

- (a) (Eq1) $ab = ba$. (Eq2) If $ad = bc$, then $cb = da$. (Eq3) Let $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$, so $ad = bc$ and $cf = de$. Multiply the first equation by f and use the second: $adf = bcf = bde$. So $d(af - be) = 0$. Since $d \in \mathbb{Z}_{>0}$, $d \neq 0$, and we may cancel d : $af = be$, that is, $(a, b) \sim (e, f)$. The hypothesis $d \neq 0$ on the **middle** pair is used here and nowhere else.
 The class of $(1, 2)$ is $\{(c, d) \in \mathbb{Z} \times \mathbb{Z}_{>0} : d = 2c\}$. Since $d > 0$, $c = d/2 > 0$, so $c \in \mathbb{Z}_{>0}$ ([Proposition 2.1.2](#) (b)), and the class is $\{(k, 2k) : k \in \mathbb{Z}_{>0}\} = \{(1, 2), (2, 4), (3, 6), \dots\}$: the ways of writing $\frac{1}{2}$ as a fraction. Each class is “one rational number”; *Number Systems: From $\mathbb{N}$ to $\mathbb{R}$* , Chapter 2 builds $\mathbb{Q}$ as this quotient set.
 (b) (Eq1) and (Eq2) hold, by the same two lines as in (a), which did not use $b, d \neq 0$. (Eq3) fails: $(1, 1) R (0, 0)$, since $1 \cdot 0 = 1 \cdot 0$, and $(0, 0) R (2, 1)$, since $0 \cdot 1 = 0 \cdot 2$; but $(1, 1) \not R (2, 1)$, since $1 \cdot 1 \neq 1 \cdot 2$. The pair $(0, 0)$ is related to every pair, so it links any two; this is exactly the cancellation of (a) failing for $d = 0$.

Exercise. B2: Computing in $\mathbb{Z}/7$

In $\mathbb{Z}/7$, compute $[3] + [5]$ and $[3][5]$, each written as $[r]$ with $r \in \{0, \dots, 6\}$. Then prove that $a \equiv b \pmod{n}$ implies $a^k \equiv b^k \pmod{n}$ for every $k \in \mathbb{N}$, and use it to find the remainder of 3^{100} on division by 7. *Hint:* $3^6 = 729$.

Solution. $[3] + [5] = [8] = [1]$, since $8 - 1 = 7$; and $[3][5] = [15] = [1]$, since $15 - 1 = 2 \cdot 7$.

Powers. Let $n \in \mathbb{Z}_{>0}$, $a, b \in \mathbb{Z}$ with $a \equiv b \pmod{n}$. By [Proposition 2.3.4](#) (d), a^k and b^k are integers. We prove $a^k \equiv b^k \pmod{n}$ by induction on $k \in \mathbb{N}$ (Logic and Proof, Axiom 4.2.1). Base: $a^0 = 1 = b^0$, and $1 \equiv 1$ by (Eq1). Step: suppose $a^k \equiv b^k$. By [Definition 2.3.2](#), $a^{k+1} = a^k a$ and $b^{k+1} = b^k b$; the product clause of [Proposition 2.5.5](#), applied to $a^k \equiv b^k$ and $a \equiv b$, gives $a^{k+1} \equiv b^{k+1}$.

The remainder. By [Proposition 2.3.4](#), $3^{100} = 3^{96} \cdot 3^4 = (3^6)^{16} \cdot 3^4$. Now $729 = 104 \cdot 7 + 1$, so $3^6 \equiv 1 \pmod{7}$, and by the power fact, $(3^6)^{16} \equiv 1^{16} = 1$. Also $3^4 = 81 = 11 \cdot 7 + 4$, so $3^4 \equiv 4$. By [Proposition 2.5.5](#), $3^{100} \equiv 1 \cdot 4 = 4 \pmod{7}$. Since $0 \leq 4 < 7$, the remainder of 3^{100} on division by 7 is 4 ([Proposition 2.5.4](#)), and $[3^{100}] = [4]$ in $\mathbb{Z}/7$. Here $1^{16} = 1$ is itself an induction: $1^0 = 1$, and $1^{k+1} = 1^k \cdot 1 = 1$.

Exercise. B3: Which rules are well defined?

Determine which of the following rules define a function. Justify your answer.

- (a) $\mathbb{Z}/n \rightarrow \mathbb{Z}/n$, $[a]_n \mapsto [a^2]_n$, for a fixed $n \in \mathbb{Z}_{>0}$.
- (b) $\mathbb{Z}/5 \rightarrow \mathbb{Z}/10$, $[a]_5 \mapsto [a]_{10}$.
- (c) $\mathbb{Z}/10 \rightarrow \mathbb{Z}/5$, $[a]_{10} \mapsto [a]_5$.

Solution. The check in each case: take $[a] = [a']$, that is, $a \equiv a'$, and decide whether the outputs agree ([Example \(A rule on classes that is not a function\)](#)).

- (a) Yes. If $a \equiv a' \pmod{n}$, the product clause of [Proposition 2.5.5](#), applied to $a \equiv a'$ twice, gives $a \cdot a \equiv a' \cdot a'$, so $[a^2]_n = [a'^2]_n$.
- (b) No. $[0]_5 = [5]_5$, since $5 - 0 = 5$. But $[0]_{10} \neq [5]_{10}$, since $5 - 0 = 5$ is not a multiple of 10 ($0 < 5 < 10$). The input $[0]_5$ would get two values.
- (c) Yes. If $a - a' = 10k$ with $k \in \mathbb{Z}$, then $a - a' = 5(2k)$ with $2k \in \mathbb{Z}$ ([Proposition 2.1.2](#) (a)), so $[a]_5 = [a']_5$. The difference from (b) is the direction: $5 \mid 10$, so a multiple of 10 is a multiple of 5, not conversely.

2.5.9.3 C. Going deeper**Exercise. C1: Exit check: classes partition**

Let $\sim$ be an equivalence relation on a set X . Prove that the equivalence classes of $\sim$ form a partition of X , and that for all $a, b \in X$, $[a] = [b]$ if and only if $a \sim b$. Write the proof without looking at the text, naming the property of $\sim$ behind each step.

Solution. Non-empty. For $a \in X$, reflexivity gives $a \sim a$, so $a \in [a]$, and $[a] \neq \emptyset$.

Cover. Every class is a subset of X by definition. Every $x \in X$ lies in the class $[x]$, as just shown. So the union of all classes is X .

$[a] = [b] \iff a \sim b$. ($\Rightarrow$) $a \in [a] = [b]$, so $a \sim b$. ($\Leftarrow$) Let $a \sim b$. If $x \in [a]$, then $x \sim a$ and $a \sim b$, so $x \sim b$ by transitivity: $[a] \subseteq [b]$. By symmetry $b \sim a$; if $x \in [b]$, then $x \sim b$ and $b \sim a$ give $x \sim a$ by transitivity: $[b] \subseteq [a]$. So $[a] = [b]$.

Disjoint or equal. Let $[a] \cap [b] \neq \emptyset$, with a common element c . Then $c \sim a$ and $c \sim b$. By symmetry $a \sim c$, and by transitivity $a \sim b$. By the previous paragraph, $[a] = [b]$. Hence two different classes are disjoint.

The three paragraphs on non-emptiness, covering and disjointness are the three clauses of [Definition 2.5.3](#), so the set of classes $X/\sim$ is a partition of X . (This is [Theorem 2.5.1](#).)

Marking scheme. (10 points)

- 2: each class is non-empty, by reflexivity.
- 1: every element lies in some class (its own), so the classes cover X .
- 4: overlapping classes are equal: a common element c with $c \sim a$ and $c \sim b$; symmetry to reverse one of them; transitivity to link a and b ; both inclusions (or the equivalence $[a] = [b] \iff a \sim b$) to conclude equality.
- 2: $[a] = [b] \iff a \sim b$, both directions.
- 1: the conclusion names the three conditions of a partition.

A proof that shows “overlapping classes are equal” by proving $[a] \subseteq [c] \subseteq [b]$ and back directly from the definition earns the 4 points for that step if each use of symmetry and transitivity is named. So does a proof that, from $c \sim a$ and $c \sim b$, applies the ($\Leftarrow$) direction of $[a] = [b] \iff a \sim b$ twice to get $[a] = [c] = [b]$, provided that equivalence was proved independently of this step, with symmetry and transitivity named. Fatal errors (0 for the step): applying transitivity to $a \sim c$ and $b \sim c$ without reversing one by symmetry; showing only $[a] \subseteq [b]$ and concluding $[a] = [b]$.

Exercise. C2: Find the error

A student proves part of [Theorem 2.5.2](#). Let $\mathcal{P}$ be a partition of X , and let $x \sim y$ mean that x and y lie in a common block of $\mathcal{P}$.

Claim. $\sim$ is transitive. *Proof.* (1) Let $x, y, z \in X$ with $x \sim y$ and $y \sim z$. (2) Since $x \sim y$, there is a block $P \in \mathcal{P}$ with $x \in P$ and $y \in P$. (3) Since $y \sim z$, y and z lie in a common block, and that block is P , so $z \in P$. (4) Hence x and z both lie in P , and $x \sim z$.

- (a) Is the claim true?
- (b) Which line fails, and which rule does it break?
- (c) Give a counterexample to the failing step, and repair the proof.

Solution.

- (a) The claim is true: it is the transitivity step (Eq3) in the proof of [Theorem 2.5.2](#) (a).

- (b) Line (3) fails. “ y and z lie in a common block” is an existential statement, and to use it the proof must name its witness by $\exists E$, with a **new** name. Line (3) gives the second witness the name P , already in use for the witness of line (2): this is entry (N4) of Logic and Proof, Example (What is not allowed), reusing one witness name for two $\exists E$ ’s (Logic and Proof, Definition 5.6.1). The tell is that the proof never uses (P2), the disjointness of the blocks; and without (P2) the claim is false, as (c) shows.
- (c) A counterexample to the step needs blocks that overlap, since for a partition the step’s conclusion happens to be true. Take the cover $\{\{1, 2\}, \{2, 3\}\}$ of $\{1, 2, 3\}$, which is not a partition, with $x = 1$, $y = 2$, $z = 3$. Then $1 \sim 2$ with $P = \{1, 2\}$, and $2 \sim 3$ through $\{2, 3\}$; line (3) concludes $3 \in \{1, 2\}$, which is false, and the argument “proves” $1 \sim 3$, although no set of the cover contains 1 and 3.

Repair: replace (3) and (4) by “Since $y \sim z$, there is a block $Q \in \mathcal{P}$ with $y \in Q$ and $z \in Q$. Then $y \in P \cap Q$, so $P \cap Q \neq \emptyset$, and $P = Q$ by (P2). Hence $x \in P$ and $z \in P$, so $x \sim z$.”

2.6 Partial Orders, Maximal and Greatest Elements

Section 2.5 studied the relations that behave like equality. This section studies the ones that behave like $\leq$: reflexive and transitive like an equivalence relation, but antisymmetric instead of symmetric, so that they have a direction. For numbers, “the largest element of a set” is clear. For sets under $\subseteq$, or for natural numbers under divisibility, two elements can be **incomparable**, neither below the other, and “largest” splits into two different notions. We separate them, prove what links them, and then take a first look at least upper bounds. That is the vocabulary *Analysis on the Real Line*, Chapter 1 needs for the supremum.

2.6.1 Partial orders

Section 2.4 found three relations with the same three properties: $\leq$ on $\mathbb{R}$, $\subseteq$ on $\mathcal{P}(X)$ (Example (Eight relations, four properties)), and $|$ on $\mathbb{N}$ (Proposition 2.4.1). Each is reflexive, antisymmetric and transitive. Each says “below” in its own sense: smaller, contained in, a divisor of. What they share is what “being below” needs, and it deserves a name, so that a fact proved once holds for all three.

A partial order says which elements lie below which, without requiring every two to be comparable.

Definition 2.6.1. Partial Order

Let X be a set. A **partial order on X** is a relation $\preceq$ on X that is reflexive, antisymmetric and transitive on X (Definition 2.4.1):

- (PO1) **for every** $x \in X$, $x \preceq x$;
 (PO2) **for all** $x, y \in X$, if $x \preceq y$ and $y \preceq x$, then $x = y$;
 (PO3) **for all** $x, y, z \in X$, if $x \preceq y$ and $y \preceq z$, then $x \preceq z$.

The pair $(X, \preceq)$ is a **partially ordered set**. Elements $x, y \in X$ are **comparable** if $x \preceq y$ or $y \preceq x$. The partial order is **total** if **every** two elements of X are comparable. The **strict**

part of $\preceq$ is the relation $\prec$ on X given by

$$x \prec y \iff x \preceq y \text{ and } x \neq y.$$

We also write $y \succeq x$ for $x \preceq y$.

In words: (PO1) everything is below itself. (PO2) is the clause that gives an order its direction: two elements below each other are the same element, so “below” cannot run both ways between different things. (PO3) lets “below” pass along a chain. “Total” adds that any two elements can be compared; “partial” is the word for not requiring it. The symbol $\preceq$ is a letter for an arbitrary partial order, chosen so that $\leq$ keeps its meaning on numbers.

The strict part is irreflexive and transitive. No x has $x \prec x$, since that would need $x \neq x$. If $x \prec y$ and $y \prec z$, then $x \preceq z$ by (PO3); and $x \neq z$, because $x = z$ would give $x \preceq y$ and $y \preceq x$, hence $x = y$ by (PO2), contradicting $x \neq y$. So $x \prec z$.

Examples.

- $(\mathbb{R}, \leq)$ is a partially ordered set, and it is total: for real numbers $x \leq y$ or $y \leq x$. Its strict part is $<$. The same holds for $(\mathbb{N}, \leq)$ and $(\mathbb{Z}, \leq)$, since restricting a total order to a subset keeps all four properties: each is a statement about all elements, and holds in particular for the elements of the subset.
- $(\mathbb{N}, |)$ is a partially ordered set by [Proposition 2.4.1](#), and it is not total: $2 \nmid 3$ and $3 \nmid 2$, so 2 and 3 are incomparable.
- $(\mathcal{P}(X), \subseteq)$ is a partially ordered set; the example after this list decides when it is total.
- Degenerate cases: equality on any set X is a partial order, in which no two **different** elements are comparable; and the empty relation is a partial order on $\emptyset$, vacuously.

Non-examples by minimal change. Divisibility on $\mathbb{Z}$ is reflexive and transitive, but not antisymmetric: $2 \mid -2$ and $-2 \mid 2$, and $2 \neq -2$ ([Proposition 2.4.1](#)). Passing from $\mathbb{N}$ to $\mathbb{Z}$ loses exactly (PO2). The relation $<$ on $\mathbb{R}$ is antisymmetric (vacuously: $x < y$ and $y < x$ never both hold) and transitive, but not reflexive, since $0 < 0$ fails. It is the strict part of $\leq$, not a partial order.

Why these clauses. Without (PO2), two different elements could each be below the other, and “the largest element” of a set could fail to be unique; the uniqueness proof in the next subsection uses (PO2) and nothing else. A relation that keeps (PO1) and (PO3) and drops (PO2) is often called a preorder; $|$ on $\mathbb{Z}$ is one. Totality is left out because the important examples, $\subseteq$ and $|$, do not have it, and much of the interest lies in what changes without it.

Example. Inclusion is a partial order, rarely total

Let X be a set. Show that $\subseteq$ is a partial order on $\mathcal{P}(X)$, and that it is total if and only if X has at most one element.

Solution. A partial order. (PO1) $A \subseteq A$ for every set A . (PO2) If $A \subseteq B$ and $B \subseteq A$, then $A = B$ by [Proposition 1.1.1](#). (PO3) is [Proposition 1.1.2](#). (These are the entries of [Example \(Eight relations, four properties\)](#).)

Total exactly when X has at most one element. Suppose X has two different elements $a \neq b$. Then $\{a\}, \{b\} \in \mathcal{P}(X)$, and $\{a\} \not\subseteq \{b\}$, since $a \in \{a\}$ and $a \notin \{b\}$; likewise $\{b\} \not\subseteq \{a\}$. So $\{a\}$ and $\{b\}$ are incomparable, and $\subseteq$ is not total. For $X = \{1, 2\}$, the incomparable pair is $\{1\}, \{2\}$.

Conversely, suppose any two elements of X are equal. Let $A \subseteq X$. If $A \neq \emptyset$, take $a \in A$; then every $x \in X$ equals a , so $x \in A$, and $A = X$. So every subset of X is $\emptyset$ or X . Any two of these are comparable, since $\emptyset \subseteq X$ (Section 1.1, after [Proposition 1.1.3](#)) and each is a subset of itself. So $\subseteq$ is total on $\mathcal{P}(X)$.

Warning. In a partial order, “ $x \preceq y$ fails” does **not** give “ $y \prec x$ ”. In $(\mathcal{P}(\{1, 2\}), \subseteq)$, $\{1\} \not\subseteq \{2\}$, and $\{2\} \subsetneq \{1\}$ is false too: the two are incomparable. The habit from numbers, “not $\leq$ means $>$ ”, is the totality of $\leq$ on $\mathbb{R}$, and it must not be carried into a partial order.

2.6.2 Bounds, greatest and maximal elements

From now on, $(X, \preceq)$ is a partially ordered set and $A \subseteq X$. The bounds come first, because both kinds of “largest” are measured against them.

Definition 2.6.2. Upper and Lower Bounds

An element $u \in X$ is an **upper bound** of A if $a \preceq u$ for every $a \in A$, and a **lower bound** of A if $u \preceq a$ for every $a \in A$. The set A is **bounded above** if it has an upper bound in X , **bounded below** if it has a lower bound in X , and **bounded** if both.

An upper bound lives in X and need not belong to A : in $(\mathbb{R}, \leq)$, 5 is an upper bound of $[0, 1]$. Every element of X is an upper bound of $\emptyset$, vacuously. In $(\mathbb{N}, |)$, an upper bound of $\{4, 6\}$ is a common multiple of 4 and 6, such as 12, 24 or 0.

Now the two notions of “largest”. The first is the familiar one: an element of the set that is above everything in the set. The second asks less: an element with nothing in the set strictly above it.

Definition 2.6.3. Greatest and Least Elements

An element g is a **greatest element** of A if $g \in A$ and $a \preceq g$ for **every** $a \in A$. An element l is a **least element** of A if $l \in A$ and $l \preceq a$ for every $a \in A$.

Definition 2.6.4. Maximal and Minimal Elements

An element m is a **maximal element** of A if $m \in A$ and there is **no** $a \in A$ with $m \prec a$. An element m is a **minimal element** of A if $m \in A$ and there is no $a \in A$ with $a \prec m$.

In words: a greatest element is an upper bound of A that belongs to A . A maximal element is one that cannot be beaten inside A , which is weaker: an element can be unbeaten simply because it is incomparable with its rivals. In symbols the difference is where the negation sits:

$$\begin{aligned} g \text{ greatest: } & g \in A \wedge \forall a \in A \ a \preceq g, \\ m \text{ maximal: } & m \in A \wedge \neg \exists a \in A \ m \prec a. \end{aligned}$$

By Logic and Proof, Proposition 6.2.3 (b), $\neg \exists a \in A \ m \prec a$ is $\forall a \in A \ \neg(m \prec a)$, and $\neg(m \prec a)$ says that $m \preceq a$ implies $m = a$. So “maximal” reads: everything in A that is above m **is** m . “Greatest” reads: everything in A is below g . The first only forbids elements above; the second demands that every element be compared.

For numbers the two coincide. In $(\mathbb{R}, \leq)$, the set $\{1, 4, 9\}$ has greatest element 9, its only maximal element, and least element 1. The interval $[0, 1)$ has least element 0 and neither a greatest nor a maximal element: for each $a \in [0, 1)$, the number $(a + 1)/2$ lies in $[0, 1)$ and is strictly above a . A one-point set $\{x\}$ has x as greatest, least, maximal and minimal element, and $\emptyset$ has none of the four, since each definition requires an element of A .

The first result explains the article in “**the** greatest element”.

Proposition 2.6.1. Greatest and Least Elements Are Unique

A subset A of a partially ordered set has at most one greatest element and at most one least element.

Proof. Let g and g' be greatest elements of A . Since $g' \in A$ and g is greatest, $g' \preceq g$; since $g \in A$ and g' is greatest, $g \preceq g'$. By (PO2), $g = g'$. For least elements the same argument applies with the inequalities reversed. ■

So we may speak of **the** greatest element and **the** least element of A when they exist. In $(\mathbb{R}, \leq)$, the greatest element of a set is its maximum, $\max A$, and the least element its minimum, $\min A$. This settles a point left open in Section 1.4: the least $q_0 \in \mathbb{Z}_{>0}$ with $rq_0 \in \mathbb{Z}$ in [Example \(Relations that are not functions\)](#) exists by [Theorem 2.1.3](#): the set of such q lies in $\mathbb{Z}_{>0} \subseteq \mathbb{N}$, and it is non-empty, since writing $r = p/q$ with $q \in \mathbb{Z}_{>0}$ (as the solution of (b) there notes) gives $rq = p \in \mathbb{Z}$. It is unique by the least-element half of the proposition, so the rule there names one value.

The second result compares the two notions.

Proposition 2.6.2. Greatest and Maximal Elements

Let $(X, \preceq)$ be a partially ordered set and $A \subseteq X$.

- (a) A greatest element of A is a maximal element of A .
- (b) If $\preceq$ is total, then a maximal element of A is a greatest element of A .

The same holds for least and minimal elements.

Idea. For (a), anything strictly above the greatest element would also be below it, and antisymmetry forbids that. For (b), a maximal element m must be compared with each $a \in A$, and totality says it can be: either $a \preceq m$, as wanted, or $m \preceq a$, and then maximality forces $a = m$.

Proof. (a) Let g be a greatest element of A . Then $g \in A$. Suppose some $a \in A$ has $g \prec a$, so $g \preceq a$ and $g \neq a$. Since g is greatest, $a \preceq g$. By (PO2), $a = g$, contradicting $g \neq a$. So no such a exists, and g is maximal.

(b) Let $\preceq$ be total and m a maximal element of A , and let $a \in A$. By totality, $a \preceq m$ or $m \preceq a$. *Case 1:* $a \preceq m$. This is what we want. *Case 2:* $m \preceq a$. If $m \neq a$, then $m \prec a$ with

$a \in A$, which maximality excludes. So $m = a$, and $a \preceq m$ by (PO1). In either case $a \preceq m$. Since $a \in A$ was arbitrary and $m \in A$, m is a greatest element.

For least and minimal elements, apply (a) and (b) to the partial order $\succeq$: its greatest and maximal elements are the least and minimal elements for $\preceq$, and it is total when $\preceq$ is. (That $\succeq$ satisfies (PO1)–(PO3), and is total when $\preceq$ is, comes from reading those clauses with the two variables exchanged. Its strict part is $x \succ y \iff y \prec x$, since $x \neq y$ is symmetric in x and y . So “no $a \in A$ with $m \succ a$ ” is “no $a \in A$ with $a \prec m$ ”, and “ $a \succeq l$ for every $a \in A$ ” is “ $l \preceq a$ for every $a \in A$ ”).

■

Part (b) uses totality once, in the case split of its second sentence, and the case split is where it fails without it. The proof of [Proposition 2.3.8](#) in Section 2.3 compares two elements at a time and keeps the larger. Its maximum half uses nothing about $\leq$ on $\mathbb{R}$ beyond reflexivity, transitivity and the comparability of any two elements, so it works word for word in any totally ordered set. The minimum half passes to $-x_1, \dots, -x_n$, which needs an ordered field; in a totally ordered set, apply the maximum half to the reverse order $\succeq$ instead, which is total by the last paragraph of the proof above. In particular the whole proposition holds in any ordered field, which is how *Analysis on the Real Line*, Chapter 1 reads it.

Without totality, (b) fails, and maximal elements need not be greatest.

Example. Maximal elements that are not greatest

Find the maximal elements of each set, and decide whether it has a greatest element: (a) $A = \{2, 3, 4, 5, 6\}$ in $(\mathbb{N}, |)$; (b) $\mathcal{B} = \{\{1\}, \{2\}\}$ in $(\mathcal{P}(\{1, 2\}), \subseteq)$.

Solution.

- (a) Within A , the only strict divisibilities are $2 \mid 4$, $2 \mid 6$ and $3 \mid 6$. So $2 \prec 4$ and $3 \prec 6$, and 2, 3 are not maximal. Nothing in A is strictly above 4, 5 or 6: every element of A is positive, and a positive multiple kn of $n \in \{4, 5, 6\}$ other than n has $k > 0$ and $k \neq 1$, so $k \geq 2$ ([Proposition 2.1.2](#) (c)), and then $kn \geq 2n \geq 8$, while every element of A is at most 6. So the maximal elements are 4, 5 and 6.

There is no greatest element. By [Proposition 2.6.2](#) (a), a greatest element would be maximal, so it would be 4, 5 or 6. But $5 \nmid 4$, $4 \nmid 5$ and $4 \nmid 6$, so none of the three is above every element of A .

- (b) $\{1\}$ and $\{2\}$ are incomparable ([Example \(Inclusion is a partial order, rarely total\)](#)), so neither is strictly above the other: both are maximal. Neither is greatest, since neither contains the other.

In (a), the minimal elements are 2, 3 and 5, and there is no least element either. The example is the minimal change from $\{1, 2, 3, 4, 5, 6\}$, which has the least element 1 and still no greatest.

Warning. “Maximal” is not “greatest”. A set can have many maximal elements and no greatest, as in [Example \(Maximal elements that are not greatest\)](#); and a set can have no maximal element at all: $\mathbb{N}$ under $\leq$, where $n + 1 \in \mathbb{N}$ (by (NS)) is strictly above n . The answer also depends on the order, not just on the set: under divisibility, $\mathbb{N}$ **has** a greatest

element, 0, since every $a \in \mathbb{N}$ divides $0 = 0 \cdot a$.

Quick check. In $\mathcal{P}(\{1, 2, 3\}) \setminus \{\{1, 2, 3\}\}$, ordered by $\subseteq$, name the maximal elements. Is there a greatest element?

Solution. The maximal elements are the three two-element sets $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$: the only subset of $\{1, 2, 3\}$ strictly containing a two-element set is $\{1, 2, 3\}$, which was removed, and every set with at most one element lies strictly inside some two-element set. There is no greatest element: it would be maximal (Proposition 2.6.2 (a)), hence a two-element set, and no two-element set contains the other two.

2.6.3 Least upper bounds, a first look

Here is the notion that the rest of the book is built on. When a set has no greatest element, its upper bounds may still have a least one, and that element plays the part of “the top of the set” from outside.

The supremum is the best upper bound: the least element of the set of upper bounds.

Definition 2.6.5. Supremum and Infimum

Let $(X, \preceq)$ be a partially ordered set and $A \subseteq X$. An element $s \in X$ is a **least upper bound**, or **supremum**, of A in X if

- (S1) s is an upper bound of A , and
- (S2) $s \preceq u$ for **every** upper bound u of A .

An element $i \in X$ is a **greatest lower bound**, or **infimum**, of A in X if i is a lower bound of A and $u \preceq i$ for every lower bound u of A .

In other words, a supremum of A is a least element of the set of upper bounds of A , and an infimum a greatest element of the set of lower bounds. By Proposition 2.6.1, applied to those sets, each is **unique** when it exists, and we write $\sup A$ and $\inf A$. The words “in X ” matter: the upper bounds are taken in X , and enlarging or shrinking X can create or destroy a supremum. The definition also carries no promise that $\sup A$ exists, or that it belongs to A .

Two quick consequences. If A has a greatest element g , then $g = \sup A$: g is an upper bound, and every upper bound u has $g \preceq u$ because $g \in A$. And a supremum that belongs to A is a greatest element of A , by (S1). So a supremum is a greatest element exactly when it lies in the set; the new cases are the ones where it lies outside.

Suprema of sets. In $(\mathcal{P}(X), \subseteq)$, let $A, B \subseteq X$. Then $\sup\{A, B\} = A \cup B$. (S1): $A \subseteq A \cup B$ and $B \subseteq A \cup B$, and $A \cup B \in \mathcal{P}(X)$. (S2): let $C \in \mathcal{P}(X)$ be an upper bound, so $A \subseteq C$ and $B \subseteq C$; every element of $A \cup B$ lies in A or in B , hence in C , so $A \cup B \subseteq C$. The supremum $A \cup B$ belongs to $\{A, B\}$ only when one of the two sets contains the other; for $\{\{1\}, \{2\}\}$ it is $\{1, 2\}$, which is not in the set. Exercise B1 extends this to every family of subsets, and adds the infimum.

Suprema under divisibility. In $(\mathbb{N}, |)$, an upper bound of $\{a, b\}$ is a common multiple of a and b . The least upper bound turns out to be the least common multiple, and proving that it is **least in the divisibility order**, not merely the smallest number, is a job for the division algorithm.

Example. The supremum under divisibility is the least common multiple

Let $a, b \in \mathbb{N}$. Find $\sup\{a, b\}$ in $(\mathbb{N}, |)$. Deduce that an integer divisible by 2 and by 3 is divisible by 6.

Solution. Case $a, b \in \mathbb{Z}_{>0}$. Let $M := \{m \in \mathbb{Z}_{>0} : a \mid m \text{ and } b \mid m\}$, the positive common multiples. It is non-empty: $ab \in \mathbb{N}$ by Lemma 2.1.1 (b), $ab > 0$ as a product of positive numbers, so $ab \in \mathbb{Z}_{>0}$ (Proposition 2.1.2 (b)), and $ab = b \cdot a = a \cdot b$ is a multiple of both. Since $M \subseteq \mathbb{Z}_{>0} \subseteq \mathbb{N}$, M has a least element L (Theorem 2.1.3), the **least common multiple** of a and b .

We show that L divides **every** integer common multiple c of a and b . By Theorem 2.2.3, $c = qL + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < L$. Write $c = as$, $L = at$ with $s, t \in \mathbb{Z}$; then $r = c - qL = a(s - qt)$ with $s - qt \in \mathbb{Z}$ (Proposition 2.1.2 (a)), so $a \mid r$, and in the same way $b \mid r$. If $r > 0$, then $r \in \mathbb{Z}_{>0}$ (Proposition 2.1.2 (b)), so $r \in M$ and $r < L$, contradicting the choice of L . Hence $r = 0$, and $c = qL$: $L \mid c$. The argument never used $c \geq 0$.

Now $L = \sup\{a, b\}$. (S1): $a \mid L$ and $b \mid L$, since $L \in M$. (S2): an upper bound $c \in \mathbb{N}$ of $\{a, b\}$ is a common multiple, so $L \mid c$ by the previous paragraph.

Case $a = 0$ or $b = 0$. Say $a = 0$. An upper bound c has $0 \mid c$, that is, $c = k \cdot 0 = 0$ for some $k \in \mathbb{Z}$. And 0 is an upper bound, since $n \mid 0 = 0 \cdot n$ for every $n \in \mathbb{N}$. So 0 is the only upper bound, and $\sup\{a, b\} = 0$, the greatest element of $(\mathbb{N}, |)$.

The instance. For $a = 2$, $b = 3$: $6 \in M$, and none of $1, 2, 3, 4, 5$ is a multiple of both 2 and 3 (1, 3, 5 are odd, and 2, 4 lie strictly between consecutive multiples of 3). So $L = 6$, and every integer n with $2 \mid n$ and $3 \mid n$ has $6 \mid n$, by the second paragraph with $c = n$.

The instance proves the fact that Section 1.2 used as a school fact: an integer divisible by 2 and by 3 is divisible by 6. It also shows the two orders on $\mathbb{N}$ disagreeing: $\sup\{4, 6\}$ is 12 under divisibility and 6 under $\leq$.

A supremum that does not exist. In $(\mathbb{Q}, \leq)$, the set $\{x \in \mathbb{Q} : x > 0, x^2 < 2\}$ is non-empty and bounded above (by 2, for instance: a rational $x > 2$ has $x^2 > 4$). It has **no** least upper bound in $\mathbb{Q}$. We do not prove this here. *Analysis on the Real Line*, Chapter 1 proves it, and then takes the existence of suprema in $\mathbb{R}$ as its axiom: the supremum of *Analysis on the Real Line*, Chapter 1, defined in an ordered field, is Definition 2.6.5 for the order $\leq$ of that field; *Analysis on the Real Line*, Chapter 1 uses it in $\mathbb{Q}$ and in $\mathbb{R}$. The ambient set decides: the same set has a supremum in $(\mathbb{R}, \leq)$, namely $\sqrt{2}$.

2.6.4 Well-ordered sets

Section 2.1 proved that every non-empty subset of $\mathbb{N}$ has a least element (Theorem 2.1.3). In the language of this section, that is a property of the partially ordered set $(\mathbb{N}, \leq)$, and it has a name.

Definition 2.6.6. Well-Ordered Set

A partially ordered set $(X, \preceq)$ is **well-ordered**, and $\preceq$ is a **well-order** on X , if $\preceq$ is total and every **non-empty** subset of X has a least element.

$(\mathbb{N}, \leq)$ is well-ordered: $\leq$ is total, and [Theorem 2.1.3](#) supplies the least elements. $(\mathbb{Z}, \leq)$ is not: $\mathbb{Z}$ itself has no least element, since for every $n \in \mathbb{Z}$ also $n - 1 \in \mathbb{Z}$ ([Proposition 2.1.2](#) (a)) and $n - 1 < n$. Nor is the set of positive rationals, which has no least element ($q/2 < q$). The word “non-empty” is needed because $\emptyset$ has no least element in any order. Totality is in fact automatic: any two elements x, y form a non-empty set $\{x, y\}$, whose least element is below the other one. Upper bounds work too: in $(\mathbb{N}, \leq)$ a non-empty set bounded above has a greatest element ([Corollary 2.1.5](#)). Exercise C1 gives a well-ordered set that does not look like $\mathbb{N}$.

Remark. In a general partially ordered set, one may ask: if every **chain** (a subset on which $\preceq$ is total) has an upper bound, must there be a maximal element? For a non-empty partially ordered set the answer “yes” is Zorn’s lemma, and it cannot be proved from what we have: it is equivalent to the axiom of choice, which Chapter 3 states. A later chapter of this volume states Zorn’s lemma, and *Axiomatic Set Theory* proves it equivalent to the axiom of choice, together with the other equivalents, among them the claim that every set has a well-order.

2.6.5 Exercises

2.6.5.1 A. Check your understanding

Exercise. A1

- (a) Let $(X, \preceq)$ be a partially ordered set and $A \subseteq X$. Write “ g is a greatest element of A ” and “ m is a maximal element of A ” in symbols, and negate each.
- (b) True or false, with a witness: every finite non-empty subset of a partially ordered set has a greatest element.
- (c) Find $\sup\{4, 6\}$ in $(\mathbb{N}, |)$ and in $(\mathbb{N}, \leq)$.
- (d) On $\mathbb{N} \times \mathbb{N}$, the **lexicographic order** is $(a, b) \preceq (c, d)$ if $a < c$, or $a = c$ and $b \leq d$. Given that it is a partial order, is it total?

Solution.

- (a) g greatest: $g \in A \wedge \forall a \in A \ a \preceq g$. Its negation, by Logic and Proof, Proposition 2.5.5 and Logic and Proof, Proposition 6.2.3 (a): $g \notin A \vee \exists a \in A \ \neg(a \preceq g)$.
 m maximal: $m \in A \wedge \neg \exists a \in A \ m \prec a$. Its negation: $m \notin A \vee \exists a \in A \ m \prec a$, by Logic and Proof, Proposition 2.5.5 and Logic and Proof, Proposition 2.5.1. In words: g fails to be greatest if some element of A is **not below** it; m fails to be maximal if some element of A is **strictly above** it. The first failure can happen through incomparability; the second cannot.
- (b) False. In $(\mathcal{P}(\{1, 2\}), \subseteq)$, the set $\{\{1\}, \{2\}\}$ is finite and non-empty and has no greatest element ([Example \(Maximal elements that are not greatest\)](#) (b)). What is true is that it has a **maximal** element: listed as $\{x_1, \dots, x_n\}$, which is what “finite” will mean in Chapter 3, it has one by Exercise R4.

- (c) Under divisibility, $\sup\{4, 6\} = 12$, the least common multiple ([Example \(The supremum under divisibility is the least common multiple\)](#)): 12 is a common multiple, and none of $1, \dots, 11$ is (the multiples of 6 below 12 are 6, and $4 \nmid 6$). Under $\leq$, $\sup\{4, 6\} = 6$, the greatest element.
- (d) Yes. Let $(a, b), (c, d) \in \mathbb{N} \times \mathbb{N}$. If $a < c$, then $(a, b) \preceq (c, d)$; if $c < a$, then $(c, d) \preceq (a, b)$; if $a = c$, then $b \leq d$ or $d \leq b$, by totality of $\leq$, and the pairs are comparable. By trichotomy these cases cover everything.

2.6.5.2 B. Practice

Exercise. B1: Suprema of families of sets

Let X be a set and $(A_i)_{i \in I}$ a family of subsets of X . In $(\mathcal{P}(X), \subseteq)$, prove that the set $\{A_i : i \in I\}$ has supremum $\bigcup_{i \in I} A_i$ and infimum $\bigcap_{i \in I} A_i$, the intersection taken in X . What are they when $I = \emptyset$?

Solution. Write $U := \bigcup_{i \in I} A_i$ and $N := \bigcap_{i \in I} A_i$ ([Definition 1.2.4](#)). Both are subsets of X , so they lie in $\mathcal{P}(X)$.

Supremum. (S1) For each $i \in I$, every element of A_i lies in U , so $A_i \subseteq U$. (S2) Let $C \in \mathcal{P}(X)$ be an upper bound: $A_i \subseteq C$ for every $i \in I$. Let $x \in U$. Then $x \in A_i$ for some $i \in I$; name it i . Since $A_i \subseteq C$, $x \in C$. So $U \subseteq C$.

Infimum. N is a lower bound: for each $i \in I$, every $x \in N$ lies in A_i , so $N \subseteq A_i$. Let $D \in \mathcal{P}(X)$ be a lower bound: $D \subseteq A_i$ for every i . Let $x \in D$. Then $x \in X$, and $x \in A_i$ for every $i \in I$, so $x \in N$. Hence $D \subseteq N$.

Empty index set. If $I = \emptyset$, then $U = \emptyset$ and $N = X$, since an intersection in X of no sets is X . This agrees with the definitions: every element of $\mathcal{P}(X)$ is an upper bound of the empty set, and the least of them is $\emptyset$; every element is a lower bound, and the greatest is X . The arguments above cover this case, since their “for every i ” clauses hold vacuously.

Exercise. B2: Two maximal elements and a least one

Give an example of a partially ordered set that has exactly two maximal elements and a least element, and check each property. Then list the properties that every answer must have.

Solution. An example. $X = \{\emptyset, \{1\}, \{2\}\}$ ordered by $\subseteq$. It is a partial order, as a subset of $(\mathcal{P}(\{1, 2\}), \subseteq)$. The least element is $\emptyset$, since $\emptyset \subseteq A$ for every $A \in X$ (Section 1.1, after [Proposition 1.1.3](#)). The maximal elements are $\{1\}$ and $\{2\}$: neither is strictly contained in an element of X . And $\emptyset$ is not maximal, since $\emptyset \subsetneq \{1\}$. So there are exactly two. Another answer: $\{1, 2, 3\}$ under divisibility, with least element 1 and maximal elements 2, 3.

What every answer must have. Let l be the least element and $m_1 \neq m_2$ the maximal elements.

- m_1 and m_2 are incomparable: if $m_1 \preceq m_2$, then $m_1 \prec m_2$, since $m_1 \neq m_2$, and m_1 would not be maximal; symmetrically for $m_2 \preceq m_1$.
- So the order is not total, and there is no greatest element: a greatest element g would

have $m_1 \preceq g$, and maximality of m_1 gives $m_1 = g$; likewise $m_2 = g$, contradicting $m_1 \neq m_2$.

- l, m_1, m_2 are three different elements: if $l = m_1$, then $m_1 = l \preceq m_2$, which the first point excludes; likewise $l \neq m_2$.

So an answer has at least three elements and two incomparable maximal ones above a common least element. A candidate failing any of these is not an answer.

Exercise. B3: Minimal is least in a total order

Let $(X, \preceq)$ be a totally ordered set and $A \subseteq X$. Prove that a minimal element of A is a least element of A , by applying [Proposition 2.6.2](#) (b) to the reverse order $\succeq$. Check each fact about $\succeq$ that you use.

Solution. The reverse order. $\succeq$ is the relation $x \succeq y \iff y \preceq x$. (PO1): $x \preceq x$, so $x \succeq x$. (PO2): if $x \succeq y$ and $y \succeq x$, then $y \preceq x$ and $x \preceq y$, so $x = y$. (PO3): if $x \succeq y$ and $y \succeq z$, then $z \preceq y$ and $y \preceq x$, so $z \preceq x$, that is, $x \succeq z$. Totality: for $x, y \in X$, $x \preceq y$ or $y \preceq x$, that is, $y \succeq x$ or $x \succeq y$. Its strict part is $x \succ y \iff y \prec x$, since $x \neq y$ is symmetric in x, y .

Translating. In the order $\succeq$, “ x lies below y ” is written $x \succeq y$, and “strictly below” is $x \succ y$. Now m is minimal for $\preceq$: $m \in A$ and no $a \in A$ has $a \prec m$, that is, no $a \in A$ has $m \succ a$. That is exactly “ m is maximal for $\succeq$ ”. Likewise “ l is greatest for $\preceq$ ” says $l \in A$ and $a \succeq l$ for every $a \in A$, that is, $l \preceq a$ for every $a \in A$: “ l is least for $\preceq$ ”.

Conclusion. Let m be minimal in A for $\preceq$. Then m is maximal in A for the total order $\succeq$, so it is greatest for $\succeq$ by [Proposition 2.6.2](#) (b), hence least for $\preceq$.

2.6.5.3 C. Going deeper

Exercise. C1: A well-order that is not like $\mathbb{N}$

Let $\preceq$ be the lexicographic order on $\mathbb{N} \times \mathbb{N}$ of Exercise A1 (d).

- Prove that $\preceq$ is a partial order.
- Prove that $(\mathbb{N} \times \mathbb{N}, \preceq)$ is well-ordered.
- Show that $(1, n) \prec (2, 1)$ for every $n \in \mathbb{N}$, and that these elements $(1, n)$ are pairwise different. In $(\mathbb{N}, \preceq)$, which elements lie strictly below a given $k \in \mathbb{N}$?

Solution.

- (PO1) $a = a$ and $b \leq b$, so $(a, b) \preceq (a, b)$. (PO2) Let $(a, b) \preceq (c, d)$ and $(c, d) \preceq (a, b)$. If $a < c$, the second relation needs $c < a$ or $c = a$, both impossible; likewise $c < a$ is impossible. So $a = c$ by trichotomy, and then $b \leq d$ and $d \leq b$ give $b = d$. (PO3) Let $(a, b) \preceq (c, d) \preceq (e, f)$. If $a < c$ or $c < e$, then $a < e$, since in each case the other comparison is $\leq$; so $(a, b) \preceq (e, f)$. Otherwise $a = c = e$ and $b \leq d \leq f$, so again $(a, b) \preceq (e, f)$.
- Totality is A1 (d). Let $S \subseteq \mathbb{N} \times \mathbb{N}$ be non-empty. The set $F := \{a \in \mathbb{N} : (a, b) \in S \text{ for some } b \in \mathbb{N}\}$ of first coordinates is non-empty, so it has a least element a_0 ([The-](#)

orem 2.1.3). The set $G := \{b \in \mathbb{N} : (a_0, b) \in S\}$ is non-empty, since $a_0 \in F$, so it has a least element b_0 . Then $(a_0, b_0) \in S$. Let $(a, b) \in S$. Then $a \in F$, so $a_0 \leq a$. If $a_0 < a$, then $(a_0, b_0) \preceq (a, b)$. If $a_0 = a$, then $b \in G$, so $b_0 \leq b$, and again $(a_0, b_0) \preceq (a, b)$. So (a_0, b_0) is the least element of S .

- (c) For $n \in \mathbb{N}$, $1 < 2$ gives $(1, n) \preceq (2, 1)$, and $(1, n) \neq (2, 1)$ since $1 \neq 2$; so $(1, n) \prec (2, 1)$. For $n \neq n'$, $(1, n) \neq (1, n')$ by the characteristic property of pairs. So one element, $(2, 1)$, has an element $(1, n)$ strictly below it for each of the natural numbers n , all different. In $(\mathbb{N}, \leq)$, the elements strictly below k are exactly $0, 1, \dots, k-1$ (none when $k = 0$), by Lemma 2.1.1 (a) and (e): they form a list ending at $k-1$. So a well-ordered set need not look like $\mathbb{N}$: here one element has a strictly smaller element for every natural number.

Exercise. C2: Infima and suprema under divisibility

Work in $(\mathbb{N}, |)$. Let $a, b \in \mathbb{N}$, not both 0, and let d be the least element of

$$D := \{ka + lb : k, l \in \mathbb{Z}\} \cap \mathbb{Z}_{>0}.$$

- (a) Explain why d exists.
- (b) Prove that $d \mid a$ and $d \mid b$.
- (c) Prove that every common divisor $c \in \mathbb{N}$ of a and b divides d , and conclude that $d = \inf\{a, b\}$.
- (d) Decide whether every non-empty subset of $\mathbb{N}$ has a supremum in $(\mathbb{N}, |)$. Then decide the same question for $(\mathbb{Z}_{>0}, |)$.

Solution (Hint 1). For (b), apply the division algorithm to a and d .

Solution (Hint 2). Write $a = qd + r$ with $0 \leq r < d$, and write $d = k_0a + l_0b$.

Solution.

- (a) $D \subseteq \mathbb{Z}_{>0} \subseteq \mathbb{N}$. It is non-empty: if $a \neq 0$, then $a = 1 \cdot a + 0 \cdot b$ and $a \in \mathbb{Z}_{>0}$, so $a \in D$; otherwise $b \neq 0$ and $b \in D$ likewise. So D has a least element by Theorem 2.1.3, unique by Proposition 2.6.1.
- (b) Write $d = k_0a + l_0b$ with $k_0, l_0 \in \mathbb{Z}$. Since $d \in \mathbb{Z}_{>0}$, Theorem 2.2.3 gives $a = qd + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < d$. Then

$$r = a - qd = (1 - qk_0)a + (-ql_0)b,$$

with integer coefficients (Proposition 2.1.2 (a)). If $r > 0$, then $r \in \mathbb{Z}_{>0}$, so $r \in D$ and $r < d$, contradicting the choice of d . Hence $r = 0$, and $a = qd$: $d \mid a$. The same argument with b in place of a gives $d \mid b$.

- (c) Let $c \in \mathbb{N}$ with $c \mid a$ and $c \mid b$: $a = cs$, $b = ct$ with $s, t \in \mathbb{Z}$. Then $d = k_0cs + l_0ct = c(k_0s + l_0t)$, so $c \mid d$. In $(\mathbb{N}, |)$, the lower bounds of $\{a, b\}$ are the common divisors in $\mathbb{N}$. By (b), d is one; by (c), every lower bound c has $c \mid d$. So d is the greatest lower

bound: $d = \inf\{a, b\}$.

- (d) *In $(\mathbb{N}, |)$: yes.* Let $T \subseteq \mathbb{N}$ be non-empty. An upper bound of T is a common multiple $c \in \mathbb{N}$ of all elements of T , and 0 is one, since $t | 0 = 0 \cdot t$ for every t .

Case 1: T has a positive common multiple. The positive common multiples form a non-empty subset of $\mathbb{N}$, so they have a least element L (Theorem 2.1.3). It is an upper bound. Let $c \in \mathbb{N}$ be any upper bound. By Theorem 2.2.3, $c = qL + r$ with $0 \leq r < L$. For each $t \in T$, $t | c$ and $t | L$, so $t | r = c - qL$ (as in Example (The supremum under divisibility is the least common multiple)). If $r > 0$, then r is a positive common multiple smaller than L , which is impossible. So $r = 0$ and $L | c$. Hence $\sup T = L$. This is the argument of Example (The supremum under divisibility is the least common multiple), which never used that the set had two elements.

Case 2: T has no positive common multiple. Then 0 is its only upper bound, and $\sup T = 0$, the least element of $\{0\}$. This case includes every T containing 0, since $0 | c$ forces $c = 0$.

In $(\mathbb{Z}_{>0}, |)$: no. (It is a partial order, as the restriction of one.) Take $T = \mathbb{Z}_{>0}$. An upper bound would be some $c \in \mathbb{Z}_{>0}$ with $t | c$ for every $t \in \mathbb{Z}_{>0}$, in particular $(c + 1) | c$: $c = k(c + 1)$ with $k \in \mathbb{Z}$. Since $c > 0$ and $c + 1 > 0$, $k > 0$ (if $k \leq 0$, then $k(c + 1) \leq 0$), so $k \geq 1$ (Proposition 2.1.2 (b)), and $c = k(c + 1) \geq c + 1$, which is impossible. So T has no upper bound in $\mathbb{Z}_{>0}$, and no supremum. Removing 0 removed the element that bounded everything.

Marking scheme. (10 points)

- 1: (a) D is non-empty (exhibits a or b in it) and $D \subseteq \mathbb{N}$, so well-ordering applies.
- 3: (b) the division algorithm with divisor $d \geq 1$; the remainder is again of the form $ka + lb$; minimality of d forces $r = 0$.
- 2: (c) a common divisor divides d , and the conclusion $d = \inf\{a, b\}$ checks both clauses of an infimum.
- 3: (d) for $\mathbb{N}$: 0 is always an upper bound; with a positive common multiple, the least one divides every upper bound (division algorithm again); without one, the supremum is 0.
- 1: (d) for $\mathbb{Z}_{>0}$: a set with no upper bound, with proof.

Fatal errors (0 for the step): concluding $d | a$ from $d \leq a$; in (d), taking the numerically smallest upper bound and not checking that it **divides** every upper bound; treating $\inf$ as the minimum under $\leq$.

2.6.5.4 R. Review

Exercise. R1

Let $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2 + 4x$. Compute $f([-3, 1])$ and $f^{-1}([-3, 0])$. Then compute $f(\mathbb{R})$, and check that $f(f^{-1}([-5, 0])) = [-5, 0] \cap f(\mathbb{R})$ by computing both sides.

Solution. Complete the square: $f(x) = (x + 2)^2 - 4$. We use, as school facts, that squares are non-negative and that every $y \geq 0$ has a square root $\sqrt{y} \geq 0$, and, by Logic and Proof, Theorem 2.4.5, that for $s, t \geq 0$, $s \leq t$ exactly when $s^2 \leq t^2$.

$f([-3, 1]) = [-4, 5]$. ($\subseteq$) If $-3 \leq x \leq 1$, then $-1 \leq x + 2 \leq 3$, so $|x + 2| \leq 3$, $0 \leq (x + 2)^2 \leq 9$, and $-4 \leq f(x) \leq 5$. ($\supseteq$) For $-4 \leq y \leq 5$, take $x := -2 + \sqrt{y + 4}$. Then $0 \leq \sqrt{y + 4} \leq 3$, so

$x \in [-2, 1] \subseteq [-3, 1]$, and $f(x) = (y + 4) - 4 = y$.

$f^{-1}([-3, 0]) = [-4, -3] \cup [-1, 0]$. For $x \in \mathbb{R}$: $-3 \leq f(x) \leq 0 \iff 1 \leq (x+2)^2 \leq 4 \iff 1 \leq |x+2| \leq 2$, and splitting by the sign of $x+2$, this holds exactly when $x+2 \in [-2, -1] \cup [1, 2]$, that is, $x \in [-4, -3] \cup [-1, 0]$.

$f(\mathbb{R}) = [-4, \infty)$, by the argument for $f([-3, 1])$: $f(x) \geq -4$ always, and $y \geq -4$ is $f(-2 + \sqrt{y+4})$.

The check. $f^{-1}([-5, 0]) = \{x : (x+2)^2 \leq 4\} = [-4, 0]$, since $(x+2)^2 - 4 \geq -4 > -5$ always. Then $f([-4, 0]) = [-4, 0]$, by the first computation's method ($x+2 \in [-2, 2]$, so $(x+2)^2 \in [0, 4]$, and each value is attained). On the other side, $[-5, 0] \cap [-4, \infty) = [-4, 0]$. Both sides are $[-4, 0]$, as Proposition 1.6.5 (b) predicts: the round trip removes exactly the part $[-5, -4)$ of $[-5, 0]$ that f never reaches.

Exercise. R2

Write in symbols, with bounded quantifiers, the statement “every element of A has an element of A strictly above it”, for a set $A \subseteq \mathbb{R}$ ordered by $\leq$. Write its negation with no $\neg$ in front of a quantifier, and decide which of the two holds for $A = \mathbb{N}$ and for $A = \{1, \dots, 5\}$.

Solution. The statement is $\forall a \in A \exists b \in A \ a < b$. By Logic and Proof, Proposition 6.2.3 (a) and then (b), its negation is

$$\exists a \in A \forall b \in A \ \neg(a < b),$$

and since $\leq$ is total on $\mathbb{R}$, $\neg(a < b)$ is $b \leq a$: the negation is $\exists a \in A \forall b \in A \ b \leq a$, “ A has a greatest element”. Passing the $\neg$ through the quantifiers turns each $\forall$ into $\exists$ and each $\exists$ into $\forall$, and leaves the bounds $a \in A$, $b \in A$ untouched; negating the bound instead is entry (N11) of Logic and Proof, Example (What is not allowed).

For $A = \mathbb{N}$ the statement holds: for $a \in \mathbb{N}$, $b := a + 1 \in \mathbb{N}$ by (NS), and $a < a + 1$. For $A = \{1, \dots, 5\}$ the negation holds, with $a = 5$: every $b \in A$ has $b \leq 5$.

Exercise. R3

Call a formula **biconditional** if it is built from atoms using $\leftrightarrow$ alone: each atom is biconditional, and if φ and ψ are biconditional, so is $(\varphi \leftrightarrow \psi)$; nothing else is. Prove that every biconditional formula is true under the valuation that makes every atom true.

Solution. Let v_T be the all-T valuation. We use Logic and Proof, Theorem 4.3.1 with the property

$$P(\varphi) : \text{if } \varphi \text{ is biconditional, then } v_T(\varphi) = T.$$

The formulas $\perp$, $\neg\varphi$ and $(\varphi \circ \psi)$ with $\circ$ one of $\wedge, \vee, \rightarrow$ are not biconditional, so P holds for them vacuously, and the corresponding clauses of the theorem need nothing more.

- (i) For an atom p , $v_T(p) = T$ by the choice of v_T .
- (ii) for $\circ = \leftrightarrow$: suppose $P(\varphi)$ and $P(\psi)$, and that $(\varphi \leftrightarrow \psi)$ is biconditional. Then φ and ψ are biconditional: by unique readability (Logic and Proof, Theorem 4.3.3), they are

the only immediate parts, and a biconditional formula that is not an atom is built by the second clause from biconditional parts. So $v_T(\varphi) = T = v_T(\psi)$, and by (V6) of Logic and Proof, Definition 4.4.2, $v_T(\varphi \leftrightarrow \psi) = T$, since the two parts have the same value.

This covers every clause, and proves the claim. In particular no biconditional formula is equivalent to $\neg p$, which is false under v_T : negation cannot be expressed with $\leftrightarrow$ alone.

Exercise. R4

Let $(X, \preceq)$ be a partially ordered set, let $n \geq 1$ be an integer, and let $x_1, \dots, x_n \in X$. Show that the set $\{x_1, \dots, x_n\}$ has a maximal element.

Solution. The tool is induction from base 1 (Theorem 2.2.1 with $n_0 = 1$), in the shape of Proposition 2.3.8: add one element at a time and keep track of a maximal element. Let $P(n)$ say: for all $x_1, \dots, x_n \in X$, the set $\{x_1, \dots, x_n\}$ has a maximal element.

Base. $\{x_1\}$: no element a of it has $x_1 \prec a$, since the only candidate is x_1 and $x_1 \prec x_1$ never holds. So x_1 is maximal.

Step. Let $n \geq 1$, assume $P(n)$, and let $x_1, \dots, x_{n+1} \in X$. Put $A := \{x_1, \dots, x_n\}$ and $A' := \{x_1, \dots, x_{n+1}\}$, where, as in Section 2.3, $\{x_1, \dots, x_n\} := \{x_k : k \in \{1, \dots, n\}\}$, now for elements of X . Since $\{1, \dots, n+1\} = \{1, \dots, n\} \cup \{n+1\}$ (Lemma 2.1.1 (e)), as in the proof of Proposition 2.3.8, $A' = A \cup \{x_{n+1}\}$. By $P(n)$, A has a maximal element m .

Case 1: $m \prec x_{n+1}$ fails. Then m is maximal in A' : $m \in A'$; no $a \in A$ has $m \prec a$, by maximality in A ; and $a = x_{n+1}$ is excluded by the case.

Case 2: $m \prec x_{n+1}$. Then x_{n+1} is maximal in A' . Suppose $x_{n+1} \prec a$ with $a \in A'$. Then $a \neq x_{n+1}$, so $a \in A$; and $m \prec x_{n+1} \prec a$ gives $m \prec a$, since the strict part of a partial order is transitive. This contradicts the maximality of m in A .

In either case A' has a maximal element, so $P(n+1)$ holds, and Theorem 2.2.1 gives $P(n)$ for every $n \geq 1$.

The look-alikes. “Take the greatest element” fails, because a finite set need not have one (Example (Maximal elements that are not greatest)). “Take the largest, as in Proposition 2.3.8” fails for the same reason: that proof keeps the larger of M and x_{n+1} , which needs the two to be comparable; here Case 1 covers incomparable elements by keeping m . “Take x_n , the last one” confuses the order of the list with the order $\preceq$: nothing makes x_n large.

Chapter 3

Cardinality: Countable and Uncountable Sets

Are there more rational numbers than natural numbers? The rationals are dense, with a third between any two, while the naturals are spread out one unit apart; and yet the answer is no, because the rationals can be listed. Are there more real numbers than rationals? Yes: no list of real numbers, however cleverly it is made, contains them all. This chapter makes “more” exact, with bijections and injections in place of counting, and proves both answers. It then shows that there is no largest size at all: by Cantor’s theorem, the set of subsets of any set is strictly bigger than the set itself. A second motivating question comes from school: *why does $0.999\dots = 1$ need a definition at all?* The answer is in Section 3.4, where a decimal expansion is defined with finite sums only, and $0.999\dots = 1$ becomes true by that definition.

On the way the chapter meets a principle that the rules of logic and the working rules for sets do not supply: making infinitely many arbitrary choices at once. The obvious proof that a countable union of countable sets is countable needs a list of every set of the family at the same time, and no rule produces one. The chapter states the missing principle as the axiom of choice, marks the line of each proof that uses it, and shows how often a rule (the least element, a formula) makes it unnecessary. The finite sets come first, with the pigeonhole principle and the facts about counting that the earlier chapters took on trust. The last section is **Gate 0**, the examination on *Logic and Proof* and the first three chapters of this volume, which a reader should pass before starting *Analysis on the Real Line*.

What you need. *Logic and Proof*, Chapters 1–5, and *Sets, Relations, and Cardinality*, Chapters 1–2. Induction (Logic and Proof, Axiom 4.2.1) from *Logic and Proof*, Chapter 3. From *Logic and Proof*, Chapter 4, the rule $\exists E$ (Logic and Proof, Definition 5.6.1): Section 3.3 shows why one use of it names one object and cannot name infinitely many. From *Logic and Proof*, Chapter 5, the catalogue of what a proof may not do (Logic and Proof, Example (What is not allowed)). From Chapter 1: sets cut out by a property (Definition 1.1.3) and Russell’s theorem (Theorem 1.1.4); power sets, families and their unions (Definition 1.2.2, Definition 1.2.3, Definition 1.2.4); ordered pairs and products (Definition 1.3.1, Definition 1.3.2); functions, the empty function and equality of functions (Definition 1.4.1, Proposition 1.4.2, Proposition 1.4.1); injective, surjective and bijective maps and their composites (Definition 1.4.3, Definition 1.4.4, Theorem 1.4.4); inverses (Theorem 1.5.4, Theorem 1.5.1, Proposition 1.5.2); images and preimages (Definition 1.6.1, Definition 1.6.2). From Chapter 2: the arithmetic of $\mathbb{N}$ and $\mathbb{Z}$ (Lemma 2.1.1, Proposition 2.1.2), well-ordering (Theorem 2.1.3), induction from any base (Theorem 2.2.1), the division algorithm (Theorem 2.2.3), definition by recursion (Theorem 2.3.1), and powers and sums (Definition 2.3.2, Proposition 2.3.4, Definition 2.3.1, Proposition 2.3.6). The arithmetic and order of $\mathbb{R}$ are school

facts in examples, as in the earlier chapters. Section 3.4 states three school facts about decimal expansions, (D1)–(D3), and Sections 3.4–3.5 use them by name without proof; Section 3.3 uses one fact from school algebra, that a non-zero polynomial of degree d has at most d real roots. Gate 0 examines all of *Logic and Proof*, Chapters 1–5, and *Sets, Relations, and Cardinality*, Chapters 1–3.

Time. About 35 hours on the core track (all of A, one B and one C per section, the exit checks); about 45 hours for everything. Both include Gate 0, about 5 hours: the definition drill, the two-hour examination, and grading it against the marking schemes.

3.0.1 Roadmap

- **Equinumerous sets and finite sets.** “The same size” defined by a bijection, finite sets and their number of elements, the pigeonhole principle and the uniqueness of that number, subsets, disjoint unions and the 2^n subsets of an n -element set, and the theorem that no finite set is equinumerous with a proper subset, which Hilbert’s hotel turns into a proof that $\mathbb{N}$ is infinite.
- **Countable sets.** Countable as listable with repetitions allowed (empty, or the image of $\mathbb{N}$), the equivalent forms “an injection into $\mathbb{N}$ ” and “finite or countably infinite”, closure under subsets, images, unions of two sets and finite products, and the countability of $\mathbb{N} \times \mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$.
- **Countable unions and the axiom of choice.** The step in the proof for countable unions that no rule justifies, the axiom of choice that licenses it, the constructions that look like choice and are not, right inverses of surjections, countable unions of countable sets, and the countability of the algebraic numbers.
- **Diagonal arguments: uncountable sets and Cantor’s theorem.** The diagonal argument for 0–1 sequences, Cantor’s theorem that no map from a set onto its power set exists, decimal expansions defined with finite sums, the uncountability of $\mathbb{R}$, and the existence of transcendental numbers.
- **Comparing sizes: the Schröder–Bernstein theorem.** Comparing sizes by injections, its exchange with surjections at the price of choice, the Schröder–Bernstein theorem proved without choice, and every interval with more than one point equinumerous with $\mathbb{R}$.
- **Infinite sets.** Every infinite set contains a countably infinite subset (a stronger use of choice), so $\mathbb{N}$ is the smallest infinite size, and a set is infinite exactly when it is equinumerous with a proper subset of itself; the review exercises reach back across this volume and *Logic and Proof*.
- **Gate 0: logic and sets examination.** A definition drill and six problems, two hours, closed book, with a marking scheme for each problem; the pass line is 42 of 60 points with no core problem at 0, and a remediation map sends each lost problem back to named sections.

3.0.2 Named moves

- **List it.** To show that A is countable, give a surjection $\mathbb{N} \rightarrow A$ or an injection $A \rightarrow \mathbb{N}$, whichever is cheaper to write down and check; repetitions in a list are free. An injection only has to keep different elements apart, as $(m, n) \mapsto 2^m 3^n$ does for $\mathbb{N} \times \mathbb{N}$.
- **The diagonal.** Given any list, build an object that differs from the n -th entry at the n -th place. It is not on the list, and since the list was arbitrary, no list contains everything.
- **Injections both ways.** To show $|A| = |B|$, find an injection $A \rightarrow B$ and an injection $B \rightarrow A$, and invoke the Schröder–Bernstein theorem; do not hunt for the bijection.
- **Name the choice.** When a proof picks one object for each of infinitely many indices with no rule that names it, say at that line that the axiom of choice is used. When a rule

names the object, such as the least element of a set of natural numbers or a formula, no choice is used.

3.1 Equinumerous Sets and Finite Sets

Chapter 1 made functions exact and sorted them into injective, surjective and bijective ones. Chapter 2 put the natural numbers on firm ground: their arithmetic, well-ordering, induction from any base and definition by recursion. This chapter asks how large a set is. For a finite set we count. For $\mathbb{N}$, $\mathbb{Q}$ or $\mathbb{R}$ counting gives no answer, and the question “are there more rational numbers than natural numbers?” needs a meaning before it can have an answer. This section gives “the same size” its meaning, a pairing, and then proves the facts about finite sets on which counting rests, starting with the pigeonhole principle.

3.1.1 Same size without counting

So far we have compared functions; now we compare sets. The everyday word is “as many as”, and we can check it without knowing any numbers. Put two piles of coins on a table, and take one coin from each pile at a time, setting the two aside as a pair. If both piles run out at the same moment, they had the same number of coins. We know this before counting either pile, and a child who cannot count past ten can do it with piles of a thousand.

The pairing is a function from the first pile to the second. Different coins are paired with different coins, so it is injective; every coin of the second pile is used, so it is surjective. Pairing makes sense for any two sets, finite or not, and that is the definition.

Same size means there is a pairing.

Definition 3.1.1. Equinumerous Sets

Let A and B be sets. We say that A and B are **equinumerous**, and write $|A| = |B|$, if **there exists** a bijection $f: A \rightarrow B$. In symbols,

$$|A| = |B| \iff \exists f \in B^A \ (f \text{ is injective} \wedge f \text{ is surjective}).$$

In words: the definition asks for **one** bijection, and it does not care which. Injectivity says that no two elements of A are paired with the same element of B ; surjectivity says that no element of B is left over. The statement $|A| = |B|$ is read as a whole, “ A and B are equinumerous”. It is a relation between two sets. We do not define an object $|A|$, “the size of A ”, standing alone, and a remark below explains why the obvious candidate is not available. For finite sets, the definition of a finite set below gives $|A|$ a meaning as a natural number, and the two uses agree.

Here are four examples, the degenerate one first.

- *The empty set.* The empty function $\emptyset \rightarrow \emptyset$ of [Proposition 1.4.2](#) (a) is injective and surjective, vacuously: there are no two inputs to compare and no output to reach. So $|\emptyset| = |\emptyset|$. Conversely, if $|A| = |\emptyset|$, there is a function $A \rightarrow \emptyset$, so $A = \emptyset$ by [Proposition 1.4.2](#) (b). The empty set is equinumerous with itself and with nothing else. This case is the test case for every statement below about sets “with n elements”, at $n = 0$.
- *Three letters and three numbers.* Let a, b, c be three different letters. The function $\{a, b, c\} \rightarrow \{1, 2, 3\}$ with $a \mapsto 1$, $b \mapsto 2$, $c \mapsto 3$ sends different letters to different numbers and reaches each of 1, 2, 3. So $|\{a, b, c\}| = |\{1, 2, 3\}|$.
- *Two intervals of different lengths.* The function $f: [0, 1] \rightarrow [0, 2]$, $f(x) = 2x$, has the

inverse $y \mapsto y/2$, since $y/2 \in [0, 1]$ for $y \in [0, 2]$, $f(y/2) = y$ and $f(x)/2 = x$. By [Theorem 1.5.4](#) it is a bijection, so $|[0, 1]| = |[0, 2]|$. Length and size are different things.

- *The natural numbers and the positive integers.* The function $\mathbb{N} \rightarrow \mathbb{Z}_{>0}$, $n \mapsto n + 1$, is a bijection. It is the Quick check near the end of this section, where it shows that $\mathbb{N}$ is infinite.

A non-example by a minimal change. Shrink $\{1, 2, 3\}$ to $\{1, 2\}$. There is no bijection $\{a, b, c\} \rightarrow \{1, 2\}$. Let g be any function $\{a, b, c\} \rightarrow \{1, 2\}$. If $g(a) = g(b)$, then g is not injective. Otherwise $g(a)$ and $g(b)$ are 1 and 2 in some order, and $g(c)$, which is 1 or 2, equals one of them. In every case two different letters share a value. So the clause that fails is injectivity, and it fails for **every** function, which is what “not equinumerous” requires. Surjectivity can hold: $a \mapsto 1$, $b \mapsto 2$, $c \mapsto 2$ reaches both numbers.

Warning. To show that two sets are **not** equinumerous, one map is never enough. The map $x \mapsto x$ from $[0, 1]$ to $[0, 2]$ is not surjective, since 2 is not a value, and yet $|[0, 1]| = |[0, 2]|$ by the map $x \mapsto 2x$ above. “Not equinumerous” is a statement about **all** functions, $\forall f \in B^A \neg(f \text{ bijective})$, and a proof of it must deal with an arbitrary f , as the non-example did.

Why this definition. To count a set is to pair it with $\{1, \dots, n\}$: saying “one, two, three” while pointing at the coins is a bijection from $\{1, 2, 3\}$ onto the pile. So the definition does not replace counting; it isolates the part of counting that still makes sense when the set is infinite. Each clause has a job. Without injectivity, the surjection $\{1, 2, 3\} \rightarrow \{1\}$ would make a three-element set as large as a one-element set. Without surjectivity, the inclusion $\{1\} \rightarrow \{1, 2, 3\}$ would do the same in the other direction.

The first result says that equinumerosity behaves like an equivalence relation.

Proposition 3.1.1. Equinumerosity Is Reflexive, Symmetric and Transitive

For all sets A , B and C :

- (a) $|A| = |A|$;
- (b) if $|A| = |B|$, then $|B| = |A|$;
- (c) if $|A| = |B|$ and $|B| = |C|$, then $|A| = |C|$.

Proof. (a) The identity id_A is injective, since $\text{id}_A(x) = \text{id}_A(y)$ says $x = y$, and surjective, since every $x \in A$ is $\text{id}_A(x)$.

(b) Let $f: A \rightarrow B$ be a bijection. By [Theorem 1.5.4](#) it has a two-sided inverse $f^{-1}: B \rightarrow A$, with $f^{-1} \circ f = \text{id}_A$ and $f \circ f^{-1} = \text{id}_B$. Read with the roles exchanged, these two equations say that f is a two-sided inverse of f^{-1} ([Definition 1.5.2](#)). So f^{-1} is a bijection $B \rightarrow A$, by [Theorem 1.5.4](#) again.

(c) If $f: A \rightarrow B$ and $g: B \rightarrow C$ are bijections, then $g \circ f: A \rightarrow C$ is a bijection by [Theorem 1.4.4](#) (c). This proves the proposition. ■

Clauses (a)–(c) have the form of (Eq1)–(Eq3) of [Definition 2.5.1](#), and yet we do not call equinumerosity an equivalence relation. An equivalence relation in Chapter 2 is a relation **on a set**

X , a subset of $X \times X$. Equinumerosity relates any two sets whatever, and there is no set of all sets for it to live on (Theorem 1.1.4). That is also why the book reserves the symbol $\sim$ for equivalence relations and writes $|A| = |B|$ here.

Remark. Why not define $|A|$ as “the collection of all sets equinumerous with A ”? Because that collection is in general not a set. The next paragraph proves this for $A = \{0\}$.

Take $A = \{0\}$. Every singleton $\{x\}$ is equinumerous with $\{0\}$, by $0 \mapsto x$. If the singletons formed a set $\mathcal{S}$, its union $\bigcup \mathcal{S}$ would be a set by (W3) of Section 1.1, and every object x would be an element of it, since $x \in \{x\} \in \mathcal{S}$. That is a set of which every set is an element, which Theorem 1.1.4 rules out.

3.1.2 Finite sets

Counting is pairing with $\{1, \dots, n\}$, so “finite” is defined by such a pairing. Recall from Section 2.1 that for $n \in \mathbb{N}$,

$$\{1, \dots, n\} := \{k \in \mathbb{N} : 1 \leq k \leq n\},$$

so that $\{1, \dots, 0\} = \emptyset$, since no k has $1 \leq k \leq 0$.

Definition 3.1.2. Finite Set, Number of Elements

Let A be a set and $n \in \mathbb{N}$. We say that A **has n elements** if $|\{1, \dots, n\}| = |A|$, that is, if **there exists** a bijection $\varphi: \{1, \dots, n\} \rightarrow A$. The set A is **finite** if it has n elements for **some** $n \in \mathbb{N}$. In that case we write $|A| = n$.

In words: φ numbers the elements of A as $\varphi(1), \varphi(2), \dots, \varphi(n)$. Injectivity says that no element gets two numbers; surjectivity says that every element gets one. So a non-empty finite set is $A = \{\varphi(1), \dots, \varphi(n)\}$, a list without repetitions. By Proposition 3.1.1 (b), “ A has n elements” is the same as $|A| = |\{1, \dots, n\}|$, so the notation $|A| = n$ is consistent with the relation of Definition 3.1.1.

There is a gap in the notation $|A| = n$. It presents n as **the** number of elements, but the definition only says that **some** n works. Could a set have 3 elements and also 4? It cannot, and a corollary of the pigeonhole principle below proves it. That corollary is the well-definedness step this definition needs. Until it is proved, read $|A| = n$ as “ A has n elements”.

Examples:

- The empty set has 0 elements: $\{1, \dots, 0\} = \emptyset$, and the empty function $\emptyset \rightarrow \emptyset$ is a bijection. Conversely, a set with 0 elements is empty, since a bijection $\emptyset \rightarrow A$ is surjective and has no values.
- $\{1, \dots, n\}$ has n elements, by $\text{id}_{\{1, \dots, n\}}$.
- For $n \in \mathbb{N}$, $\{0, \dots, n\}$ has $n + 1$ elements: $k \mapsto k - 1$ maps $\{1, \dots, n + 1\}$ to $\{0, \dots, n\}$, because $k - 1 \in \mathbb{N}$ for $k \neq 0$ (Lemma 2.1.1 (c)) and $k - 1 \leq n$. It has the inverse $j \mapsto j + 1$, so it is a bijection (Theorem 1.5.4).
- The three-letter set $\{a, b, c\}$ has 3 elements, by the inverse of the bijection of the second example after Definition 3.1.1.

Definition 3.1.3. Infinite Set

A set A is **infinite** if it is not finite: **for every** $n \in \mathbb{N}$ there is **no** bijection $\{1, \dots, n\} \rightarrow A$.

To prove a set infinite we must rule out every n at once, and for every n every candidate bijection. Nothing so far does that, not even for $\mathbb{N}$. The tool is the pigeonhole principle, and the last theorem of this section turns it into a usable test.

Shrinking the codomain. One small device is used throughout the chapter. Let $f: X \rightarrow Y$ be a function, and let $C \subseteq Y$ contain every value $f(x)$. The graph of f is then a subset of $X \times C$, and with the codomain C it is a function $X \rightarrow C$ with the same values (Definition 1.4.1). We call it “ f with codomain C ”. It is injective if and only if f is, since injectivity speaks only of values. In particular, f with codomain $f(X)$ is surjective (Definition 1.6.1). So an **injective** $f: X \rightarrow Y$ is a bijection $X \rightarrow f(X)$ once its codomain is shrunk, and $|X| = |f(X)|$.

3.1.3 The pigeonhole principle

If more pigeons than holes fly into a dovecote, some hole gets two pigeons. In Chapter 1 we found one function from a three-element set to a two-element set that is not injective (Example (Deciding injective and surjective) (c)), and observed without proof that none can be. Here is the general fact, for the standard sets $\{1, \dots, m\}$ and $\{1, \dots, n\}$.

Theorem 3.1.2. Pigeonhole Principle

Let $m, n \in \mathbb{N}$ with $m > n$. Then there is **no** injection $\{1, \dots, m\} \rightarrow \{1, \dots, n\}$.

Idea. Induct on the number n of holes. In the step there are $n + 1$ holes, and we look at the last one. If no pigeon sits in hole $n + 1$, all the pigeons sit in n holes, and the hypothesis applies at once. If pigeon j sits there, there is only one, by injectivity. Move pigeon j to the hole of pigeon m , and send pigeon m away. That leaves $m - 1$ pigeons in n holes, still one to a hole, and $m - 1 > n$. The hypothesis is used at $m - 1$, not at m , so the property we induct on must say “for every m ”. *Before reading on: self-explain the construction of the injection g in Case 2: why are its values in $\{1, \dots, n\}$, and why is it injective?*

Proof. For $n \in \mathbb{N}$ let $P(n)$ be the statement

for every $m \in \mathbb{N}$ with $m > n$, there is no injection $\{1, \dots, m\} \rightarrow \{1, \dots, n\}$.

We prove $P(n)$ for every $n \in \mathbb{N}$ by Logic and Proof, Axiom 4.2.1.

Base case $P(0)$. Let $m \in \mathbb{N}$ with $m > 0$. Then $m \geq 1$ (Lemma 2.1.1 (a)), so $1 \in \{1, \dots, m\}$ and the domain is non-empty, while $\{1, \dots, 0\} = \emptyset$. By Proposition 1.4.2 (b) there is no function $\{1, \dots, m\} \rightarrow \emptyset$ at all, so there is no injection.

Inductive step. Let $n \in \mathbb{N}$ and suppose $P(n)$. Let $m \in \mathbb{N}$ with $m > n + 1$, and suppose, for a contradiction, that $f: \{1, \dots, m\} \rightarrow \{1, \dots, n + 1\}$ is injective. There are two cases.

Case 1: $n + 1$ is not a value of f . For every i , $f(i) \leq n + 1$ and $f(i) \neq n + 1$, so $f(i) < n + 1$, and $f(i) \leq n$ by Lemma 2.1.1 (e). So f with codomain $\{1, \dots, n\}$ is an injection $\{1, \dots, m\} \rightarrow \{1, \dots, n\}$, with $m > n + 1 > n$. This contradicts $P(n)$.

Case 2: $f(j) = n + 1$ for some $j \in \{1, \dots, m\}$. Since $m > n + 1 \geq 1$, we have $m \neq 0$, so $m - 1 \in \mathbb{N}$ by [Lemma 2.1.1](#) (c); and $m - 1 > n$, by subtracting 1 from $m > n + 1$. Define $\sigma: \{1, \dots, m - 1\} \rightarrow \{1, \dots, m\}$ by

$$\sigma(i) := \begin{cases} i & \text{if } i \neq j, \\ m & \text{if } i = j. \end{cases}$$

(The second line applies only when $j \leq m - 1$; if $j = m$, then $\sigma(i) = i$ for every i .) The map σ is injective: two different inputs other than j keep their different values, and if one input is j , its value is m , while the other has value $i \leq m - 1 < m$. Moreover $\sigma(i) \neq j$ for every i : either $\sigma(i) = i \neq j$, or $\sigma(i) = m$, which happens only when $j \leq m - 1 < m$.

Let $g := f \circ \sigma$. It is injective by [Theorem 1.4.4](#) (a). For every i , $\sigma(i) \neq j$, and f is injective with $f(j) = n + 1$, so $g(i) = f(\sigma(i)) \neq n + 1$; as in Case 1, $g(i) \in \{1, \dots, n\}$. So g with codomain $\{1, \dots, n\}$ is an injection $\{1, \dots, m - 1\} \rightarrow \{1, \dots, n\}$, with $m - 1 > n$. This contradicts $P(n)$, applied to the number $m - 1$.

In both cases the injection f cannot exist, so $P(n+1)$ holds. By Logic and Proof, Axiom 4.2.1, $P(n)$ holds for every $n \in \mathbb{N}$. This proves the theorem. ■

The theorem is about the standard sets only, but bijections carry it to any finite sets. Part (b) below is its mirror image for surjections.

Corollary 3.1.3. Pigeonhole for Finite Sets

Let $m, n \in \mathbb{N}$, let A have m elements, and let B have n elements.

- (a) If $m > n$, there is no injection $A \rightarrow B$.
- (b) If $m < n$, there is no surjection $A \rightarrow B$.

Idea. For (a), the bijections of [Definition 3.1.2](#) turn an injection $A \rightarrow B$ into an injection $\{1, \dots, m\} \rightarrow \{1, \dots, n\}$, which [Theorem 3.1.2](#) forbids. For (b), a surjection $g: A \rightarrow B$ has a right inverse h , and h is an injection $B \rightarrow A$, forbidden by (a). The right inverse is built by a rule: send b to the element of A with the **smallest** number among those that g sends to b .

Proof. Let $\varphi: \{1, \dots, m\} \rightarrow A$ and $\psi: \{1, \dots, n\} \rightarrow B$ be bijections ([Definition 3.1.2](#)). By [Theorem 1.5.4](#), ψ has an inverse $\psi^{-1}: B \rightarrow \{1, \dots, n\}$, which is a bijection: the equations $\psi^{-1} \circ \psi = \text{id}$ and $\psi \circ \psi^{-1} = \text{id}$ say that ψ is a two-sided inverse of ψ^{-1} ([Definition 1.5.2](#)), so [Theorem 1.5.4](#) applies.

(a) Suppose $m > n$ and $f: A \rightarrow B$ is injective. Then $\psi^{-1} \circ f \circ \varphi: \{1, \dots, m\} \rightarrow \{1, \dots, n\}$ is injective by [Theorem 1.4.4](#) (a), which contradicts [Theorem 3.1.2](#).

(b) Suppose $m < n$ and $g: A \rightarrow B$ is surjective. For $b \in B$ let

$$S_b := \{i \in \{1, \dots, m\} : g(\varphi(i)) = b\}.$$

It is non-empty: since g is onto, $b = g(a)$ for some $a \in A$, and $a = \varphi(i)$ for some i , since φ

is onto. By [Theorem 2.1.3](#), S_b has a least element. Define $h: B \rightarrow A$ by $h(b) := \varphi(\min S_b)$. Then $g(h(b)) = b$ for every $b \in B$, because $\min S_b \in S_b$. So $g \circ h = \text{id}_B$: the function g is a left inverse of h , and h is injective by [Theorem 1.5.1](#) (a). Now h is an injection from the n -element set B to the m -element set A , with $n > m$, which contradicts (a). This proves the corollary. ■

The function h in (b) is defined by a **rule**, “the least element”, applied to each b at once, so its existence needs nothing beyond specification. Section 3.3 meets surjections for which no such rule is available, and there the book adds a new principle.

Two special cases settle earlier loose ends. With $m = n + 1$, part (a) says that a set with n elements has no $n + 1$ different elements: there is no injection $\{1, \dots, n + 1\} \rightarrow A$. That is the second school fact that Logic and Proof, Exercise (C2: A sentence with no finite model, used later) took on trust. With $m = 3$, $n = 2$ in (a) and $m = 2$, $n = 3$ in (b), the corollary gives the two observations after [Example \(Deciding injective and surjective\)](#) in general.

Here is a typical use of (a). Among any three integers, two have an even sum. Let A be a set of three integers, and let $p: A \rightarrow \{0, 1\}$ send each a to its remainder r on division by 2 ([Theorem 2.2.3](#)), an integer with $0 \leq r < 2$; then $r \in \mathbb{N}$ by [Proposition 2.1.2](#) (b), and $r \leq 1$ by [Lemma 2.1.1](#) (e), so $r \in \{0, 1\}$. The set $\{0, 1\}$ has 2 elements, so by (a), p is not injective: there are $a \neq b$ in A with the same remainder r . Writing $a = 2q + r$ and $b = 2q' + r$,

$$a + b = 2(q + q' + r),$$

which is even, since $q + q' + r \in \mathbb{Z}$ ([Proposition 2.1.2](#) (a)). The pigeons are the integers and the holes are the remainders; the whole skill is choosing the holes.

The next corollary is the well-definedness promised after [Definition 3.1.2](#).

Corollary 3.1.4. The Number of Elements Is Unique

Let A be a set and $m, n \in \mathbb{N}$. If A has m elements and A has n elements, then $m = n$.

Proof. Let $\varphi: \{1, \dots, m\} \rightarrow A$ and $\psi: \{1, \dots, n\} \rightarrow A$ be bijections. If $m > n$, then $\psi^{-1} \circ \varphi: \{1, \dots, m\} \rightarrow \{1, \dots, n\}$ is injective ([Theorem 1.5.4](#), [Theorem 1.4.4](#) (a)), which contradicts [Theorem 3.1.2](#). Exchanging the roles of φ and ψ rules out $m < n$ in the same way. By trichotomy, $m = n$. ■

So a finite set has exactly one number of elements, and from now on $|A| = n$ means “**the** number of elements of A is n ”.

3.1.4 Subsets and unions of finite sets

Three facts about finite sets are used everywhere below: the sizes of disjoint sets add, a subset is no larger than the whole, and removing one point lowers the size by one. The first uses nothing else from this section, so it comes first.

Proposition 3.1.5. Disjoint Unions

Let $m, n \in \mathbb{N}$, let A have m elements, let B have n elements, and suppose $A \cap B = \emptyset$. Then $A \cup B$ has $m + n$ elements.

Idea. Number A with $1, \dots, m$ and then carry on numbering B with $m + 1, \dots, m + n$: shift the numbering of B up by m .

Proof. Let $\varphi: \{1, \dots, m\} \rightarrow A$ and $\psi: \{1, \dots, n\} \rightarrow B$ be bijections. Note $m + n \in \mathbb{N}$ by Lemma 2.1.1 (b). For $k \in \{1, \dots, m + n\}$ with $k > m$, we have $m + 1 \leq k$ by Lemma 2.1.1 (e), so $k - m \in \mathbb{N}$ with $1 \leq k - m \leq n$ (Lemma 2.1.1 (d)). So we may define $\chi: \{1, \dots, m + n\} \rightarrow A \cup B$ by

$$\chi(k) := \begin{cases} \varphi(k) & \text{if } k \leq m, \\ \psi(k - m) & \text{if } k > m. \end{cases}$$

Injective. Let $\chi(k) = \chi(k')$. If $k, k' \leq m$, then $\varphi(k) = \varphi(k')$ gives $k = k'$. If $k, k' > m$, then $\psi(k - m) = \psi(k' - m)$ gives $k - m = k' - m$, so $k = k'$. If $k \leq m < k'$, then $\chi(k) \in A$ and $\chi(k') \in B$, and $A \cap B = \emptyset$, so $\chi(k) \neq \chi(k')$: this case cannot occur, nor can its mirror image.

Surjective. Let $x \in A \cup B$. If $x \in A$, then $x = \varphi(i)$ with $1 \leq i \leq m$, and $\chi(i) = x$. If $x \in B$, then $x = \psi(j)$ with $1 \leq j \leq n$; put $k := m + j$, so $m < k \leq m + n$, and $\chi(k) = \psi(j) = x$.

So χ is a bijection, and $A \cup B$ has $m + n$ elements. ■

Now subsets. The proof is an induction on the size of the whole set, and it must be set up with care: the step peels off one element, and the obvious way to say “peel off one element” is the corollary after it, on removing a point. But that corollary is proved **from** this proposition. So the step peels off $\varphi(n + 1)$ by restricting the bijection φ directly, and cites nothing proved later.

Proposition 3.1.6. Subsets of Finite Sets

Let $n \in \mathbb{N}$, let A have n elements, and let $S \subseteq A$. Then S is finite and has k elements for some $k \leq n$. If $S \neq A$, then $k < n$.

Idea. Remove the last element $\varphi(n + 1)$ from A . Either it is outside S , and S sits inside the smaller set; or it is in S , and S is its part in the smaller set plus one point, which Proposition 3.1.5 counts.

Proof. For $n \in \mathbb{N}$ let $P(n)$ be the statement

for every set A with n elements and every $S \subseteq A$, the set S has k elements for some $k \leq n$, and $k < n$ if $S \neq A$.

We prove $P(n)$ for every n by Logic and Proof, Axiom 4.2.1.

Base case $P(0)$. A set A with 0 elements is empty, so $S \subseteq \emptyset$ gives $S = \emptyset = A$, which has $0 \leq 0$ elements; the second clause is vacuous.

Inductive step. Let $n \in \mathbb{N}$ and suppose $P(n)$. Let A have $n + 1$ elements, with a bijection $\varphi: \{1, \dots, n + 1\} \rightarrow A$, and let $S \subseteq A$. Put

$$A' := \varphi(\{1, \dots, n\}), \quad a := \varphi(n + 1).$$

The restriction of φ to $\{1, \dots, n\}$, with codomain A' , is injective and surjective, so A' has n elements. Every element of A is $\varphi(i)$ with $i \leq n$ or $i = n + 1$ ([Lemma 2.1.1](#) (e)), so $A = A' \cup \{a\}$. And $a \notin A'$: if $a = \varphi(i)$ with $i \leq n$, injectivity would give $i = n + 1$.

Case $a \notin S$. Then $S \subseteq A'$, and $P(n)$, applied to A' and S , shows that S has $k \leq n < n + 1$ elements. In this case $S \neq A$, since $a \in A \setminus S$, and indeed $k < n + 1$, as required.

Case $a \in S$. Then $S = (S \cap A') \cup \{a\}$, and the union is disjoint because $a \notin A'$. By $P(n)$, applied to A' and $S \cap A' \subseteq A'$, the set $S \cap A'$ has j elements for some $j \leq n$. The set $\{a\}$ has 1 element (by $1 \mapsto a$). By [Proposition 3.1.5](#), S has $j + 1 \leq n + 1$ elements. If $S \neq A$, then $S \cap A' \neq A'$, since otherwise $S = A' \cup \{a\} = A$; so $j < n$ by $P(n)$, and $j + 1 < n + 1$.

In both cases $P(n + 1)$ holds. By Logic and Proof, Axiom 4.2.1, $P(n)$ holds for every $n \in \mathbb{N}$. This proves the proposition. ■

Read contrapositively, with [Corollary 3.1.4](#): a subset of an n -element set that itself has n elements is the whole set. (If $S \neq A$, its number of elements is $k < n$, and by uniqueness it is not n .) That is the first school fact that Logic and Proof, Exercise (C1: True in every finite model, used later) took on trust.

Corollary 3.1.7. Removing a Point

Let $n \in \mathbb{N}$, let A have $n + 1$ elements, and let $a \in A$. Then $A \setminus \{a\}$ has n elements.

Proof. Since $a \in A$, the set $A \setminus \{a\}$ is a proper subset of A , so by [Proposition 3.1.6](#) it has k elements for some $k < n + 1$. Now $A = (A \setminus \{a\}) \cup \{a\}$ is a disjoint union, and $\{a\}$ has 1 element, so A has $k + 1$ elements by [Proposition 3.1.5](#). By [Corollary 3.1.4](#), $k + 1 = n + 1$, so $k = n$. ■

Every later induction on the number of elements uses this corollary in the same way: its step starts from an **arbitrary** set A with $n + 1$ elements, picks some $a \in A$ (there is one, since $n + 1 \geq 1$), and applies the hypothesis to $A \setminus \{a\}$, which has n elements. It never presents the set as “ X with one point added” from the outset, which would prove the claim only for sets that come in that form. Here is the first such induction. Chapter 1 promised the count of subsets of a finite set.

Proposition 3.1.8. Counting Subsets

Let $n \in \mathbb{N}$. A set with n elements has 2^n subsets: if $|A| = n$, then $|\mathcal{P}(A)| = 2^n$.

Idea. Fix one element a . Every subset either leaves a out or contains it, and adding a pairs the subsets of the first kind with those of the second. So the count doubles with each new element.

Proof. We use Logic and Proof, Axiom 4.2.1 on n . *Base case.* A set with 0 elements is $\emptyset$, and $\mathcal{P}(\emptyset) = \{\emptyset\}$ has $1 = 2^0$ element (Definition 2.3.2).

Inductive step. Let $n \in \mathbb{N}$, and suppose every set with n elements has 2^n subsets. Let A have $n + 1$ elements, pick $a \in A$, and put $X := A \setminus \{a\}$. By Corollary 3.1.7, X has n elements, so $\mathcal{P}(X)$ has 2^n elements by the hypothesis (here $2^n \in \mathbb{N}$ by Proposition 2.3.4 (d), so the count makes sense; likewise $2^{n+1} \in \mathbb{N}$). Let

$$T := \{S \cup \{a\} : S \in \mathcal{P}(X)\},$$

the image of $\mathcal{P}(X)$ under $S \mapsto S \cup \{a\}$, a function $\mathcal{P}(X) \rightarrow \mathcal{P}(A)$.

$\mathcal{P}(A)$ is the disjoint union of $\mathcal{P}(X)$ and T . Every member of T contains a , and no member of $\mathcal{P}(X)$ does, since $a \notin X$; so the union is disjoint. Both sets consist of subsets of A . Conversely, let $U \subseteq A$. If $a \notin U$, then $U \subseteq X$. If $a \in U$, then $U = (U \setminus \{a\}) \cup \{a\}$ with $U \setminus \{a\} \subseteq X$, so $U \in T$.

T has 2^n elements. The map $S \mapsto S \cup \{a\}$, with codomain T , is surjective by the definition of T . It is injective: for $S \subseteq X$ we have $a \notin S$, so $(S \cup \{a\}) \setminus \{a\} = S$, and S can be read back from $S \cup \{a\}$. So it is a bijection $\mathcal{P}(X) \rightarrow T$, and composing a bijection $\{1, \dots, 2^n\} \rightarrow \mathcal{P}(X)$ with it (Theorem 1.4.4 (c)) shows that T has 2^n elements.

By Proposition 3.1.5, $\mathcal{P}(A)$ has $2^n + 2^n = 2^n \cdot 2 = 2^{n+1}$ elements (Definition 2.3.2). By Logic and Proof, Axiom 4.2.1, the claim holds for every $n \in \mathbb{N}$. ■

For any set Y , the map $S \mapsto \mathbf{1}_S$, with inverse $f \mapsto \{y \in Y : f(y) = 1\}$, is a bijection $\mathcal{P}(Y) \rightarrow \{0, 1\}^Y$ onto the set of functions $Y \rightarrow \{0, 1\}$ (the correspondence of Section 1.4 between subsets and indicator functions). Taking Y with n elements, a set with n elements also has exactly 2^n functions to $\{0, 1\}$, and the two counting problems of Chapter 1 are one. Exercise C2 counts all functions between finite sets.

3.1.5 No finite set is equinumerous with a proper subset

For finite sets, “a proper part is smaller” is a theorem.

Theorem 3.1.9. A Finite Set Has No Proper Copy

Let A be a finite set and $B \subsetneq A$. Then A and B are not equinumerous: there is no bijection $A \rightarrow B$.

Proof. Let A have n elements, with a bijection $\varphi: \{1, \dots, n\} \rightarrow A$. By Proposition 3.1.6, B has k elements for some $k < n$, since $B \neq A$. Suppose $g: A \rightarrow B$ were a bijection. Then

$g \circ \varphi: \{1, \dots, n\} \rightarrow B$ is a bijection ([Theorem 1.4.4 \(c\)](#)), so B also has n elements, and [Corollary 3.1.4](#) gives $k = n$. This contradicts $k < n$. ■

Turned around, the theorem is a test for infinity: **a set that is equinumerous with a proper subset of itself is infinite**. Here is its first use. For the even natural numbers we write

$$2\mathbb{N} := \{2n : n \in \mathbb{N}\},$$

the image of $\mathbb{N}$ under $n \mapsto 2n$.

Example. Hilbert's Hotel

A hotel has one room for every natural number, and every room is taken. A coach arrives with guests for the odd-numbered rooms: the guest in room n is asked to move to room $2n$. Show that this is a bijection $\mathbb{N} \rightarrow 2\mathbb{N}$, that $2\mathbb{N}$ is a proper subset of $\mathbb{N}$, and deduce that $\mathbb{N}$ is infinite.

Solution. Let $d: \mathbb{N} \rightarrow 2\mathbb{N}$, $d(n) := 2n$. Its values lie in $2\mathbb{N}$ by definition, and $2\mathbb{N} \subseteq \mathbb{N}$ by [Lemma 2.1.1 \(b\)](#). It is injective, since $2n = 2n'$ gives $n = n'$, and surjective, since every element of $2\mathbb{N}$ is $2n = d(n)$ for some $n \in \mathbb{N}$. So $|\mathbb{N}| = |2\mathbb{N}|$.

The inclusion is proper because $1 \notin 2\mathbb{N}$: if $n = 0$, then $2n = 0$; if $n \neq 0$, then $n \geq 1$ ([Lemma 2.1.1 \(a\)](#)) and $2n \geq 2$. So $2n \neq 1$ for every $n \in \mathbb{N}$.

If $\mathbb{N}$ were finite, [Theorem 3.1.9](#) would forbid a bijection from $\mathbb{N}$ onto its proper subset $2\mathbb{N}$. So $\mathbb{N}$ is infinite. Every guest still has a room, and the odd rooms are now free.

The finite intuition “a proper part is smaller” fails for $\mathbb{N}$, and it fails exactly because $\mathbb{N}$ is infinite. Section 3.6 proves that it fails for **every** infinite set.

Warning. $A \subsetneq B$ does not make A smaller than B when B is infinite: $2\mathbb{N} \subsetneq \mathbb{N}$ and $|2\mathbb{N}| = |\mathbb{N}|$. Equinumerous is not “one is a subset of the other”, in either direction.

Warning. “Finite” does not mean “bounded”. The set $E := \{1/n : n \in \mathbb{Z}_{>0}\}$ lies in $[0, 1]$, and it is infinite: $x \mapsto x/(1+x)$ sends $1/n$ to $1/(n+1)$, and it is a bijection from E onto $\{1/n : n \geq 2\}$, with inverse $y \mapsto y/(1-y)$. That set is a proper subset of E , since $1 \in E$ and $1/n \leq 1/2$ for $n \geq 2$.

Quick check. Is $n \mapsto n+1$ a bijection $\mathbb{N} \rightarrow \mathbb{N} \setminus \{0\}$? What does that show?

Solution. Yes. Its values lie in $\mathbb{N} \setminus \{0\}$, since $n+1 \geq 1$. Its inverse is $m \mapsto m-1$, which is defined on $\mathbb{N} \setminus \{0\}$ because $m-1 \in \mathbb{N}$ for $m \neq 0$ ([Lemma 2.1.1 \(c\)](#)); the two maps undo each other, so it is a bijection by [Theorem 1.5.4](#). So $\mathbb{N}$ is equinumerous with its proper subset

$\mathbb{N} \setminus \{0\} = \mathbb{Z}_{>0}$, and by [Theorem 3.1.9](#), $\mathbb{N}$ is infinite, a second time. In the hotel, every guest moves up one room, and room 0 is free.

Remark. By [Definition 3.1.2](#), a non-empty finite set A with n elements is $\{x_1, \dots, x_n\}$, with $x_k := \varphi(k)$. So when its elements are real numbers it has a maximum and a minimum, by [Proposition 2.3.8](#). Section 2.3 announced this passage from a finite set to a list.

3.1.6 Exercises

3.1.6.1 A. Check your understanding

Exercise. A1

- Define “ A is finite”, and write its negation in symbols, with every quantifier shown.
- True or false, with a reason: a subset of an infinite set is infinite; a set that has an infinite subset is infinite.
- How many injections $\{1, 2, 3\} \rightarrow \{1, 2\}$ are there?
- Let $f: A \rightarrow B$, where A and B are finite and A has more elements than B . What does the pigeonhole principle guarantee about f ?

Solution.

- A is finite if there are $n \in \mathbb{N}$ and a bijection $\varphi: \{1, \dots, n\} \rightarrow A$ ([Definition 3.1.2](#)): $\exists n \in \mathbb{N} \exists \varphi \in A^{\{1, \dots, n\}}$ (φ injective $\wedge$ φ surjective). By Logic and Proof, [Proposition 6.2.3](#), applied to each quantifier, and De Morgan, its negation is

$$\forall n \in \mathbb{N} \forall \varphi \in A^{\{1, \dots, n\}} (\varphi \text{ not injective} \vee \varphi \text{ not surjective}),$$

which is [Definition 3.1.3](#).

- The first is false: $\emptyset \subseteq \mathbb{N}$, and $\emptyset$ is finite while $\mathbb{N}$ is infinite ([Example \(Hilbert’s Hotel\)](#)). The second is true: if B were finite, every subset of B would be finite by [Proposition 3.1.6](#); so a set with an infinite subset is not finite.
- None. By [Theorem 3.1.2](#) with $m = 3 > 2 = n$, there is no injection $\{1, 2, 3\} \rightarrow \{1, 2\}$.
- That f is not injective ([Corollary 3.1.3](#) (a)): there are two **different** elements $a \neq a'$ of A with $f(a) = f(a')$. It does not say which two, and it says nothing when A has at most as many elements as B .

3.1.6.2 B. Practice

Exercise. B1: Moving the guests

Give explicit bijections $\mathbb{N} \rightarrow \mathbb{N} \setminus \{0, 1, 2\}$, $\mathbb{N} \rightarrow \{n \in \mathbb{N} : n \geq 5\}$ and $\mathbb{N} \rightarrow \{2n + 1 : n \in \mathbb{N}\}$, and prove that each is a bijection by writing down its inverse.

Solution. First. For $m \in \mathbb{N}$: if $m \notin \{0, 1, 2\}$, then $m \geq 3$. Indeed $m \neq 0$ gives $m \geq 1$ (Lemma 2.1.1 (a)); with $m \neq 1$, $m > 1$, so $m \geq 2$ (Lemma 2.1.1 (e)); with $m \neq 2$, $m \geq 3$ likewise. Conversely every $m \geq 3$ lies outside $\{0, 1, 2\}$. So $\mathbb{N} \setminus \{0, 1, 2\} = \{m \in \mathbb{N} : m \geq 3\}$. Let $f(n) := n + 3$; then $f(n) \in \mathbb{N}$ and $f(n) \geq 3$. Let $g(m) := m - 3$ for $m \geq 3$; then $g(m) \in \mathbb{N}$ by Lemma 2.1.1 (d). Since $g(f(n)) = n$ and $f(g(m)) = m$, the map f is a bijection by Theorem 1.5.4.

Second. The same with 5 in place of 3: $n \mapsto n + 5$, with inverse $m \mapsto m - 5$, which lies in $\mathbb{N}$ for $m \geq 5$ by Lemma 2.1.1 (d).

Third. Let $O := \{2n + 1 : n \in \mathbb{N}\}$, the odd natural numbers, and $f(n) := 2n + 1$, which lies in O by definition. Every $m \in O$ is $2n + 1$ for exactly one $n \in \mathbb{N}$: one by the definition of O , and only one since $2n + 1 = 2n' + 1$ gives $n = n'$. So $g(m) := (m - 1)/2$ is that n , a natural number. Then $g(f(n)) = n$ and $f(g(m)) = m$, so f is a bijection $\mathbb{N} \rightarrow O$. Each of the three sets is a proper subset of $\mathbb{N}$ equinumerous with $\mathbb{N}$, which shows again that $\mathbb{N}$ is infinite.

Exercise. B2: Choosing the holes

- (a) Prove that among any 6 different integers, two have the same remainder on division by 5.
- (b) Prove that among any 5 different elements of $\{1, \dots, 8\}$, two have sum 9.

Hint: in (b), group $\{1, \dots, 8\}$ into pairs.

Solution.

- (a) Let A be a set of 6 integers. By Theorem 2.2.3 with divisor 5, each $a \in A$ has a unique remainder $r(a)$, an integer with $0 \leq r(a) < 5$; then $r(a) \in \mathbb{N}$ by Proposition 2.1.2 (b), and $r(a) \leq 4$ by discreteness, Proposition 2.1.2 (c), so $r(a) \in \{0, \dots, 4\}$. This defines $r: A \rightarrow \{0, \dots, 4\}$. The set $\{0, \dots, 4\}$ has 5 elements (the third example after Definition 3.1.2). Since $6 > 5$, Corollary 3.1.3 (a) shows that r is not injective: there are $a \neq a'$ in A with $r(a) = r(a')$.
- (b) Let $A \subseteq \{1, \dots, 8\}$ have 5 elements. The holes are the four pairs $\{1, 8\}$, $\{2, 7\}$, $\{3, 6\}$, $\{4, 5\}$, each of the form $\{x, 9 - x\}$. Define $p: A \rightarrow \{1, 2, 3, 4\}$ by $p(x) := \min\{x, 9 - x\}$, the smaller element of the pair containing x ; for $1 \leq x \leq 8$ this lies in $\{1, 2, 3, 4\}$, since $x \leq 4$ or $9 - x \leq 4$. The codomain has 4 elements and $5 > 4$, so by Corollary 3.1.3 (a) there are $x \neq y$ in A with $p(x) = p(y)$. Then x and y lie in the same pair $\{c, 9 - c\}$, where $c = p(x)$, and being different they are its two elements, so $x + y = 9$. Five is the least number that works: $\{1, 2, 3, 4\}$ has 4 elements, and no two of them sum to 9, since their largest sum is $3 + 4 = 7$.

Exercise. B3: Finite unions

Let $n \in \mathbb{N}$, and let $A_1, \dots, A_n$ be finite sets, that is, a family $(A_i)_{i \in \{1, \dots, n\}}$ of finite sets. Prove that $\bigcup_{i \in \{1, \dots, n\}} A_i$ is finite.

Solution. For $n \in \mathbb{N}$ let $P(n)$ be “for every family $(A_i)_{i \in \{1, \dots, n\}}$ of finite sets, the union $U_n := \bigcup_{i \in \{1, \dots, n\}} A_i$ is finite”. We use Logic and Proof, Axiom 4.2.1.

Base case. For $n = 0$ the index set is $\{1, \dots, 0\} = \emptyset$, and the union of the empty family is $\emptyset$ (Definition 1.2.4: no x lies in some A_i), which has 0 elements.

Step. Suppose $P(n)$, and let $(A_i)_{i \in \{1, \dots, n+1\}}$ be a family of finite sets. Since $\{1, \dots, n+1\} = \{1, \dots, n\} \cup \{n+1\}$, we have $U_{n+1} = U_n \cup A_{n+1}$, and U_n is finite by $P(n)$, applied to the first n sets; say it has p elements. Write

$$U_{n+1} = U_n \cup (A_{n+1} \setminus U_n),$$

a disjoint union. The set $A_{n+1} \setminus U_n$ is a subset of the finite set A_{n+1} , so it is finite by Proposition 3.1.6; say it has q elements. By Proposition 3.1.5, U_{n+1} has $p + q$ elements. By Logic and Proof, Axiom 4.2.1, $P(n)$ holds for every n .

The detour through $A_{n+1} \setminus U_n$ is needed because Proposition 3.1.5 is about **disjoint** sets, and U_n and A_{n+1} may overlap.

3.1.6.3 C. Going deeper

Exercise. C1: Infinite means arbitrarily large finite pieces

Prove that a set A is infinite if and only if for every $n \in \mathbb{N}$ there is an injection $\{1, \dots, n\} \rightarrow A$. In the direction $(\Rightarrow)$, say whether your proof makes infinitely many choices at once, and why.

Solution (Hint 1). For $(\Leftarrow)$, suppose A has m elements and use the hypothesis at one well-chosen n . For $(\Rightarrow)$, induct on n .

Solution (Hint 2). For $(\Rightarrow)$: let $P(n)$ be “there is an injection $\{1, \dots, n\} \rightarrow A$ ”; the base case is the empty function.

Solution. $(\Leftarrow)$ Suppose the condition holds and, for a contradiction, that A has m elements, with a bijection $\varphi: \{1, \dots, m\} \rightarrow A$. Take $n := m+1$ and an injection $f: \{1, \dots, m+1\} \rightarrow A$. By Theorem 1.5.4, φ has an inverse $\varphi^{-1}: A \rightarrow \{1, \dots, m\}$, which is a bijection, since φ is a two-sided inverse of it (Definition 1.5.2, Theorem 1.5.4). Then $\varphi^{-1} \circ f: \{1, \dots, m+1\} \rightarrow \{1, \dots, m\}$ is injective (Theorem 1.4.4 (a)), which contradicts Theorem 3.1.2. So A is infinite.

$(\Rightarrow)$ Let A be infinite. For $n \in \mathbb{N}$ let $P(n)$ be “there is an injection $\{1, \dots, n\} \rightarrow A$ ”. We use Logic and Proof, Axiom 4.2.1.

Base case. $\{1, \dots, 0\} = \emptyset$, and the empty function $\emptyset \rightarrow A$ (Proposition 1.4.2 (a)) is injective, vacuously.

Step. Suppose $P(n)$, and let $f: \{1, \dots, n\} \rightarrow A$ be an injection. With its codomain shrunk to its image $F := f(\{1, \dots, n\})$, it is a bijection, so F has n elements. Since A is infinite, $F \neq A$; and $F \subseteq A$, so there is some $a \in A \setminus F$. Define $g: \{1, \dots, n+1\} \rightarrow A$ by $g(k) := f(k)$ for $k \leq n$ and $g(n+1) := a$; this is a function because $\{1, \dots, n+1\} = \{1, \dots, n\} \cup \{n+1\}$

and $n + 1 \notin \{1, \dots, n\}$ (Lemma 2.1.1 (e)). It is injective: two inputs $\leq n$ are separated by f , and $g(n + 1) = a \notin F$ differs from every $g(k)$ with $k \leq n$. So $P(n + 1)$ holds. By Logic and Proof, Axiom 4.2.1, $P(n)$ holds for every n .

No infinitely many choices are made. The step uses two existence statements, $P(n)$ and “ $A \setminus F \neq \emptyset$ ”, and names one witness for each: one $\exists E$ each, inside the proof of the single implication $P(n) \Rightarrow P(n + 1)$. Its conclusion, $P(n + 1)$, is again an existence statement, so the witnesses never leave the step. Induction then gives $P(n)$ for each n **separately**: for each n **some** injection exists, and different n may use unrelated ones. Gluing the steps into one injection $\mathbb{N} \rightarrow A$, with a new point picked at every stage, would require all the picks at once, with no rule saying which point to pick. That is a different statement; Section 3.6 proves it with the principle of choice of Section 3.3.

Marking scheme. (10 points) - 2: ($\Leftarrow$) supposes A has m elements and uses the hypothesis at $n = m + 1$ to contradict the pigeonhole principle (Theorem 3.1.2 or Corollary 3.1.3 (a)). - 1: ($\Rightarrow$) sets up an induction on n with the base case $n = 0$ (the empty function); a base $n = 1$ earns the point only if $n = 0$ is also covered (by the empty function). - 3: in the step, shows that the image of the given injection has n elements and is therefore not A , and takes a point outside it. - 2: extends the injection by that point and checks injectivity, including the new value against the old ones. - 2: explains that each step makes finitely many choices inside one implication, so no principle of choice is used, and that building one injection on all of $\mathbb{N}$ would be a different statement. Fatal errors (0 for the step): picking “a new point for every n ” to define one function $\mathbb{N} \rightarrow A$ and calling it induction; taking a point outside the image without showing the image is not all of A .

Exercise. C2: Counting products and functions

Let $m, n \in \mathbb{N}$, let A have m elements, and let B have n elements.

- Prove that $A \times B$ has mn elements.
- Prove that the set B^A of functions $A \rightarrow B$ has n^m elements.
- Check your formula in (b) against Proposition 1.4.2 in the cases $m = 0$ and $m \geq 1$, $n = 0$.

Solution (Hint 1). Induct on m , with B fixed, and remove one point a from A by Corollary 3.1.7.

Solution (Hint 2). For (b), a function on A is the same as a function on $A \setminus \{a\}$ together with one value at a : consider $f \mapsto (f|_{A \setminus \{a\}}, f(a))$.

Solution.

- Fix B with n elements. For $m \in \mathbb{N}$ let $P(m)$ be “for every set A with m elements, $A \times B$ has mn elements” ($mn \in \mathbb{N}$ by Lemma 2.1.1 (b)). *Base case.* If A has 0 elements, then $A = \emptyset$, and $A \times B = \emptyset$, since a pair in it would need a first coordinate in $\emptyset$. It has $0 = 0 \cdot n$ elements. *Step.* Suppose $P(m)$, let A have $m + 1$ elements, pick $a \in A$, and

put $A' := A \setminus \{a\}$, which has m elements by [Corollary 3.1.7](#). Then

$$A \times B = (A' \times B) \cup (\{a\} \times B),$$

and the union is disjoint, since first coordinates in A' differ from a ([Definition 1.3.1](#)). The first set has mn elements by $P(m)$. The map $B \rightarrow \{a\} \times B$, $b \mapsto (a, b)$, is injective by [Definition 1.3.1](#) and surjective by [Definition 1.3.2](#), so $\{a\} \times B$ has n elements. By [Proposition 3.1.5](#), $A \times B$ has $mn + n = (m + 1)n$ elements. By Logic and Proof, Axiom 4.2.1, $P(m)$ holds for every m , and B was arbitrary.

- (b) Fix B with n elements, and let $Q(m)$ be “for every set A with m elements, B^A has n^m elements” ($n^m \in \mathbb{N}$ by [Proposition 2.3.4](#) (d)). *Base case.* $A = \emptyset$, and by [Proposition 1.4.2](#) (a), $B^\emptyset$ has exactly one element, the empty function; $n^0 = 1$ ([Definition 2.3.2](#)), also when $n = 0$. *Step.* Suppose $Q(m)$, let A have $m + 1$ elements, pick $a \in A$, and let $A' := A \setminus \{a\}$, with m elements. Define

$$R: B^A \rightarrow B^{A'} \times B, \quad R(f) := (f|_{A'}, f(a)).$$

Injective: if $R(f) = R(f')$, then f and f' agree on A' and at a , hence on $A = A' \cup \{a\}$, and they have the same domain and codomain, so $f = f'$ by [Proposition 1.4.1](#).

Surjective: given $(h, b) \in B^{A'} \times B$, let f have graph $G \cup \{(a, b)\}$, where G is the graph of h , and codomain B . Each $x \in A'$ has exactly one partner, its h -value, because (a, b) has first coordinate $a \notin A'$; and a has exactly one, b , because every pair of G has first coordinate in A' . So $f \in B^A$, and $R(f) = (h, b)$. Hence R is a bijection. By $Q(m)$, $B^{A'}$ has n^m elements, so by (a), applied to the sets $B^{A'}$ and B , the product $B^{A'} \times B$ has $n^m \cdot n = n^{m+1}$ elements ([Definition 2.3.2](#)). Since R is a bijection, $|B^A| = |B^{A'} \times B|$ ([Definition 3.1.1](#)), and $|\{1, \dots, n^{m+1}\}| = |B^{A'} \times B|$. By [Proposition 3.1.1](#) (b) and (c), $|\{1, \dots, n^{m+1}\}| = |B^A|$, that is, B^A has n^{m+1} elements. By Logic and Proof, Axiom 4.2.1, $Q(m)$ holds for every m .

- (c) For $m = 0$: $B^\emptyset$ has one element by [Proposition 1.4.2](#) (a), and $n^0 = 1$, even for $n = 0$. For $m \geq 1$ and $n = 0$: $A \neq \emptyset$, so there is no function $A \rightarrow \emptyset$ by [Proposition 1.4.2](#) (b), and $B^A = \emptyset$ has 0 elements; and $0^m = 0^{m-1} \cdot 0 = 0$. Both agree with (b). With $B = \{0, 1\}$, (b) recovers the count 2^m of the paragraph after [Proposition 3.1.8](#).

Marking scheme. (10 points) - 2: sets up each induction on m with B fixed, and handles $m = 0$ (the empty product; the empty function). - 2: in each step, takes an **arbitrary** A with $m + 1$ elements, removes one point by [Corollary 3.1.7](#), and applies the hypothesis to the rest. - 2: (a) splits $A \times B$ into two disjoint pieces and counts $\{a\} \times B$ by a bijection with B . - 3: (b) gives a bijection between B^A and $B^{A'} \times B$, checking injectivity (equality of functions) and surjectivity (the extended graph is a function). - 1: (c) checks both degenerate cases against [Proposition 1.4.2](#). - Direct route for (a): composing the given bijections with a bijection $\{1, \dots, m\} \times \{1, \dots, n\} \rightarrow \{1, \dots, mn\}$, e.g. $(i, j) \mapsto (i-1)n + j$, with injectivity and surjectivity checked (for example by [Theorem 2.2.3](#)) and the case $m = 0$ or $n = 0$ handled, earns the 2 points of the (a) line and the (a) halves (1 point each) of the two induction lines. Fatal errors (0 for the step): “each of the m elements has n possible values, so n^m ” with no bijection or induction; a step that applies the hypothesis to a set not shown to have m elements.

3.2 Countable Sets

Section 3.1 made “the same size” exact and settled the finite sets. It also showed that $\mathbb{N}$ is infinite, because it is equinumerous with its proper subset $2\mathbb{N}$ (Example (Hilbert’s Hotel)). The next question is which sets can be put into a list indexed by $\mathbb{N}$. *Analysis on the Real Line*, Chapter 1 already uses the word for these sets, **countable**, in its proof that the real numbers cannot be listed. This section defines it, proves three equivalent forms of it, shows that countable sets are closed under subsets, images, unions of two sets and finite products, and lists the integers and the rationals.

3.2.1 Lists

Can we list the rational numbers as $q_0, q_1, q_2, \dots$ so that every rational number appears somewhere? Density makes it look impossible. Between any two rationals $p < q$ lies a third, $(p+q)/2$, so no rational has a next one, and a list in increasing order cannot even get from q_0 to q_1 . And yet a list need not be in increasing order. It only has to reach every element at some finite position. The rationals can be listed in that sense, and so can many sets that look larger than $\mathbb{N}$.

A list indexed by $\mathbb{N}$ is a function on $\mathbb{N}$, its n -th entry being the value at n . Reaching every element is surjectivity.

Countable means listable, repetitions allowed.

Definition 3.2.1. Countable Set

A set A is **countable** if $A = \emptyset$ or **there exists** a surjection $f: \mathbb{N} \rightarrow A$. In the second case

$$A = f(\mathbb{N}) = \{f(0), f(1), f(2), \dots\},$$

that is, the elements of A can be listed as $x_0, x_1, x_2, \dots$, with $x_n := f(n)$, repetitions allowed. A set that is not countable is **uncountable**. In symbols, A is countable if and only if

$$A = \emptyset \vee \exists f \in A^{\mathbb{N}} \quad \forall a \in A \quad \exists n \in \mathbb{N} \quad f(n) = a.$$

In words: a non-empty set is countable when some list $x_0, x_1, x_2, \dots$ reaches every one of its elements, each at a **finite** position $n \in \mathbb{N}$. The list may repeat an element as often as it likes, and it need not follow any order. The first clause is there because there is no function at all from $\mathbb{N}$ to $\emptyset$ (Proposition 1.4.2 (b)), so the second clause can never hold for the empty set; without the first clause, $\emptyset$ would be uncountable. For a non-empty set the definition says that it is the image of some function from $\mathbb{N}$, which is how *Analysis on the Real Line*, Chapter 1 states it.

Examples, starting from the smallest:

- $\emptyset$ is countable, by the first clause.
- A non-empty finite set is countable. For three different objects a, b, c , the list $a, b, c, c, c, \dots$, that is $f(0) := a$, $f(1) := b$ and $f(n) := c$ for $n \geq 2$, is a surjection $\mathbb{N} \rightarrow \{a, b, c\}$. Repetitions pad a short list out to the length of $\mathbb{N}$.
- $\mathbb{N}$ is countable: $\text{id}_{\mathbb{N}}$ is a surjection.
- The set of squares $\{n^2 : n \in \mathbb{N}\}$ is countable: it is the image of $\mathbb{N}$ under $n \mapsto n^2$, and a function is a surjection onto its image (Section 3.1, “Shrinking the codomain”). In the same way $2\mathbb{N}$ is countable.

A non-example by a minimal change. The list $n \mapsto 2n$ is a list of $2\mathbb{N}$. Read it as a list of $\mathbb{N}$, keeping the rule and changing only the target set, and it fails: the clause “for every $a \in \mathbb{N}$ there is an n with $2n = a$ ” fails at $a = 1$. This shows that **this** list misses 1, not that $\mathbb{N}$ is uncountable; one map proves nothing about all maps (the warning after the non-example in Section 3.1). Sets that are genuinely uncountable appear in Section 3.4.

Why repetitions are allowed. Two reasons. First, finite sets are then countable with no special case: pad the list by repeating an element, as above. Second, the closure proofs below become short. A function applied to a list gives a list of the image, which may have repetitions even if the first list had none; interleaving lists of two overlapping sets repeats their common elements. If repetitions were forbidden, every such proof would need a step that deletes them. The definition puts that work in one place, the theorem on three ways to say countable below, and does it once.

Sometimes we need the stronger statement that a set has exactly the size of $\mathbb{N}$.

Definition 3.2.2. Countably Infinite Set

A set A is **countably infinite** if $|A| = |\mathbb{N}|$, that is, if there is a bijection $A \rightarrow \mathbb{N}$ (equivalently, by Proposition 3.1.1 (b), a bijection $\mathbb{N} \rightarrow A$).

So $\mathbb{N}$, $2\mathbb{N}$ and $\mathbb{Z}_{>0}$ are countably infinite (Section 3.1), and a countably infinite set is countable, since a bijection is a surjection.

Warning. In this book “countable” **includes** the finite sets and $\emptyset$. Some books use “countable” to mean “countably infinite” and say “at most countable” for our “countable”. Before quoting a theorem about countable sets from elsewhere, check which convention it uses. Here, $\{0, 1\}$ is countable and is not countably infinite.

3.2.2 Subsets of the natural numbers

Every countable set will turn out to be finite or countably infinite, with nothing in between. The heart of that is a fact about subsets of $\mathbb{N}$: an infinite one can be listed in increasing order without gaps.

Lemma 3.2.1. Subsets of the Natural Numbers

Every subset $S \subseteq \mathbb{N}$ is finite or countably infinite.

Idea. If S is infinite, list it in increasing order: s_0 is the least element of S , and s_{n+1} is the least element of S above s_n . Well-ordering supplies each least element, provided the set is non-empty, and it is non-empty because S has elements beyond any bound, being infinite. Recursion turns the rule into a function. The list misses nothing: if $t \in S$ were skipped, it would lie strictly between two consecutive entries $s_{n-1} < t < s_n$, and then s_n would not be the least element of S above s_{n-1} .

Proof. If S is finite there is nothing to prove, so let S be infinite. In particular $S \neq \emptyset$.

Every $x \in \mathbb{N}$ has an element of S above it. Let $x \in \mathbb{N}$, and suppose that no $s \in S$ has $s > x$. Then $S \subseteq \{0, \dots, x\}$, a set with $x + 1$ elements (the third example after Definition 3.1.2), so S is finite by Proposition 3.1.6, a contradiction. Hence $T_x := \{s \in S : s > x\}$ is non-empty, and by Theorem 2.1.3 it has a least element.

The list. Define $g: \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ by $g(n, x) := \min T_x$; by the previous paragraph this is defined for every $x \in \mathbb{N}$. By Theorem 2.3.1, with $X := \mathbb{N}$ and $a := \min S$ (which exists by Theorem 2.1.3), there is a function $n \mapsto s_n$ from $\mathbb{N}$ to $\mathbb{N}$ with

$$s_0 = \min S, \quad s_{n+1} = \min\{s \in S : s > s_n\} \quad (n \in \mathbb{N}).$$

Every s_n lies in S : $s_0 \in S$, and $s_{n+1} \in T_{s_n} \subseteq S$.

Increasing. By construction $s_{n+1} > s_n$ for every n . Hence $s_m < s_n$ whenever $m < n$, by induction on n : there is no $m < 0$ (Lemma 2.1.1 (a)), and if $m < n + 1$, then $m \leq n$ (Lemma 2.1.1 (e)), so $s_m \leq s_n < s_{n+1}$, using the hypothesis when $m < n$. In particular $n \mapsto s_n$ is injective, since $m \neq n$ gives $m < n$ or $n < m$. Also $s_n \geq n$ for every n : $s_0 \geq 0$ (Lemma 2.1.1 (a)), and if $s_n \geq n$, then $s_{n+1} > s_n \geq n$, so $s_{n+1} \geq n + 1$ (Lemma 2.1.1 (e)).

Onto S . Let $t \in S$. The set $\{n \in \mathbb{N} : s_n \geq t\}$ is non-empty, since $s_t \geq t$, so it has a least element n by Theorem 2.1.3. If $n = 0$, then $s_0 = \min S \leq t$ and $s_0 \geq t$, so $s_0 = t$. If $n \neq 0$, then $n - 1 \in \mathbb{N}$ (Lemma 2.1.1 (c)), and $s_{n-1} < t$ by the minimality of n . So $t \in T_{s_{n-1}}$, and $s_n = \min T_{s_{n-1}} \leq t$. With $s_n \geq t$ this gives $s_n = t$.

So $n \mapsto s_n$, with codomain S , is a bijection $\mathbb{N} \rightarrow S$, and S is countably infinite. ■

3.2.3 Three ways to say countable

Being countable (empty or listable by $\mathbb{N}$), having an injection into $\mathbb{N}$, and being finite or the size of $\mathbb{N}$ all describe the same sets. Having the three forms side by side is what makes countability easy to prove: we may use whichever is cheapest for the set in hand.

Theorem 3.2.2. Characterizations of Countable Sets

For every set A , the following are equivalent.

- (i) A is countable.
- (ii) There is an injection $A \rightarrow \mathbb{N}$.
- (iii) A is finite or countably infinite.

Idea. From a list to an injection: send each element to the **first** position at which it appears. From an injection to (iii): the injection makes A equinumerous with a subset of $\mathbb{N}$, and Lemma 3.2.1 sorts those. From (iii) back to a list: pad a finite list by repetition, as in the examples above.

Proof.

- (i) $\Rightarrow$ (ii). If $A = \emptyset$, the empty function $\emptyset \rightarrow \mathbb{N}$ is injective, vacuously. Otherwise let $g: \mathbb{N} \rightarrow A$ be a surjection. For $a \in A$, the preimage $g^{-1}(\{a\}) \subseteq \mathbb{N}$ is non-empty, since g is onto, so it has a least element by [Theorem 2.1.3](#). Define $h: A \rightarrow \mathbb{N}$ by

$$h(a) := \min g^{-1}(\{a\}).$$

Then $g(h(a)) = a$ for every $a \in A$, that is, $g \circ h = \text{id}_A$, so h is injective by [Theorem 1.5.1 \(a\)](#).

- (ii) $\Rightarrow$ (iii). Let $h: A \rightarrow \mathbb{N}$ be injective. With its codomain shrunk to $h(A)$, it is a bijection $A \rightarrow h(A)$ (Section 3.1), so $|A| = |h(A)|$. By [Lemma 3.2.1](#), $h(A) \subseteq \mathbb{N}$ is finite or countably infinite. If $h(A)$ has n elements, there is a bijection $\{1, \dots, n\} \rightarrow h(A)$, and following it by the inverse of $A \rightarrow h(A)$ gives a bijection $\{1, \dots, n\} \rightarrow A$ ([Proposition 3.1.1 \(b\)](#) for the inverse, [Theorem 1.4.4 \(c\)](#)); so A has n elements. If $|h(A)| = |\mathbb{N}|$, then $|A| = |\mathbb{N}|$ by [Proposition 3.1.1 \(c\)](#).

- (iii) $\Rightarrow$ (i). If A is countably infinite, a bijection $\mathbb{N} \rightarrow A$ is a surjection. Let A be finite with n elements. If $n = 0$, then $A = \emptyset$, which is countable. If $n \geq 1$, let $\varphi: \{1, \dots, n\} \rightarrow A$ be a bijection and define $f: \mathbb{N} \rightarrow A$ by

$$f(k) := \begin{cases} \varphi(k) & \text{if } 1 \leq k \leq n, \\ \varphi(1) & \text{if } k = 0 \text{ or } k > n. \end{cases}$$

Every $a \in A$ is $\varphi(k) = f(k)$ for some $k \in \{1, \dots, n\}$, so f is a surjection. This proves the theorem. ■

The step (i) $\Rightarrow$ (ii) picks one element $h(a)$ of $g^{-1}(\{a\})$ for **every** $a \in A$ at once, and it needs no new principle for that, because there is a rule that names the element: the least one. That works because the preimages are subsets of $\mathbb{N}$, which is well-ordered. Section 3.3 meets a step that looks the same and has no such rule, and it is the point at which the book adds the axiom of choice.

In practice, (ii) is often the cheapest test. An injection into $\mathbb{N}$ only has to keep different elements apart, while a surjection has to reach every element, and (iii) asks for a bijection, which must do both.

3.2.4 Closure properties

Three closure properties follow at once, each from the form of countability that suits it.

Proposition 3.2.3. Subsets of Countable Sets

Every subset of a countable set is countable.

Proof. Let B be countable and $A \subseteq B$. By [Theorem 3.2.2](#), (i) $\Rightarrow$ (ii), there is an injection $h: B \rightarrow \mathbb{N}$. Its restriction $h|_A: A \rightarrow \mathbb{N}$ is injective, since it has the same values. By (ii) $\Rightarrow$ (i), A is countable.

Proposition 3.2.4. Images of Countable Sets

Let A be countable and $f: A \rightarrow B$ a function. Then $f(A)$ is countable. In particular, if f is surjective, then B is countable.

Proof. If $A = \emptyset$, then $f(A) = \emptyset$, which is countable. Otherwise let $g: \mathbb{N} \rightarrow A$ be a surjection. The function f with codomain $f(A)$ is surjective (Section 3.1), so its composite with g is a surjection $\mathbb{N} \rightarrow f(A)$ by [Theorem 1.4.4](#) (b). If f is surjective, then $f(A) = B$.

For unions we interleave two lists, taking the entries of the first list at the even positions and those of the second at the odd positions. That needs the fact that every natural number is even or odd, in exactly one way.

Proposition 3.2.5. Union of Two Countable Sets

If A and B are countable, then $A \cup B$ is countable.

Idea. Two lists $a_0, a_1, \dots$ and $b_0, b_1, \dots$ merge into one, $a_0, b_0, a_1, b_1, \dots$: the entry at position $2q$ comes from the first list and the entry at position $2q + 1$ from the second. For this to define a function, every position must be even or odd, and not both.

Proof. If $A = \emptyset$, then $A \cup B = B$, and if $B = \emptyset$, then $A \cup B = A$; both are countable. So let $g: \mathbb{N} \rightarrow A$ and $h: \mathbb{N} \rightarrow B$ be surjections.

Claim. Every $a \in \mathbb{N}$ can be written as $a = 2q + r$ with $q \in \mathbb{N}$ and $r \in \{0, 1\}$, and the pair (q, r) is unique. In particular, no natural number is both of the form $2q$ and of the form $2q' + 1$ with $q, q' \in \mathbb{N}$.

Proof. By [Theorem 2.2.3](#) with divisor 2, there are unique integers q, r with $a = 2q + r$ and $0 \leq r < 2$. Then $r \in \mathbb{N}$ by [Proposition 2.1.2](#) (b), and $r \leq 1$ by [Lemma 2.1.1](#) (e), so $r \in \{0, 1\}$. If $q < 0$, then $q + 1 \leq 0$ by the discreteness clause [Proposition 2.1.2](#) (c), so $q \leq -1$ and $a = 2q + r \leq -2 + 1 < 0$, which contradicts $a \geq 0$ ([Lemma 2.1.1](#) (a)). So $q \geq 0$, and $q \in \mathbb{N}$ by [Proposition 2.1.2](#) (b). Uniqueness among pairs with $q \in \mathbb{N}$ follows from uniqueness among all integer pairs. The last sentence is the uniqueness of r .

Define $k: \mathbb{N} \rightarrow A \cup B$ by

$$k(2q) := g(q), \quad k(2q + 1) := h(q) \quad (q \in \mathbb{N}).$$

By the claim, each $a \in \mathbb{N}$ falls under exactly one of the two clauses, with exactly one q , so k is a function. It is onto: an element of A is $g(q) = k(2q)$ for some q , and an element of B is $h(q) = k(2q + 1)$ for some q , where $2q$ and $2q + 1$ lie in $\mathbb{N}$ by [Lemma 2.1.1](#) (b). So $A \cup B$ is

countable. ■

If A and B overlap, the interleaved list repeats their common elements, and the definition does not mind. By induction on the number of sets, a union of finitely many countable sets is countable. A union of **infinitely** many is a different matter, and Section 3.3 is about it.

3.2.5 Pairs of natural numbers

The set $\mathbb{N} \times \mathbb{N}$ looks two-dimensional, and a first attempt to list it shows where the difficulty lies.

Example. Listing the rows one after another

Arrange $\mathbb{N} \times \mathbb{N}$ in rows, the m -th row being $(m, 0), (m, 1), (m, 2), \dots$. Consider the “list”

$(0, 0), (0, 1), (0, 2), \dots$, then $(1, 0), (1, 1), \dots$, then $(2, 0), \dots$

Is it a surjection $\mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$?

Solution. No. A list indexed by $\mathbb{N}$ gives each entry a position $n \in \mathbb{N}$. Here position n holds $(0, n)$, for every $n \in \mathbb{N}$: that already defines a function $f(n) := (0, n)$ on all of $\mathbb{N}$. No position is left for $(1, 0)$: there is no $n \in \mathbb{N}$ with $f(n) = (1, 0)$, since the first coordinate of every value is 0. The word “then” asks for positions after all the natural numbers, and there are none. So the claim that this is a list of $\mathbb{N} \times \mathbb{N}$ is false. It is a list of the first row only.

Warning. A list must give every entry a **finite** position. “First all of row 0, then all of row 1” never reaches row 1, because row 0 uses up every position. The same mistake hides in any listing that runs through an infinite block before starting the next one.

The conclusion that $\mathbb{N} \times \mathbb{N}$ is countable is nevertheless true.

Theorem 3.2.6. The Plane of Natural Numbers Is Countable

The map $p: \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$, $p(m, n) := 2^m 3^n$, is injective. Hence $\mathbb{N} \times \mathbb{N}$ is countable.

Try it first (ten minutes): list $\mathbb{N} \times \mathbb{N}$, or inject it into $\mathbb{N}$. Which is less work to write down and check?

Idea. The rows-first list fails, as we saw. A list that works runs along the diagonals $m + n = 0, 1, 2, \dots$, each of which is finite; but writing a formula for the position of (m, n) and checking that every position is used takes real work. By [Theorem 3.2.2](#) it is enough to **inject** $\mathbb{N} \times \mathbb{N}$ into $\mathbb{N}$, and an injection only has to be a code from which m and n can be read back. In $2^m 3^n$, the power of 2 records m and the power of 3 records n . To read them back we do not cite unique factorization into primes, which we have not proved. Parity does the job: cancel the smaller power of 2; if any factor 2 is left on one side, that side is even and the other is

odd.

Proof. First, $p(m, n) \in \mathbb{N}$ by [Proposition 2.3.4](#) (d) and [Lemma 2.1.1](#) (b). Next, every power 3^k with $k \in \mathbb{N}$ is odd, that is, $3^k = 2j + 1$ for some $j \in \mathbb{N}$: by Logic and Proof, Axiom 4.2.1, $3^0 = 1 = 2 \cdot 0 + 1$, and if $3^k = 2j + 1$, then $3^{k+1} = 3^k \cdot 3 = 2(3j + 1) + 1$ ([Definition 2.3.2](#)).

Let $(m, n), (m', n') \in \mathbb{N} \times \mathbb{N}$ with $2^m 3^n = 2^{m'} 3^{n'}$. Suppose $m \neq m'$. The hypothesis is symmetric in the two pairs, so we may assume $m > m'$. Then $m - m' \in \mathbb{N}$ and $m - m' \geq 1$ ([Lemma 2.1.1](#) (d), (e)), and $2^m = 2^{m'} 2^{m-m'}$ by [Proposition 2.3.4](#) (a). Canceling $2^{m'}$, which is positive by [Proposition 2.3.4](#) (e), gives

$$2^{m-m'} 3^n = 3^{n'}.$$

The left side is even: $2^{m-m'} = 2^{m-m'-1} \cdot 2$ by [Definition 2.3.2](#), where $m - m' - 1 \in \mathbb{N}$ by [Lemma 2.1.1](#) (c), so the left side is 2 times the natural number $2^{m-m'-1} 3^n$. The right side is odd, as shown above. No natural number is both even and odd, by the uniqueness of the remainder in [Theorem 2.2.3](#) with divisor 2 (the claim in the proof of [Proposition 3.2.5](#)). This contradiction shows $m = m'$.

Canceling 2^m now gives $3^n = 3^{n'}$. By [Proposition 2.3.4](#) (e) with $u = 3 > 1$, $n < n'$ would give $3^n < 3^{n'}$, and $n' < n$ would give $3^{n'} < 3^n$; so $n = n'$. Hence $(m, n) = (m', n')$, and p is injective. By [Theorem 3.2.2](#), (ii) $\Rightarrow$ (i), $\mathbb{N} \times \mathbb{N}$ is countable. ■

Alternatively, one can list $\mathbb{N} \times \mathbb{N}$ along its diagonals: first the one pair with $m + n = 0$, then the two pairs with $m + n = 1$, then the three with $m + n = 2$, and so on. Each diagonal is finite, so every pair is reached after finitely many steps, and the list never stalls on an infinite block. Some readers find this picture more convincing than the code $2^m 3^n$. The injection is shorter to write out and to check, which is why the proof uses it.

The theorem passes to products of any countable sets. Chapter 1 defines X^n for one set X ; for several sets we first say what the product is. For an integer $k \geq 2$ and sets $A_1, \dots, A_k$, let $U := \bigcup_{i \in \{1, \dots, k\}} A_i$ and

$$A_1 \times \cdots \times A_k := \{x \in U^k : x = (x_1, \dots, x_k) \text{ with } x_i \in A_i \text{ for every } i\},$$

the k -tuples whose i -th coordinate lies in A_i . It is a set by (W3), (W5) and (W1) of Section 1.1. For $k = 2$ it is $A_1 \times A_2$ of [Definition 1.3.2](#), a pair being a 2-tuple, and if every A_i is the same set X , it is X^k .

Corollary 3.2.7. Products of Countable Sets

- (a) If A and B are countable, then $A \times B$ is countable.
- (b) Let $k \geq 2$ be an integer. If $A_1, \dots, A_k$ are countable, then $A_1 \times \cdots \times A_k$ is countable.

Idea. For two factors, injections $A \rightarrow \mathbb{N}$ and $B \rightarrow \mathbb{N}$ turn $A \times B$ into a subset of $\mathbb{N} \times \mathbb{N}$, and the code $2^m 3^n$ does the rest. More factors are added one at a time: a $(k + 1)$ -tuple is a k -tuple followed by one more coordinate.

Proof. (a) By [Theorem 3.2.2](#), (i) $\Rightarrow$ (ii), there are injections $i: A \rightarrow \mathbb{N}$ and $j: B \rightarrow \mathbb{N}$. The map $A \times B \rightarrow \mathbb{N} \times \mathbb{N}$, $(a, b) \mapsto (i(a), j(b))$, is injective: if $(i(a), j(b)) = (i(a'), j(b'))$, then $i(a) = i(a')$ and $j(b) = j(b')$ by [Definition 1.3.1](#), so $a = a'$ and $b = b'$. Following it by the injection p of [Theorem 3.2.6](#) gives an injection $A \times B \rightarrow \mathbb{N}$ ([Theorem 1.4.4](#) (a)), and (ii) $\Rightarrow$ (i) concludes. No case split is needed: if $A = \emptyset$ or $B = \emptyset$, then $A \times B = \emptyset$ ([Proposition 1.3.1](#)), so the map $(a, b) \mapsto (i(a), j(b))$ and its composite with p are empty functions ([Proposition 1.4.2](#) (a)), which are injective vacuously.

(b) We use [Theorem 2.2.1](#) with $n_0 = 2$ on the statement “for all countable sets $A_1, \dots, A_k$, the set $A_1 \times \dots \times A_k$ is countable”. *Base case* $k = 2$. This is (a). *Step.* Let $k \geq 2$, suppose the statement for k , and let $A_1, \dots, A_{k+1}$ be countable. The map

$$((a_1, \dots, a_k), a_{k+1}) \mapsto (a_1, \dots, a_{k+1})$$

is a function, since a k -tuple determines its coordinates (Section 1.3), and it sends $(A_1 \times \dots \times A_k) \times A_{k+1}$ onto $A_1 \times \dots \times A_{k+1}$: a $(k+1)$ -tuple with $x_i \in A_i$ is the image of $((x_1, \dots, x_k), x_{k+1})$. By the hypothesis, $A_1 \times \dots \times A_k$ is countable, so the domain is countable by (a), and the $(k+1)$ -fold product is countable by [Proposition 3.2.4](#). By [Theorem 2.2.1](#) the statement holds for every $k \geq 2$. ■

The map in the step is in fact a bijection, by the characteristic property of tuples (Section 1.3). The proof uses that property only to see that the map is well defined, and of the map it uses only that it is onto; injectivity is never needed. Tuples of different lengths are different objects, and the book has no 1-tuples, so the corollary starts at $k = 2$; a single countable set needs no corollary.

Quick check. Let L be the set of the 26 letters a–z. Is the set of all strings of exactly 5 letters, that is L^5 , countable? Which results do you use?

Solution. Yes. The set L is finite, so it is countable by [Theorem 3.2.2](#), (iii) $\Rightarrow$ (i). The strings of length 5 are the 5-tuples of letters, $L^5 = L \times L \times L \times L \times L$, which is countable by [Corollary 3.2.7](#) (b) with $k = 5$. (It is in fact finite, with 26^5 elements: the map in the step of the proof of [Corollary 3.2.7](#) (b) is a bijection $L^k \times L \rightarrow L^{k+1}$ (the remark above), so Exercise C2 (a) of Section 3.1, applied for $k = 2, 3, 4$ after $L^2 = L \times L$, counts it. Countability needs less than the count.)

All strings of every finite length together form the union of the sets L^k over k , a union of infinitely many countable sets. Section 3.3 takes up such unions.

3.2.6 The integers and the rationals

Here is the answer to the question that opened the section. Recall from Section 2.1 that $\mathbb{Z} := \mathbb{N} \cup \{-n : n \in \mathbb{N}\}$, $\mathbb{Z}_{>0} := \{n \in \mathbb{N} : n \neq 0\}$ and $\mathbb{Q} := \{m/n : m \in \mathbb{Z}, n \in \mathbb{Z}_{>0}\}$.

Theorem 3.2.8. The Integers and the Rationals Are Countable

$\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ are countable. Each of them is countably infinite.

Idea. For $\mathbb{Z}$, interleave the non-negative and the negative integers, as in [Proposition 3.2.5](#), but with explicit lists, so that the result is a bijection: $0, -1, 1, -2, 2, -3, \dots$. For $\mathbb{Q}$, do not try to list the fractions directly. Every rational is m/n for some pair $(m, n) \in \mathbb{Z} \times \mathbb{Z}_{>0}$, and that set of pairs is countable, so $\mathbb{Q}$ is the image of a countable set.

Proof. $\mathbb{N}$. The identity $\text{id}_{\mathbb{N}}$ is a bijection $\mathbb{N} \rightarrow \mathbb{N}$.

$\mathbb{Z}$. By the claim in the proof of [Proposition 3.2.5](#), every $n \in \mathbb{N}$ is $2k$ or $2k + 1$ for exactly one $k \in \mathbb{N}$, and not both. So we may define $z: \mathbb{N} \rightarrow \mathbb{Z}$ by

$$z(2k) := k, \quad z(2k + 1) := -(k + 1) \quad (k \in \mathbb{N}),$$

which begins $z(0) = 0, z(1) = -1, z(2) = 1, z(3) = -2, z(4) = 2$. It is injective: $z(2k) = k \geq 0$, while $z(2k + 1) = -(k + 1) \leq -1 < 0$ ([Lemma 2.1.1](#) (a)), so an even and an odd input never share a value; and $k = k'$ or $-(k + 1) = -(k' + 1)$ gives $k = k'$. It is surjective: let $m \in \mathbb{Z}$. If $m \in \mathbb{N}$, then $m = z(2m)$. Otherwise $m = -n$ with $n \in \mathbb{N}$, and $n \neq 0$, since $-0 = 0 \in \mathbb{N}$; so $n - 1 \in \mathbb{N}$ by [Lemma 2.1.1](#) (c), and $m = -((n - 1) + 1) = z(2(n - 1) + 1)$. So z is a bijection.

$\mathbb{Q}$. The set $\mathbb{Z}_{>0} \subseteq \mathbb{N}$ is countable by [Proposition 3.2.3](#), and $\mathbb{Z}$ is countable, so $\mathbb{Z} \times \mathbb{Z}_{>0}$ is countable by [Corollary 3.2.7](#) (a). The map $(m, n) \mapsto m/n$ from $\mathbb{Z} \times \mathbb{Z}_{>0}$ to $\mathbb{Q}$ is defined, since $n \geq 1$ is not 0, and it is surjective by the definition of $\mathbb{Q}$. By [Proposition 3.2.4](#), $\mathbb{Q}$ is countable. It is infinite: $\mathbb{N} \subseteq \mathbb{Q}$, since $n = n/1$, and $\mathbb{N}$ is infinite ([Example \(Hilbert's Hotel\)](#)), so $\mathbb{Q}$ is not finite by [Proposition 3.1.6](#). By [Theorem 3.2.2](#), (i) $\Rightarrow$ (iii), $\mathbb{Q}$ is countably infinite. ■

The list of $\mathbb{Q}$ that the proof produces repeats every rational infinitely often: $1/1, 2/2, 3/3, \dots$ are all 1. The definition allows it, and that is what made the proof three lines. No line of the proof, or of the results it cites, uses the axiom of choice of Section 3.3: each surjection or injection it relies on is either given by a formula or named by a single $\exists E$, and the only selections made for many elements at once are least elements ([Theorem 2.1.3](#)). That $\mathbb{Q}$ can be listed is the fact *Analysis on the Real Line*, Chapter 1 quotes when it compares $\mathbb{Q}$ with $\mathbb{R}$: the rationals, dense as they are, can be listed, while *Analysis on the Real Line*, Chapter 1 shows that the reals cannot.

3.2.7 Exercises

3.2.7.1 A. Check your understanding

Exercise. A1

- (a) Define “ A is countable”, and say why the empty set is mentioned separately in the definition.
- (b) True or false, with a reason: every infinite subset of $\mathbb{Q}$ is countably infinite; every

countable set is finite or equinumerous with $\mathbb{N}$.

- (c) Which of the three forms (i)–(iii) of [Theorem 3.2.2](#) is quickest to check for $\{n^2 : n \in \mathbb{N}\}$, and why?
- (d) Where does the proof of (i) $\Rightarrow$ (ii) in [Theorem 3.2.2](#) use [Theorem 2.1.3](#), and to which set is it applied?

Solution.

- (a) A is countable if $A = \emptyset$ or there is a surjection $\mathbb{N} \rightarrow A$ ([Definition 3.2.1](#)). The empty set is mentioned separately because there is no function $\mathbb{N} \rightarrow \emptyset$ at all ([Proposition 1.4.2](#) (b)), so the surjection clause never holds for $\emptyset$.
- (b) Both are true. An infinite $A \subseteq \mathbb{Q}$ is countable by [Proposition 3.2.3](#) and [Theorem 3.2.8](#); by [Theorem 3.2.2](#), (i) $\Rightarrow$ (iii), it is finite or countably infinite, and it is not finite. The second statement is (i) $\Rightarrow$ (iii) itself, with “countably infinite” unfolded by [Definition 3.2.2](#).
- (c) Form (i): the set is given as the image of $\mathbb{N}$ under $n \mapsto n^2$, and a function is a surjection onto its image, so nothing is left to check. Form (ii) is almost as quick: the inclusion into $\mathbb{N}$ is injective. Form (iii) needs a bijection with $\mathbb{N}$, which costs a proof that $n \mapsto n^2$ is injective on $\mathbb{N}$.
- (d) In defining $h(a) := \min g^{-1}(\{a\})$: well-ordering is applied to the preimage $g^{-1}(\{a\}) \subseteq \mathbb{N}$, which is non-empty because g is onto. It gives every a a definite position, the first one at which a appears in the list, so no choice is needed.

3.2.7.2 B. Practice

Exercise. B1

Prove that each of the following sets is countable.

- (a) $2\mathbb{N} \times \mathbb{Z}$.
- (b) $\{(m, n) \in \mathbb{N} \times \mathbb{N} : m < n\}$.
- (c) $\mathbb{Q} \times \mathbb{Q}$, the points of the plane with rational coordinates.
- (d) The set of open intervals with rational endpoints, that is, of the sets $\{x \in \mathbb{R} : a < x < b\}$ with $a, b \in \mathbb{Q}$ and $a < b$.

Solution.

- (a) $2\mathbb{N} \subseteq \mathbb{N}$ is countable by [Proposition 3.2.3](#), and $\mathbb{Z}$ is countable by [Theorem 3.2.8](#); so $2\mathbb{N} \times \mathbb{Z}$ is countable by [Corollary 3.2.7](#) (a).
- (b) It is a subset of $\mathbb{N} \times \mathbb{N}$, which is countable by [Theorem 3.2.6](#), so it is countable by [Proposition 3.2.3](#).
- (c) $\mathbb{Q}$ is countable ([Theorem 3.2.8](#)), so $\mathbb{Q} \times \mathbb{Q}$ is countable by [Corollary 3.2.7](#) (a).
- (d) Let $D := \{(a, b) \in \mathbb{Q} \times \mathbb{Q} : a < b\}$, countable by (c) and [Proposition 3.2.3](#). Define $J : D \rightarrow \mathcal{P}(\mathbb{R})$ by $J(a, b) := \{x \in \mathbb{R} : a < x < b\}$. The set of intervals in question is the image $J(D)$, so it is countable by [Proposition 3.2.4](#). (Here $(a, b) \in D$ is an ordered

pair, and the interval it names is written in set-builder notation, so the two readings of “ (a, b) ” never meet.)

Exercise. B2: Putting a proof in order

The lines (i)–(ix) below include the seven lines of a proof that $2^m 3^n = 2^{m'} 3^{n'}$ implies $m = m'$ and $n = n'$, for $m, n, m', n' \in \mathbb{N}$, shuffled, together with two lines that do not belong. Write the seven lines in an order that makes a correct proof, and say why each of the other two lines fails.

- (i) The left side is even, because $m - m' \geq 1$ by the supposition; the right side is odd; no natural number is both ([Theorem 2.2.3](#), uniqueness of the remainder for divisor 2). This is a contradiction.
- (ii) Suppose $2^m 3^n = 2^{m'} 3^{n'}$.
- (iii) The left side is odd and the right side is even, a contradiction.
- (iv) Canceling 2^m gives $3^n = 3^{n'}$.
- (v) Suppose $m \neq m'$; by symmetry, say $m > m'$.
- (vi) Since $k \mapsto 3^k$ is strictly increasing on $\mathbb{N}$, $n = n'$.
- (vii) Since 2 and 3 are prime, unique factorization gives $m = m'$ and $n = n'$.
- (viii) Canceling $2^{m'}$ gives $2^{m-m'} 3^n = 3^{n'}$.
- (ix) So $m = m'$.

Solution. The order is (ii), (v), (viii), (i), (ix), (iv), (vi). No other order is admissible. Line (ii) introduces the four numbers and the equation. Lines (v), (viii) and (i) are one proof by contradiction: (v) opens it, (viii) uses the supposition $m > m'$ to cancel, and (i) closes it with a contradiction, so (ix), its conclusion, must come right after (i). Line (iv) cancels 2^m on both sides, which needs $m = m'$, so it follows (ix); and (vi) uses the equation of (iv).

Line (iii) reverses the parities. The left side $2^{m-m'} 3^n$ has a factor 2 because $m - m' \geq 1$, so it is even; the right side $3^{n'}$ is odd. A line asserting the opposite is false: for $m = 1$, $m' = 0$, $n = n' = 0$, the left side is 2 and the right side is 1.

Line (vii) would end the proof at once, but it cites unique factorization into primes, which the book has neither proved nor taken as a school fact: Section 2.2 sets it aside as a fact it does not prove or use (after [Definition 2.2.2](#)), and the idea before [Theorem 3.2.6](#) declines to cite it. So (vii) cites a line that is not available, (N9) of Logic and Proof, Example (What is not allowed). The theorem is true, and the conclusion of (vii) is true; the error is in the citation, which is exactly why the proof in the text goes through parity instead.

Exercise. B3: A surjection by formula

Give an explicit surjection $\mathbb{N} \rightarrow \mathbb{N} \times \{0, 1\}$ and prove that it is one. List the properties that any correct answer must have, so that a different answer can be checked against them.

Solution. By the claim in the proof of [Proposition 3.2.5](#), every $a \in \mathbb{N}$ is $2q + r$ for exactly one pair (q, r) with $q \in \mathbb{N}$ and $r \in \{0, 1\}$. Define $s: \mathbb{N} \rightarrow \mathbb{N} \times \{0, 1\}$ by $s(2q + r) := (q, r)$: so $s(0) = (0, 0)$, $s(1) = (0, 1)$, $s(2) = (1, 0)$, $s(3) = (1, 1)$. The uniqueness of the pair makes s a function. It is surjective: given $(q, r) \in \mathbb{N} \times \{0, 1\}$, the number $a := 2q + r$ lies in $\mathbb{N}$ ([Lemma 2.1.1](#) (b)), and $s(a) = (q, r)$. (It is even a bijection, by the uniqueness of the pair.)

A correct answer must:

- assign to **every** $a \in \mathbb{N}$ exactly one value, so that it is a function on all of $\mathbb{N}$;
- take its values in $\mathbb{N} \times \{0, 1\}$;
- reach every pair $(q, 0)$ **and** every pair $(q, 1)$, for every $q \in \mathbb{N}$, each at a finite position.

For instance $s'(2q) := (q, 1)$, $s'(2q + 1) := (q, 0)$ also qualifies. The rows-first attempt $a \mapsto (a, 0)$, “then” the pairs $(q, 1)$, fails the third property ([Example \(Listing the rows one after another\)](#)).

3.2.7.3 C. Going deeper

Exercise. C1: Finite subsets of the natural numbers

Let $\mathcal{F}$ be the set of all finite subsets of $\mathbb{N}$. Prove that $\mathcal{F}$ is countable, without using any result about unions of infinitely many sets.

Solution (Hint 1). Inject $\mathcal{F}$ into $\mathbb{N}$ by sending F to $S_N(F) := \sum_{k=0}^N \mathbf{1}_F(k) 2^k$, for any N with $F \subseteq \{0, \dots, N\}$ (the value does not depend on N): the number whose binary digits are the indicator of F .

Solution (Hint 2). To prove injectivity, let m be the largest element at which two different sets $F \neq G$ differ (the largest element of $(F \setminus G) \cup (G \setminus F)$, a non-empty finite subset of $\mathbb{N}$). The terms above m are the same for F and G . Then use $\sum_{k=0}^{m-1} 2^k = 2^m - 1 < 2^m$.

Solution. The set $\mathcal{F} \subseteq \mathcal{P}(\mathbb{N})$ exists by specification. We build an injection $\sigma: \mathcal{F} \rightarrow \mathbb{N}$ in four steps. For $N \in \mathbb{N}$ and $F \subseteq \{0, \dots, N\}$ let

$$S_N(F) := \sum_{k=0}^N \mathbf{1}_F(k) 2^k,$$

a finite sum as in [Definition 2.3.1](#) (indexed from 0), where $\mathbf{1}_F: \mathbb{N} \rightarrow \{0, 1\}$ is the indicator function of F . Each term is 0 or 2^k , and $2^k \in \mathbb{N}$ by [Proposition 2.3.4](#) (d). So $S_N(F) \in \mathbb{N}$, by induction on N through the recursive clause of [Definition 2.3.1](#), using [Lemma 2.1.1](#) (b) at each step.

Step 1: a bound. Each term is at most 2^k , so

$$0 \leq S_N(F) \leq \sum_{k=0}^N 2^k = 2^{N+1} - 1,$$

by [Proposition 2.3.6](#) with $r = 2$ (shifting the index by one).

Step 2: one more term. Splitting off the last term (Section 2.3), for $F \subseteq \{0, \dots, N+1\}$,

$$S_{N+1}(F) = S_N(F \cap \{0, \dots, N\}) + \mathbf{1}_F(N+1) 2^{N+1}.$$

In particular, if $F \subseteq \{0, \dots, N\}$, then $S_{N+1}(F) = S_N(F)$. By induction on $N' \geq N$, $S_{N'}(F) = S_N(F)$ for every $N' \geq N$: the value does not depend on which bound N we use.

Step 3: injective for a fixed N . We prove by Logic and Proof, Axiom 4.2.1 on N : if $F, G \subseteq \{0, \dots, N\}$ and $S_N(F) = S_N(G)$, then $F = G$. For $N = 0$, F and G are $\emptyset$ or $\{0\}$, with values 0 and 1. For the step, let $F, G \subseteq \{0, \dots, N+1\}$ with $S_{N+1}(F) = S_{N+1}(G)$, and write $F' := F \cap \{0, \dots, N\}$, $G' := G \cap \{0, \dots, N\}$. If $N+1 \in F$ and $N+1 \notin G$, then Steps 2 and 1 give

$$S_{N+1}(F) \geq 2^{N+1} > 2^{N+1} - 1 \geq S_N(G') = S_{N+1}(G),$$

a contradiction; the case $N+1 \in G \setminus F$ is the same with F and G exchanged. So $N+1$ lies in both or in neither, the last terms in Step 2 agree, and $S_N(F') = S_N(G')$. By the hypothesis, $F' = G'$. Since $\{0, \dots, N+1\} = \{0, \dots, N\} \cup \{N+1\}$, it follows that $F = G$.

Step 4: the injection. Every $F \in \mathcal{F}$ lies in some $\{0, \dots, N\}$: if $F = \emptyset$, take $N = 0$; otherwise F is a non-empty finite set of real numbers, so it has a maximum $M \in F \subseteq \mathbb{N}$ ([Proposition 2.3.8](#), as in the remark at the end of Section 3.1), and $F \subseteq \{0, \dots, M\}$. Define $\sigma(F) := S_N(F)$ for any such N ; by Step 2 this does not depend on N . If $\sigma(F) = \sigma(G)$, take N at least as large as a bound for F and a bound for G ; then $S_N(F) = S_N(G)$, and Step 3 gives $F = G$. So σ is injective, and $\mathcal{F}$ is countable by [Theorem 3.2.2](#), (ii) $\Rightarrow$ (i).

For example $\sigma(\emptyset) = 0$ (the empty set contributes no terms), $\sigma(\{0, 2\}) = 1 + 4 = 5$ and $\sigma(\{1, 3\}) = 2 + 8 = 10$: $\sigma(F)$ is the number whose binary numeral has a 1 exactly in the places listed in F . Step 3 is the uniqueness of binary numerals, and Step 4 uses no principle of choice: every value is given by a formula.

Marking scheme. (10 points) - 2: defines a map from the finite subsets to $\mathbb{N}$ explicitly, including the value at $\emptyset$, and checks that its values are natural numbers. - 2: makes the map well defined, either by a bound-independent definition (as in Step 2) or by summing over the elements of F listed in increasing order. - 4: proves injectivity, either by comparing the terms at the largest element at which F and G differ, using $\sum_{k=0}^{m-1} 2^k < 2^m$, or by an induction on the bound as in Step 3, or by an equivalent proved uniqueness argument for binary numerals. - 2: concludes with [Theorem 3.2.2](#), (ii) $\Rightarrow$ (i). - Route by surjection: an explicit map $\mathbb{N} \rightarrow \mathcal{F}$ that reads off binary digits (for example, $D(n)$ is the set of $k \leq n$ for which the quotient of n by 2^k in [Theorem 2.2.3](#) is odd) earns the same points: 2 for defining the map, 2 for checking that each value is a finite subset of $\mathbb{N}$, 4 for a proof that every $F \in \mathcal{F}$ is a value, and 2 for concluding by [Definition 3.2.1](#). Fatal errors (0 for the step): writing $\mathcal{F}$ as the union over n of the sets of n -element subsets and citing a result on countable unions (the task excludes it); asserting injectivity, or surjectivity on the second route, “by uniqueness of binary expansions” with no proof.

Exercise. C2: Find the error

Here is a “proof” that every subset of a countable set is countable.

- Let B be countable and $A \subseteq B$. (1) If $A = \emptyset$, then A is countable by definition. (2) Otherwise $B \neq \emptyset$, so there is a surjection $g: \mathbb{N} \rightarrow B$. (3) Define $h: \mathbb{N} \rightarrow A$

by $h(n) := g(n)$. (4) Let $a \in A$. Then $a \in B$, so $a = g(n)$ for some $n \in \mathbb{N}$, and $h(n) = a$. (5) Hence h is a surjection onto A , and A is countable.

Is the conclusion true? Does the argument prove it? If not, name the line that fails, the rule it breaks, and give a counterexample **to that line**. Then repair the proof.

Solution. The conclusion is true: it is [Proposition 3.2.3](#). The argument does not prove it.

Line (3) fails. It does not define a function from $\mathbb{N}$ to A : [Definition 1.4.1](#) requires, for **every** $n \in \mathbb{N}$, exactly one value **in** A , and when $g(n) \notin A$ the formula offers no value in A at all. Counterexample to line (3): $B = \mathbb{N}$, $g = \text{id}_{\mathbb{N}}$, $A = \{1\}$. Then “ $h(2) = g(2) = 2$ ”, and $2 \notin A$, so h is not a function into A . Lines (1), (2) and (4) are correct, and (5) would follow if h were a function $\mathbb{N} \rightarrow A$.

Repair. In the case $A \neq \emptyset$, fix one element $a_0 \in A$ (one $\exists E$) and define

$$h(n) := \begin{cases} g(n) & \text{if } g(n) \in A, \\ a_0 & \text{if } g(n) \notin A. \end{cases}$$

Now h is a function $\mathbb{N} \rightarrow A$, and line (4) shows that it is onto: for $a \in A$, some n has $g(n) = a \in A$, and then $h(n) = g(n) = a$. The repaired proof uses that A is non-empty, which line (1) had already split off; the flawed line ignored the elements of $B \setminus A$, and the default value a_0 is where they go. The proof in the text avoids the issue by restricting an **injection** instead, since a restriction of a function is always a function.

Marking scheme. (10 points) - 2: says the conclusion is true, citing [Proposition 3.2.3](#) or an equivalent proof. - 3: names line (3) and the broken requirement: a function into A needs a value in A at every n ([Definition 1.4.1](#)). - 2: gives a concrete instance of the failure of line (3), with a specific n whose value leaves A . - 3: repairs the proof with a default value in A (or by restricting an injection), and checks surjectivity. Fatal errors (0 for the step): calling the conclusion false; blaming line (4) or (5); a “counterexample” that shows a subset that is not countable.

3.3 Countable Unions and the Axiom of Choice

Section 3.2 listed $\mathbb{N} \times \mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$, and showed that a union of two countable sets is countable ([Proposition 3.2.5](#)). The next question is a union of **countably many** countable sets. The natural proof is short, and it contains one step that none of the rules met so far can justify. This section isolates that step, adds the principle that licenses it, the axiom of choice, and then shows how to tell a proof that needs the axiom from one that only seems to.

3.3.1 A union of countably many lists

Let $A_0, A_1, A_2, \dots$ be countable sets, a family $(A_n)_{n \in \mathbb{N}}$ in the sense of [Definition 1.2.3](#). We want to prove that $\bigcup_{n \in \mathbb{N}} A_n$ is countable.

Here is the proof most people write first. List each set: $A_n = \{a_{n,0}, a_{n,1}, a_{n,2}, \dots\}$. Put $a_{n,m}$ in row n and column m of an infinite grid. Every element of the union stands somewhere in the

grid, and the grid is indexed by $\mathbb{N} \times \mathbb{N}$, which is countable (Theorem 3.2.6). So the union is the image of a countable set, and it is countable.

Every step but the first, “list each set”, is a result of Section 3.2. Where do the lists come from?

3.3.2 The obstacle, rule by rule

For a non-empty A_n , “ A_n is countable” says that **there exists** a surjection $\mathbb{N} \rightarrow A_n$ (Definition 3.2.1). To use it we apply $\exists E$ (Logic and Proof, Definition 5.6.1): we name the object and argue inside the box of that name. One $\exists E$ names **one** surjection, for **one** n .

- **Finitely many sets.** The proof of Proposition 3.2.5 used two $\exists E$ ’s; for $A_0, \dots, A_9$ we could nest ten boxes.
- **All $n \in \mathbb{N}$ at once.** The grid needs a list of A_n for every n , assembled into one object: a function that sends each n to a surjection $\mathbb{N} \rightarrow A_n$. A proof is a finite object. It cannot contain one $\exists E$ for each of infinitely many indices, and no rule of *Logic and Proof*, Chapter 4 makes infinitely many choices in one step.

Two escape routes fail. *Recursion* (Theorem 2.3.1) defines a function on all of $\mathbb{N}$ at once, but from a rule g given in advance, each value determined by the previous one. The list of A_{n+1} is not determined by the list of A_n , and a rule producing a list of every A_{n+1} would already be the missing object. *Induction* on k , with Proposition 3.2.5 at each step, proves “for every k , $A_0 \cup \dots \cup A_k$ is countable” (Exercise C3 writes it out), but that is a statement about each k separately; nothing makes the lists for different k fit together.

Section 1.5 met the same wall: a right inverse of a surjection must pick a preimage of **every** point at once, and the rules could do that only for finitely many points, or when a rule names the preimage.

3.3.3 The axiom of choice

None of the working rules (W1)–(W6) of Section 1.1 forms a set out of infinitely many arbitrary picks. The principle we add is the simplest statement that does.

From any family of non-empty sets, one may pick one element from each set, all at once.

Axiom 3.3.1. Axiom of Choice

Let I be a set, and let $(A_i)_{i \in I}$ be a family of sets such that A_i is **non-empty** for **every** $i \in I$. Then **there exists** a function

$$c: I \rightarrow \bigcup_{i \in I} A_i \quad \text{with} \quad c(i) \in A_i \quad \text{for every } i \in I.$$

Such a function c is called a **choice function** for the family.

In symbols, with Y^X the set of functions of Section 1.4 and $U := \bigcup_{i \in I} A_i$:

$$\forall i \in I (A_i \neq \emptyset) \rightarrow \exists c \in U^I \quad \forall i \in I \quad c(i) \in A_i.$$

In words: the hypothesis says that each set of the family has an element; the conclusion says that **one** function picks an element of every A_i at the same time. The codomain, a set by (W3), only gives c somewhere to take its values. The axiom asserts that c exists and gives no recipe for it.

Its status. For finitely many indices the axiom adds nothing (proved below); for infinitely many it is a genuinely new principle. It is independent of the other axioms of set theory: Gödel (1938) showed that adding it creates no contradiction that was not there already, and Cohen (1963) that it cannot be proved from them. A later chapter states its best-known equivalent, Zorn’s lemma; the companion volume on set theory proves it equivalent to the axiom, together with the other equivalents such as the well-ordering theorem, and surveys the independence results. This book uses the axiom openly on the main line, among other places: here, for countable unions and right inverses; in Section 3.6, for infinite subsets that can be listed; in *Analysis on the Real Line*, Chapter 1, for sup approximations (the remark below); in *Measure and Integration*, Chapter 1, for the Vitali set; in *Functional Analysis*, Chapter 4, through Zorn’s lemma, for the Hahn–Banach theorem; and in *Functional Analysis*, Chapter 6, for Tychonoff’s theorem. Each use is flagged at its line.

Remark. A corollary in *Analysis on the Real Line*, Chapter 1 gives, for each $n \geq 1$ separately, some $a_n \in A$ with $a_n > \sup A - \frac{1}{n}$: one $\exists E$ per n , and no axiom. Forming the **sequence** $(a_n)_{n \geq 1}$ from these is a choice function for the family $(\{a \in A : a > \sup A - \frac{1}{n}\})_{n \in \mathbb{Z}_{>0}}$, an instance of [Axiom 3.3.1](#) with the countable index set $I = \mathbb{Z}_{>0}$. *Analysis on the Real Line*, Chapter 1 flags exactly this step.

Examples.

1. *The empty family.* If $I = \emptyset$, the empty function $\emptyset \rightarrow \bigcup_{i \in \emptyset} A_i = \emptyset$ ([Proposition 1.4.2 \(a\)](#)) is a choice function: the condition “for every $i \in \emptyset$ ” holds vacuously. No axiom is needed.
2. *A family with a rule.* Let $I = \mathbb{Z}_{>0}$ and $A_n := \{k \in \mathbb{N} : n \mid k, k \neq 0\}$, the positive multiples of n . Each A_n contains n . The rule $c(n) := n$ is a choice function, defined by specification ([Definition 1.1.3](#)); no axiom is needed.
3. *A family with no rule in sight.* Let I be the set of non-empty subsets of $\mathbb{R}$, and $A_S := S$ for $S \in I$. A choice function picks an element of every non-empty set of reals. “The least element” fails, since $(0, 1)$ has none, and no rule that works for every S is known. Here the axiom supplies what the other rules do not.

A non-example by minimal change. In Example 2, add the index $n = 0$. The positive multiples of 0 form the empty set ($0 \mid k$ means $k = 0$), so $A_0 = \emptyset$, and no function can have $c(0) \in A_0$. Only the clause “ A_i is non-empty for every i ” fails, and without it the conclusion is false. That is why the proofs below remove the empty sets from a family before they choose. (Dropping “all at once” instead would leave nothing beyond the hypothesis.)

3.3.4 When choice is not needed

Two kinds of construction look like choice and are not; telling them apart from a real use is the skill this section is for.

Finitely many sets. For every finite set I and every family $(A_i)_{i \in I}$ of non-empty sets, a choice function exists, with no appeal to the axiom. For $I = \{1, 2, 3\}$, three applications of $\exists E$ name $a_i \in A_i$, and $\{(1, a_1), (2, a_2), (3, a_3)\}$ is the graph of a choice function. For a finite set of arbitrary size we induct on its number of elements ([Definition 3.1.2](#)). Let $P(n)$ be: “for every set I with n elements and every family $(A_i)_{i \in I}$ of non-empty sets, there is a choice function”. *Base case* $P(0)$: a set with 0 elements is the image of a bijection from $\{1, \dots, 0\} = \emptyset$, so it is empty, and the empty function is a choice function (Example 1). *Inductive step*: assume $P(n)$, and let I have $n + 1$ elements. Then $I \neq \emptyset$, since a bijection $\{1, \dots, n + 1\} \rightarrow I$ has a value at 1; pick $i_0 \in I$ by $\exists E$. The set $I' := I \setminus \{i_0\}$ has n elements ([Corollary 3.1.7](#)), so $P(n)$ gives a choice

function c' for $(A_i)_{i \in I'}$. Since $A_{i_0} \neq \emptyset$, one more $\exists E$ names $a_0 \in A_{i_0}$. The graph of c' together with the pair (i_0, a_0) , with codomain $\bigcup_{i \in I} A_i$, is a function on $I' \cup \{i_0\} = I$, because $i_0 \notin I'$; and it is a choice function. So $P(n+1)$ holds, and Logic and Proof, Axiom 4.2.1 gives $P(n)$ for every $n \in \mathbb{N}$. Each step uses two $\exists E$'s and no new principle; the induction says nothing about an infinite I , which has no number of elements.

A rule that picks. If one description names an element of each A_i , the choice function is defined by specification (W1), as the set of pairs (i, a) with a the element named for i . Typical rules: the **least element** of a non-empty set of natural numbers (Theorem 2.1.3); the **non-negative** square root of a positive real; in a bisection, **the left half** if it has the property in question, and otherwise the right half. The first rule is how Theorem 3.2.2 passed from a surjection $\mathbb{N} \rightarrow A$ to an injection $A \rightarrow \mathbb{N}$ with no axiom: the least preimage is named, not chosen.

Warning. **Choice is not “picking an element”.** “Let $a \in A$ ” is a single $\exists E$ and needs no axiom; neither does picking from ten sets. The axiom is needed only for picking **one element from each of infinitely many sets, with no rule**. A proof that says “choose” need not use it, and one that never says “choose” may: “for each n , let g_n be a list of A_n ” is a use, however it is worded.

3.3.5 Right inverses

The first consequence settles the question Section 1.5 left open. By Proposition 1.5.2, a function with a right inverse is surjective; the converse is where choice enters.

Theorem 3.3.2. Surjections Have Right Inverses

Let $f: X \rightarrow Y$ be a surjective function. Then f has a right inverse: there is a function $g: Y \rightarrow X$ with $f(g(y)) = y$ for every $y \in Y$.

Idea. A right inverse must name a preimage of every y . The preimages of the points of Y form a family of sets indexed by Y , each non-empty because f is onto. A choice function for that family is exactly a right inverse.

Proof. For $y \in Y$, let $P_y := f^{-1}(\{y\}) = \{x \in X : f(x) = y\}$ (Definition 1.6.2). This defines a family $(P_y)_{y \in Y}$ of subsets of X , indexed by Y . Each P_y is non-empty, since f is surjective (Definition 1.4.4). By Axiom 3.3.1 there is a function $c: Y \rightarrow \bigcup_{y \in Y} P_y$ with $c(y) \in P_y$ for every $y \in Y$. The union is a subset of X , so the graph of c is a subset of $Y \times X$, and with codomain X it is a function $g: Y \rightarrow X$ with $g(y) = c(y)$ (Definition 1.4.1). For every $y \in Y$, $g(y) \in P_y$, that is, $f(g(y)) = y$. So $f \circ g = \text{id}_Y$, and g is a right inverse of f (Definition 1.5.1 (b)).

■

Together with Proposition 1.5.2: a function is surjective **if and only if** it has a right inverse. The theorem is in fact equivalent to the axiom: a right inverse of the projection $(i, a) \mapsto i$, from the set of pairs with $a \in A_i$ onto I , is a choice function in disguise.

Quick check. Let $f: \mathbb{Z} \rightarrow \mathbb{N}$, $f(k) := |k|$. It is surjective. Does the existence of a right inverse of f need [Axiom 3.3.1](#)?

Solution. No. Each $n \in \mathbb{N}$ has the preimages n and $-n$, and the rule “take the non-negative one” names one of them: $g(n) := n$. Then $f(g(n)) = |n| = n$, so $g: \mathbb{N} \rightarrow \mathbb{Z}$ is a right inverse, defined by a formula. The axiom is needed only when no such rule is available.

3.3.6 Countable unions

Now the theorem the section opened with. The second form, with any countable index set, is the one used in practice.

Theorem 3.3.3. Countable Unions of Countable Sets

- (a) Let $(A_n)_{n \in \mathbb{N}}$ be a family of sets such that A_n is countable for every $n \in \mathbb{N}$. Then $\bigcup_{n \in \mathbb{N}} A_n$ is countable.
- (b) Let I be a countable set and $(A_i)_{i \in I}$ a family of sets such that A_i is countable for every $i \in I$. Then $\bigcup_{i \in I} A_i$ is countable.

Idea. This is the grid of the opening, with the lists supplied by the axiom. ① Throw away the empty sets: the axiom needs non-empty ones, and they add nothing to the union. ② For each remaining n , the surjections $\mathbb{N} \rightarrow A_n$ form a non-empty set; choose one from each, all at once. ③ The chosen lists turn the grid $J \times \mathbb{N}$ into a surjection onto the union, and the grid is countable by Section 3.2. Part (b) is part (a) after renumbering the index set by a list.

Before reading on: which line uses choice, and why can $\exists E$ not replace it?

Proof.

- (a) Let $U := \bigcup_{n \in \mathbb{N}} A_n$ and $J := \{n \in \mathbb{N} : A_n \neq \emptyset\}$.
If $J = \emptyset$, every A_n is empty, so $U = \emptyset$, which is countable by [Definition 3.2.1](#).
Suppose $J \neq \emptyset$. For $n \in J$, let

$$S_n := \{h \in A_n^{\mathbb{N}} : h \text{ is surjective}\},$$

the set of surjections $\mathbb{N} \rightarrow A_n$ ([Definition 1.1.3](#), with $A_n^{\mathbb{N}}$ the set of functions of Section 1.4). Since A_n is countable and non-empty, $S_n \neq \emptyset$ ([Definition 3.2.1](#)). By [Axiom 3.3.1](#), applied to the family $(S_n)_{n \in J}$, there is a function $c: J \rightarrow \bigcup_{n \in J} S_n$ with $c(n) \in S_n$ for every $n \in J$. This is the only line of the proof that uses the axiom.

Define $F: J \times \mathbb{N} \rightarrow U$ by $F(n, m) := c(n)(m)$. It takes values in U : $c(n)$ maps into $A_n \subseteq U$. It is surjective: let $x \in U$; then $x \in A_n$ for some $n \in \mathbb{N}$ ([Definition 1.2.4](#)), so $A_n \neq \emptyset$ and $n \in J$; since $c(n)$ is a surjection onto A_n , there is $m \in \mathbb{N}$ with $c(n)(m) = x$, that is, $F(n, m) = x$.

The set J is countable by [Proposition 3.2.3](#), as a subset of $\mathbb{N}$, which is countable (the identity is a surjection). So $J \times \mathbb{N}$ is countable by [Corollary 3.2.7](#) (a), and it is non-

empty, since it contains $(n_0, 0)$ for any $n_0 \in J$. By [Definition 3.2.1](#) there is a surjection $s: \mathbb{N} \rightarrow J \times \mathbb{N}$. Then $F \circ s: \mathbb{N} \rightarrow U$ is surjective ([Theorem 1.4.4 \(b\)](#)), and U is countable.

- (b) Let $U := \bigcup_{i \in I} A_i$. If $I = \emptyset$, then $U = \emptyset$, which is countable. Otherwise [Definition 3.2.1](#) gives a surjection $s: \mathbb{N} \rightarrow I$; fix one. For $n \in \mathbb{N}$ put $B_n := A_{s(n)}$, a countable set. Then $\bigcup_{n \in \mathbb{N}} B_n = U$. ($\subseteq$) An element of B_n lies in $A_{s(n)}$ with $s(n) \in I$. ($\supseteq$) If $x \in A_i$ with $i \in I$, then $i = s(n)$ for some $n \in \mathbb{N}$, because s is onto, and $x \in B_n$. By (a), U is countable. ■

The line in bold is the only use of the axiom. The surjections s of parts (a) and (b) are single objects, each named by one $\exists E$. What needed the axiom was a surjection **for every** $n \in J$ at once: $\exists E$ supplies $c(n)$ only inside the box of a fixed n , and F needs all of them outside every such box. The set J is there because for $A_n = \emptyset$ there is no function $\mathbb{N} \rightarrow A_n$ at all ([Proposition 1.4.2 \(b\)](#)), so $S_n = \emptyset$. The theorem completes the closure properties of countable sets: subsets, images, finite products, countable unions.

Neither hypothesis can be dropped: Section 3.4 shows that $\mathbb{R}$ is not countable, and $\mathbb{R}$ is the union of the single set $\mathbb{R}$, and also of the countable sets $\{x\}$, $x \in \mathbb{R}$.

Warning. The choice is of the lists, not of preimages. A surjection $\mathbb{N} \rightarrow A$ gives least preimages by a rule ([Theorem 3.2.2](#)), but only once the surjection is **in hand**. In [Theorem 3.3.3](#), what is missing is the surjection itself, one for each A_n , and the sets A_n need not be sets of natural numbers, so no “least” is available to name one.

3.3.7 Unions without choice

Many countable unions met in practice come with their lists, and then no axiom is used.

Example. Unions that come with their lists

- (a) Show that $\mathbb{Q} = \bigcup_{n \in \mathbb{Z}_{>0}} Q_n$, where $Q_n := \{m/n : m \in \mathbb{Z}\}$, and list the union without using [Axiom 3.3.1](#).
 (b) Does the proof in [Theorem 3.2.8](#) that $\mathbb{Q}$ is countable use [Axiom 3.3.1](#)?

Solution.

- (a) By the definition $\mathbb{Q} := \{m/n : m \in \mathbb{Z}, n \in \mathbb{Z}_{>0}\}$, a number is rational exactly when it lies in Q_n for some $n \in \mathbb{Z}_{>0}$, so the two sets are equal. Fix the bijection $z: \mathbb{N} \rightarrow \mathbb{Z}$ of [Theorem 3.2.8](#) ($z(2k) = k$, $z(2k+1) = -(k+1)$). Then $k \mapsto z(k)/n$ lists Q_n , and the whole family of lists is the single function

$$L: \mathbb{Z}_{>0} \times \mathbb{N} \rightarrow \mathbb{Q}, \quad L(n, k) := z(k)/n,$$

defined by specification. It is onto $\mathbb{Q}$: for m/n with $m \in \mathbb{Z}$, $n \in \mathbb{Z}_{>0}$, take k with $z(k) = m$. Since $\mathbb{Z}_{>0} \times \mathbb{N}$ is countable and non-empty ([Proposition 3.2.3](#), [Corollary 3.2.7](#)), composing L with a surjection $\mathbb{N} \rightarrow \mathbb{Z}_{>0} \times \mathbb{N}$ lists $\mathbb{Q}$. The grid is the same as in [Theorem 3.3.3](#); the difference is that the rows were written down, not chosen.

- (b) No, and its place before [Axiom 3.3.1](#) does not settle this, since by the warning in “When choice is not needed” a proof can use the axiom without naming it; we check the lines. The maps z and $(m, n) \mapsto m/n$ are formulas. The results cited name finitely many objects by $\exists E$: one injection in [Proposition 3.2.3](#), two in [Corollary 3.2.7](#) (a), one surjection in [Proposition 3.2.4](#). The passage from a surjection to an injection in [Theorem 3.2.2](#) takes the least preimage, a rule; the return from an injection to a list goes through [Lemma 3.2.1](#), whose list is defined by recursion from least elements ([Theorem 2.1.3](#)), again a rule. No line picks from infinitely many sets without a rule.

The pattern is general: if one formula in n lists every A_n , the family of lists is a single function, and no axiom is used. Choice enters only when each A_n is known to be countable as a bare existence statement.

3.3.8 Algebraic numbers

The theorem pays off at once. A real number is called **algebraic** if it is a root of a non-zero polynomial with integer coefficients. To keep this inside what the book has defined, we speak of coefficient tuples. For $d \geq 1$ and $c = (c_0, c_1, \dots, c_d) \in \mathbb{Z}^{d+1}$ (a $(d+1)$ -tuple, entries numbered from 0), let

$$p_c(x) := \sum_{k=0}^d c_k x^k \quad (x \in \mathbb{R})$$

([Definition 2.3.1](#), [Definition 2.3.2](#)), and call c **non-zero** if $c_k \neq 0$ for some k . Then $x \in \mathbb{R}$ is **algebraic** if $p_c(x) = 0$ for some $d \geq 1$ and some non-zero $c \in \mathbb{Z}^{d+1}$. Constant polynomials are left out without loss, since a non-zero constant has no roots.

For example, $\sqrt{2}$ is algebraic, a root of $x^2 - 2$ ($c = (-2, 0, 1)$). Every rational m/n is algebraic, a root of $nx - m$ ($c = (-m, n)$). So is $\sqrt{2} + \sqrt{3}$, a root of $x^4 - 10x^2 + 1$.

We use one fact from school algebra, not proved here: **if $c \in \mathbb{Z}^{d+1}$ is non-zero, then p_c has at most d real roots.** (It follows from the factor theorem.) In particular the set $R(c) := \{x \in \mathbb{R} : p_c(x) = 0\}$ is finite.

Theorem 3.3.4. The Algebraic Numbers Are Countable

The set of algebraic real numbers is countable.

Idea. Sort the polynomials by degree. For each d , the coefficient tuples form a countable set, and each polynomial contributes finitely many roots: a countable union of finite sets. Then take the union over d . Both unions are instances of [Theorem 3.3.3](#) (b).

Proof. Let Alg be the set of algebraic real numbers, and for $d \geq 1$ let Z_d be the set of non-zero tuples in $\mathbb{Z}^{d+1}$.

Step 1. Let $d \geq 1$. The set $\mathbb{Z}$ is countable ([Theorem 3.2.8](#)), and $\mathbb{Z}^{d+1}$ is the product of $d+1 \geq 2$ factors all equal to $\mathbb{Z}$, so it is countable by [Corollary 3.2.7](#) (b). Hence $Z_d \subseteq \mathbb{Z}^{d+1}$ is countable ([Proposition 3.2.3](#)).

Step 2. For $c \in Z_d$, the root set $R(c)$ is finite by the school fact above, hence countable ([The-](#)

orem 3.2.2, (iii) $\Rightarrow$ (i)). So $\text{Alg}_d := \bigcup_{c \in Z_d} R(c)$, the set of roots of the non-zero polynomials of degree at most d , is countable by [Theorem 3.3.3](#) (b), with the countable index set Z_d .

Step 3. By definition, $\text{Alg} = \bigcup_{d \in \mathbb{Z}_{>0}} \text{Alg}_d$. The index set $\mathbb{Z}_{>0} \subseteq \mathbb{N}$ is countable ([Proposition 3.2.3](#)). So Alg is countable by [Theorem 3.3.3](#) (b). ■

The proof cites [Theorem 3.3.3](#) twice, and so it uses [Axiom 3.3.1](#). Here the axiom can be avoided (not proved in the text; nothing later uses it): list each finite root set in increasing order, which is a rule, and replace the two unions by one fixed enumeration of all integer coefficient tuples. Exercise C2 asks you to carry this out. The countability of Alg has a striking consequence once we know that $\mathbb{R}$ is not countable: Section 3.4 deduces that some real numbers are not algebraic.

3.3.9 Exercises

3.3.9.1 A. Check your understanding

Exercise. A1

These parts are about the proof of [Theorem 3.3.3](#).

- Which line of the proof of part (a) uses [Axiom 3.3.1](#)? What are the index set I and the sets A_i of the axiom at that line?
- Why does the proof introduce the set J ?
- Run the proof on $A_n := \{n, n+1\}$ for $n \in \mathbb{N}$, with the lists $c(n)(m) := n$ for odd m and $c(n)(m) := n+1$ for even m . What is J , and what is $F(3, 2)$?
- State the idea of the proof in two sentences.
- Is [Axiom 3.3.1](#) needed to show that $\bigcup_{n \in \mathbb{N}} A_n$ is countable when every A_n is a subset of $\mathbb{N}$? Justify your answer.

Solution.

- The line “By [Axiom 3.3.1](#), applied to the family $(S_n)_{n \in J}$, there is a function c on J with $c(n) \in S_n$ ”. There $I = J$, the set of indices with $A_n \neq \emptyset$, and the sets of the axiom are the S_n , the sets of surjections $\mathbb{N} \rightarrow A_n$. The axiom chooses lists, not elements of the A_n .
- The axiom applies only to families of **non-empty** sets. If $A_n = \emptyset$, there is no function $\mathbb{N} \rightarrow A_n$ at all ([Proposition 1.4.2](#) (b)), so $S_n = \emptyset$, and the family $(S_n)_{n \in \mathbb{N}}$ would not qualify. Empty sets add nothing to the union, so restricting to J loses nothing.
- Every A_n is non-empty, so $J = \mathbb{N}$. The pair $(3, 2)$ has $m = 2$ even, so $F(3, 2) = c(3)(2) = 3 + 1 = 4$. (Each $c(n)$ is onto $\{n, n+1\}$: $c(n)(1) = n$ and $c(n)(0) = n+1$.)
- Choose, with the axiom, a list of each non-empty A_n , all at once. The lists turn the countable grid $J \times \mathbb{N}$ into a surjection onto the union, so the union is countable.
- No. The union is a subset of $\mathbb{N}$, so it is countable by [Proposition 3.2.3](#). (Even the proof can run without the axiom: $h_A(m) := m$ for $m \in A$ and $h_A(m) := \min A$ otherwise ([Theorem 2.1.3](#)) is a rule that maps $\mathbb{N}$ onto any non-empty $A \subseteq \mathbb{N}$.)

3.3.9.2 B. Practice

Exercise. B1: Which constructions need choice?

Determine which of the following constructions need [Axiom 3.3.1](#). Justify your answer.

- (a) Choose a real number in each interval $(n, n + 1)$, $n \in \mathbb{N}$.
- (b) Choose an element of each non-empty subset of $\mathbb{N}$.
- (c) Choose an element of each non-empty subset of $\mathbb{R}$.
- (d) Choose one of the two square roots of each positive real number.

Solution.

- (a) No. The rule $n \mapsto n + \frac{1}{2}$ names an element of each interval, since $n < n + \frac{1}{2} < n + 1$.
- (b) No. The rule “the least element” names one, by [Theorem 2.1.3](#): $S \mapsto \min S$ is a choice function, defined by specification.
- (c) Yes, as far as this book is concerned. The rule “the least element” fails, since $(0, 1)$ has no least element, and no rule is known that names an element of every non-empty set of reals. A proof that needs such a function uses the axiom, and says so.
- (d) No. The rule “the positive one” names one root of each $y > 0$: $y \mapsto \sqrt{y}$.

Exercise. B2: Finite sequences

Prove that the set of all finite sequences of natural numbers is countable, and say whether your proof uses [Axiom 3.3.1](#). Here a finite sequence of length $k \in \mathbb{N}$ is a function $\{1, \dots, k\} \rightarrow \mathbb{N}$, represented by its graph, a subset of $\mathbb{Z}_{>0} \times \mathbb{N}$; the sequence of length 0 is $\emptyset$.

Hint: [Theorem 2.3.1](#), together with one fixed surjection $\mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$.

Solution. Let Seq be the set of $x \in \mathcal{P}(\mathbb{Z}_{>0} \times \mathbb{N})$ such that x is the graph of a function $\{1, \dots, k\} \rightarrow \mathbb{N}$ for some $k \in \mathbb{N}$ ([Definition 1.1.3](#), [Definition 1.2.2](#)). This k is determined by x , since $\{1, \dots, k\}$ is the set of first coordinates of x , and k is its number of elements ([Corollary 3.1.4](#)); call it the **length** of x , and let Seq_k be the set of elements of length k . So $\text{Seq} = \bigcup_{k \in \mathbb{N}} \text{Seq}_k$, and $\text{Seq}_0 = \{\emptyset\}$.

With choice. Each Seq_k is countable (for $k \geq 2$ it is in bijection with $\mathbb{N}^k$, by $x \mapsto (x(1), \dots, x(k))$; Seq_1 with $\mathbb{N}$; Seq_0 has one element), so [Theorem 3.3.3](#) (a) applies. This route uses [Axiom 3.3.1](#), through the theorem.

Without choice. Fix one surjection $\sigma: \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$ ([Theorem 3.2.6](#), [Definition 3.2.1](#); a single $\exists E$), and write $\sigma(n) = (\sigma_1(n), \sigma_2(n))$. For $x \in \text{Seq}$ of length k and $m \in \mathbb{N}$, let $x \frown m := x \cup \{(k+1, m)\}$, the sequence x with m appended; it has length $k+1$. Every $y \in \text{Seq}_{k+1}$ is $x \frown m$ for x the restriction of y to $\{1, \dots, k\}$ and $m = y(k+1)$.

Let $X := \text{Seq}^{\mathbb{N}}$, the set of functions $\mathbb{N} \rightarrow \text{Seq}$, let $a \in X$ be the constant function with value $\emptyset$, and define $g: \mathbb{N} \times X \rightarrow X$ by

$$g(k, s)(n) := s(\sigma_1(n)) \frown \sigma_2(n).$$

By [Theorem 2.3.1](#) there is a function $k \mapsto s_k$ from $\mathbb{N}$ to X with $s_0 = a$ and $s_{k+1}(n) = s_k(\sigma_1(n)) \frown \sigma_2(n)$ for all $k, n \in \mathbb{N}$.

Claim: for every $k \in \mathbb{N}$, s_k maps $\mathbb{N}$ onto Seq_k . By Logic and Proof, Axiom 4.2.1 on k . For $k = 0$, every value of s_0 is $\emptyset$, and $\text{Seq}_0 = \{\emptyset\}$. Assume the claim for k . Each value $s_{k+1}(n) = s_k(\sigma_1(n)) \frown \sigma_2(n)$ has length $k + 1$, because $s_k(\sigma_1(n))$ has length k . Let $y \in \text{Seq}_{k+1}$, say $y = x \frown m$ with $x \in \text{Seq}_k$. By the hypothesis there is j with $s_k(j) = x$, and since σ is onto there is n with $\sigma(n) = (j, m)$. Then $s_{k+1}(n) = s_k(j) \frown m = y$. This proves the claim.

Now $n \mapsto s_{\sigma_1(n)}(\sigma_2(n))$ maps $\mathbb{N}$ onto Seq : for $y \in \text{Seq}_k$, the claim gives j with $s_k(j) = y$, and some n has $\sigma(n) = (k, j)$. So Seq is countable. One surjection was named, and the lists s_k were defined from it by recursion: no choice is used. Either route earns full credit if the use of choice is reported correctly.

Exercise. B3

A **rational point** of the plane is a point of $\mathbb{Q} \times \mathbb{Q}$. Prove that the set of lines in the plane that pass through two different rational points is countable. Does your proof use [Axiom 3.3.1](#)?

Solution. Let $D := \{(p, q) \in (\mathbb{Q} \times \mathbb{Q}) \times (\mathbb{Q} \times \mathbb{Q}) : p \neq q\}$. The set $\mathbb{Q} \times \mathbb{Q}$ is countable by [Theorem 3.2.8](#) and [Corollary 3.2.7](#); so is its product with itself, again by [Corollary 3.2.7](#), and so is the subset D ([Proposition 3.2.3](#)). For $(p, q) \in D$, let $\text{line}(p, q)$ be the line through p and q , as a subset of $\mathbb{R}^2$: $\text{line}(p, q) := \{p + t(q - p) : t \in \mathbb{R}\}$. Two different points lie on exactly one line (school geometry), so the lines in question form the image $\text{line}(D)$, countable by [Proposition 3.2.4](#). The map line is a formula, so no choice is used.

3.3.9.3 C. Going deeper

Exercise. C1: Exit check: countable unions

Prove that if A_n is countable for every $n \in \mathbb{N}$, then so is $\bigcup_{n \in \mathbb{N}} A_n$. Point to the step of your proof that uses [Axiom 3.3.1](#).

Solution. Let $U := \bigcup_{n \in \mathbb{N}} A_n$ and $J := \{n \in \mathbb{N} : A_n \neq \emptyset\}$. If $J = \emptyset$, then $U = \emptyset$, which is countable. Let $J \neq \emptyset$. For $n \in J$, the set S_n of surjections $\mathbb{N} \rightarrow A_n$ is non-empty ([Definition 3.2.1](#)). By [Axiom 3.3.1](#), applied to $(S_n)_{n \in J}$, there is a function c on J with $c(n) \in S_n$ for every $n \in J$: a list of every A_n at once, which no finite number of $\exists E$'s gives. This is the step that uses the axiom.

Put $F(n, m) := c(n)(m) \in A_n \subseteq U$ for $(n, m) \in J \times \mathbb{N}$. It is onto U : if $x \in U$, then $x \in A_n$ for some $n \in \mathbb{N}$ ([Definition 1.2.4](#)), so $n \in J$, and $c(n)(m) = x$ for some m , as $c(n)$ is onto A_n . Since $J \subseteq \mathbb{N}$ is countable ([Proposition 3.2.3](#)), $J \times \mathbb{N}$ is countable ([Corollary 3.2.7](#) (a)) and non-empty, so some $s: \mathbb{N} \rightarrow J \times \mathbb{N}$ is onto ([Definition 3.2.1](#)), and so is $F \circ s: \mathbb{N} \rightarrow U$ ([Theorem 1.4.4](#) (b)). This is the proof of [Theorem 3.3.3](#) (a).

Marking scheme. (10 points) - 1: handles the empty sets: the case where every A_n is empty, and the restriction to the non-empty ones before choosing. - 3: obtains, by [Axiom 3.3.1](#), a surjection onto A_n for every n simultaneously, and says at that line that the axiom is used.

- 2: defines a map from a set of pairs (or another index set) to the union. - 2: shows that this map is onto the union. - 2: shows that the index set is countable and non-empty, citing the product result, and concludes. Through injections $\iota_n: A_n \rightarrow \mathbb{N}$: 1 for noting that they exist for every n , empty A_n included (the empty function), so nothing is discarded; 3 for choosing all ι_n by [Axiom 3.3.1](#), flagged; 2 for $x \mapsto (n_x, \iota_{n_x}(x))$ into $\mathbb{N} \times \mathbb{N}$, with n_x the least n with $x \in A_n$ ([Theorem 2.1.3](#)); 2 for its injectivity; 2 for concluding with [Theorem 3.2.6](#) and [Theorem 3.2.2](#), (ii) $\Rightarrow$ (i). Fatal errors (0 for the step): “for each n let f_n be a list of A_n ” with no mention of choice, an unflagged infinite $\exists E$ that no rule licenses (0 for the choice step); listing $\mathbb{N} \times \mathbb{N}$ row by row, as in [Example \(Listing the rows one after another\)](#) (0 for the countability step).

Exercise. C2: Algebraic numbers without choice

Prove [Theorem 3.3.4](#) without using [Axiom 3.3.1](#).

Solution (Hint 1). Use [Exercise \(B2: Finite sequences\)](#) to obtain, without choice, one surjection from $\mathbb{N}$ onto all coefficient tuples of all lengths together.

Solution (Hint 2). For an algebraic r , let $n(r)$ be the least index whose tuple has r as a root; it exists by [Theorem 2.1.3](#).

Solution. Use the notation of [Theorem 3.3.4](#), and let Alg be the set of algebraic numbers.

One list of all coefficient tuples. Let ZSeq be defined as Seq in [Exercise \(B2: Finite sequences\)](#), with $\mathbb{Z}$ in place of $\mathbb{N}$, and let $z: \mathbb{N} \rightarrow \mathbb{Z}$ be the bijection of [Theorem 3.2.8](#). Sending $x \in \text{Seq}$ to the sequence $i \mapsto z(x(i))$ maps Seq onto ZSeq (use z^{-1} entrywise to find a preimage), so ZSeq is countable without choice, by that exercise and [Proposition 3.2.4](#). Since ZSeq is non-empty (it contains $\emptyset$, the sequence of length 0), fix one surjection $\mu: \mathbb{N} \rightarrow \text{ZSeq}$ ([Definition 3.2.1](#)), a single $\exists E$. A sequence y of length $d + 1 \geq 2$ is read as the tuple $c = (y(1), \dots, y(d + 1)) \in \mathbb{Z}^{d+1}$, and every such tuple arises; write $p_y := p_c$ and $R(y) := R(c)$. Call $n \in \mathbb{N}$ **good for** $r \in \mathbb{R}$ if $\mu(n)$ has length at least 2, is non-zero, and $p_{\mu(n)}(r) = 0$. Then r is algebraic if and only if some n is good for r , since μ is onto.

An injection, by rules. For $r \in \text{Alg}$, let $n(r)$ be the **least** $n \in \mathbb{N}$ that is good for r ([Theorem 2.1.3](#)), and let

$$\rho(r) := \text{the number of elements of } \{s \in R(\mu(n(r))) : s < r\},$$

well defined because this set is a subset of the finite set $R(\mu(n(r)))$ ([Proposition 3.1.6](#), [Corollary 3.1.4](#)). Both are named by rules, so $r \mapsto (n(r), \rho(r))$ is a function $\text{Alg} \rightarrow \mathbb{N} \times \mathbb{N}$, defined by specification.

It is injective. Let $r, r' \in \text{Alg}$ with $n(r) = n(r')$ and $\rho(r) = \rho(r')$; call this common index n , and suppose $r \neq r'$, say $r < r'$. Both are roots of $p_{\mu(n)}$. The set $\{s \in R(\mu(n)) : s < r\}$ is a subset of $\{s \in R(\mu(n)) : s < r'\}$, and a proper one, since r lies in the second and not in the first. By [Proposition 3.1.6](#), the first has fewer elements: $\rho(r) < \rho(r')$, a contradiction. So $r = r'$.

Following it with the injection $(m, k) \mapsto 2^m 3^k$ of [Theorem 3.2.6](#) gives an injection $\text{Alg} \rightarrow \mathbb{N}$ ([Theorem 1.4.4](#) (a)), so Alg is countable by [Theorem 3.2.2](#), (ii) $\Rightarrow$ (i). Apart from the single surjection μ , every object was named by a rule, so [Axiom 3.3.1](#) is not used.

Marking scheme. (10 points) - 3: produces, without choice, one surjection (or injection) relating $\mathbb{N}$ to the coefficient tuples of **all** degrees at once, citing [Exercise \(B2: Finite sequences\)](#) or an explicit coding. - 3: assigns to each algebraic number its data (an index of a polynomial and a position among its roots) by **rules**, such as a least index and a count or the increasing order of the finite root set. - 4: proves that the resulting map is injective (or a map onto Alg surjective), concludes, and says why no step uses choice. Fatal errors (0 for the step): citing [Theorem 3.3.3](#), which uses the axiom; “for each polynomial (or each d), take a list” with no rule given.

Exercise. C3: Finite unions

- (a) Prove, without [Axiom 3.3.1](#), that for every finite set I and every family $(A_i)_{i \in I}$ of countable sets, $\bigcup_{i \in I} A_i$ is countable.
- (b) Explain in two sentences why your induction does not prove [Theorem 3.3.3](#).

Solution.

- (a) We use Logic and Proof, [Axiom 4.2.1](#) on the number of elements of I . Let $P(n)$ be: “for every set I with n elements and every family $(A_i)_{i \in I}$ of countable sets, $\bigcup_{i \in I} A_i$ is countable”.

Base case. A set I with 0 elements is empty, so $\bigcup_{i \in I} A_i = \emptyset$, which is countable.

Inductive step. Assume $P(n)$, and let I have $n + 1$ elements and $(A_i)_{i \in I}$ be a family of countable sets. Since $n + 1 \geq 1$, $I \neq \emptyset$; pick $i_0 \in I$. Then $I' := I \setminus \{i_0\}$ has n elements ([Corollary 3.1.7](#)), so $B := \bigcup_{i \in I'} A_i$ is countable by $P(n)$, applied to the family $(A_i)_{i \in I'}$. Since $I = I' \cup \{i_0\}$, an element lies in some A_i with $i \in I$ exactly when it lies in B or in A_{i_0} , so

$$\bigcup_{i \in I} A_i = B \cup A_{i_0},$$

which is countable by [Proposition 3.2.5](#). So $P(n + 1)$ holds, and Logic and Proof, [Axiom 4.2.1](#) gives $P(n)$ for every n . Each step makes finitely many $\exists E$'s, so no choice is used.

- (b) The induction proves the statement for each finite number n of sets, and $\mathbb{N}$ is not finite ([Example \(Hilbert's Hotel\)](#)), so no instance of $P(n)$ is about the family $(A_n)_{n \in \mathbb{N}}$. Writing $\bigcup_{n \in \mathbb{N}} A_n$ as the union of the finite unions $A_0 \cup \dots \cup A_k$ does not help, since that needs a list of every finite union at once, which is a choice of infinitely many lists again.

3.4 Diagonal Arguments: Uncountable Sets and Cantor's Theorem

Every infinite set met so far has turned out to be countable: $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, finite products of countable sets, countable unions of them ([Theorem 3.3.3](#)), and even the algebraic numbers ([Theorem 3.3.4](#)).

Is every infinite set countable? This section shows that the answer is no, by one argument used three times. Given any list, it builds an object that the list misses. The argument proves that the set of 0–1 sequences is uncountable, that no set can be mapped onto the set of its subsets (Cantor's theorem), and, with decimals, that $\mathbb{R}$ is uncountable.

3.4.1 A sequence that is not on the list

Recall from Section 1.4 that $\{0, 1\}^{\mathbb{N}}$ is the set of all functions $\mathbb{N} \rightarrow \{0, 1\}$: the sequences of zeros and ones, indexed from 0. It is infinite, and it looks no bigger than $\mathbb{N} \times \mathbb{N}$, which is countable. To show that it is **not** countable we must show, by [Definition 3.2.1](#), that it is non-empty and that **no** function $f: \mathbb{N} \rightarrow \{0, 1\}^{\mathbb{N}}$ is onto. So we take an arbitrary list $f(0), f(1), f(2), \dots$ of sequences and look for a sequence that is not on it.

Write the list as a table: row n is the sequence $f(n)$, and the entry in row n and column k is the number $f(n)(k)$. Here is the start of one such table, with the entries in places $(0, 0), (1, 1), (2, 2), (3, 3)$ boxed:

place	0	1	2	3	4	...
$f(0)$	0	1	0	1	0	...
$f(1)$	1	1	1	1	1	...
$f(2)$	0	0	0	0	0	...
$f(3)$	1	0	0	1	1	...
$\vdots$						
g	1	0	1	0		...

The boxed entries run down the **diagonal** of the table. The bottom row g is the diagonal with every entry flipped: $g(n) = 1 - f(n)(n)$. It cannot be row 0, because it differs from $f(0)$ in place 0; it cannot be row 1, because it differs from $f(1)$ in place 1; and so on down the table. Nothing in this reasoning used the particular entries.

Theorem 3.4.1. The 0–1 Sequences Are Uncountable

The set $\{0, 1\}^{\mathbb{N}}$ of all sequences of zeros and ones is uncountable.

Proof. Let $f: \mathbb{N} \rightarrow \{0, 1\}^{\mathbb{N}}$ be any function. Define $g: \mathbb{N} \rightarrow \{0, 1\}$ by

$$g(n) := 1 - f(n)(n) \quad (n \in \mathbb{N});$$

its values lie in $\{0, 1\}$, since $1 - 0 = 1$ and $1 - 1 = 0$. So $g \in \{0, 1\}^{\mathbb{N}}$. For every $n \in \mathbb{N}$, $g(n) \neq f(n)(n)$, since no $b \in \{0, 1\}$ has $1 - b = b$. Hence $g \neq f(n)$: if $g = f(n)$, then $g(n) = f(n)(n)$, which is false; the two functions differ at place n [=E, $\neg$ I; [Proposition 1.4.1](#)]. So g is not a value of f , and f is not surjective. Since f was arbitrary, no function $\mathbb{N} \rightarrow \{0, 1\}^{\mathbb{N}}$ is surjective. The set $\{0, 1\}^{\mathbb{N}}$ is non-empty, since it contains the zero sequence. By [Definition 3.2.1](#), $\{0, 1\}^{\mathbb{N}}$ is not countable. ■

This is the move of the section, and it has a name.

The diagonal. *Given any list, build an object that differs from the n -th entry at the n -th place. The new object is not on the list, because it disagrees with every entry somewhere, and the place of disagreement is chosen by the entry's own position.*

Quick check. Run the proof on the list $f(n) := \mathbf{1}_{\{n\}}$, the sequence with a 1 in place n and 0 elsewhere. What is g ? Is it on the list?

Solution. Here $f(n)(n) = \mathbf{1}_{\{n\}}(n) = 1$ for every n , so $g(n) = 1 - 1 = 0$ for every n : g is the zero sequence. It is indeed not on the list, since every $f(n)$ has a 1 in place n .

Warning. **The diagonal object depends on the list.** “The diagonal sequence” is not one sequence: each list has its own, and a different list may well contain the g built from this one (the list $g, f(0), f(1), \dots$ does). And the conclusion “uncountable” needs the list to be **arbitrary**. Running the construction on one list, or on a few, shows only that those lists miss something; concluding “so no list works” from them is the error (N1) of Logic and Proof, Example (What is not allowed). The proof above starts from “let f be **any** function”, and that is what makes its conclusion universal.

3.4.2 Cantor's theorem

The table had a row for each $n \in \mathbb{N}$ and a column for each place $k \in \mathbb{N}$. Nothing about the diagonal needs the rows and the columns to be indexed by $\mathbb{N}$: it needs only that they are indexed by the **same** set. A subset S of a set X is a row with one entry for each $x \in X$, “in” or “out”, the values of $\mathbf{1}_S$. So the argument applies to any set and the set of its subsets.

Theorem 3.4.2. Cantor's Theorem

Let X be a set. Then there is **no** surjection $X \rightarrow \mathcal{P}(X)$.

Try it first (ten minutes): given a function $f: X \rightarrow \mathcal{P}(X)$, which subset of X could be missed by f ? Try $X = \{1, 2, 3\}$ with a function of your own first.

Idea. The first attempts look for a subset that is “too big” or “too small” to be a value, such as X itself or $\emptyset$, and fail: f may well take both values. Size is the wrong idea. A set D differs from $f(x)$ as soon as the two disagree about **one** element, and the only element we know anything about in connection with $f(x)$ is x itself. So we make D disagree with $f(x)$ at x : put x into D exactly when $x \notin f(x)$. This is the diagonal of the table, flipped.

Before reading on: where is surjectivity used?

Proof. Let $f: X \rightarrow \mathcal{P}(X)$ be a function, and let

$$D := \{x \in X : x \notin f(x)\},$$

a subset of X by [Definition 1.1.3](#), so $D \in \mathcal{P}(X)$ ([Definition 1.2.2](#)). Suppose f is surjective. Then there is $d \in X$ with $f(d) = D$ ([Definition 1.4.4](#)). If $d \in D$, then $d \notin f(d)$ by the definition of D , that is, $d \notin D$, a contradiction. Hence $d \notin D$. Then $d \notin f(d)$, and since $d \in X$, the definition of D gives $d \in D$, a contradiction again. So f is not surjective. This proves the theorem. ■

Surjectivity was used once, to obtain d ; everything else holds for every function f . Read without the supposition, the proof shows that $D \neq f(x)$ for **every** $x \in X$: the two sets disagree about x . Three consequences follow.

- (i) *Russell's theorem is the same argument.* In [Theorem 1.1.4](#), the set $R_A = \{x \in A : x \notin x\}$ is the set D built from the rule $f(x) := x$: each $x \in A$ is tested against itself. The proof that D differs from every $f(x)$ becomes the proof that R_A differs from every element of A , that is, $R_A \notin A$. (The rule $x \mapsto x$ need not map into $\mathcal{P}(A)$, but the argument never used that $f(x) \subseteq X$, only the test “ $x \in f(x)$ ”.)
- (ii) *The power set is strictly bigger.* The map $x \mapsto \{x\}$ is an injection $X \rightarrow \mathcal{P}(X)$, since $\{x\} = \{y\}$ gives $x = y$. So $\mathcal{P}(X)$ has at least as many elements as X , in the sense that X is in bijection with a subset of it, and by the theorem it is not equinumerous with X , since a bijection would be a surjection. Section 3.5 names this relation: X is strictly smaller than $\mathcal{P}(X)$.
- (iii) *There is no largest size.* $\mathbb{N}, \mathcal{P}(\mathbb{N}), \mathcal{P}(\mathcal{P}(\mathbb{N})), \dots$: each set in this sequence is strictly bigger than the one before.

Corollary 3.4.3. The Subsets of the Naturals

The power set $\mathcal{P}(\mathbb{N})$ is uncountable.

Proof. The set $\mathcal{P}(\mathbb{N})$ is non-empty, since $\emptyset \in \mathcal{P}(\mathbb{N})$, and by [Theorem 3.4.2](#) with $X = \mathbb{N}$ there is no surjection $\mathbb{N} \rightarrow \mathcal{P}(\mathbb{N})$. By [Definition 3.2.1](#), $\mathcal{P}(\mathbb{N})$ is not countable. ■

The two diagonal proofs are one. By the correspondence of Section 1.4, $S \mapsto \mathbf{1}_S$ is a bijection $\mathcal{P}(\mathbb{N}) \rightarrow \{0, 1\}^{\mathbb{N}}$, with inverse $b \mapsto \{n \in \mathbb{N} : b(n) = 1\}$ ([Theorem 1.5.4](#)). Given $f: \mathbb{N} \rightarrow \mathcal{P}(\mathbb{N})$, the sequence $g(n) = 1 - \mathbf{1}_{f(n)}(n)$ of [Theorem 3.4.1](#) is 1 exactly when $n \notin f(n)$, so $g = \mathbf{1}_D$ for the set D of Cantor's proof.

3.4.3 Decimal expansions

The real numbers should be next. A real number in $[0, 1)$ is written as a decimal $0.d_1d_2d_3\dots$, a sequence of digits, so the diagonal suggests itself: list the numbers, write their decimals as the rows of a table, and change the diagonal digits. Two questions must be answered first. What does it **mean** that a sequence of digits is the decimal of x ? And what is $0.999\dots$? School says it equals 1, and a diagonal that produces $0.4999\dots$ when $0.5000\dots$ is on the list has produced a number that is on the list.

An infinite sum would answer the first question, but this book has not yet defined one. Finite sums are enough. The first n digits of x should pin x down to within 10^{-n} .

A *digit sequence* is a decimal of x when each of its finite truncations is within one unit of its last place below x .

Definition 3.4.1. Decimal Expansion

A **digit sequence** is a function $k \mapsto d_k$ from $\mathbb{Z}_{>0}$ to $\{0, 1, \dots, 9\}$. For $n \geq 1$ let

$$t_n := \sum_{k=1}^n d_k 10^{-k}$$

(Definition 2.3.1), where $10^{-k} := (1/10)^k$ (Definition 2.3.2). The digit sequence is a **decimal expansion** of $x \in [0, 1]$ if **for every** $n \geq 1$,

$$t_n \leq x \leq t_n + 10^{-n}.$$

In words: t_n is the number written with the first n digits, $0.d_1 \dots d_n$, and the condition says that x lies in the closed interval of length 10^{-n} starting at t_n , for every n . The digits are indexed from 1, so that d_k is the digit in the k -th place after the point. No infinite sum appears: every t_n is a finite sum.

Examples.

1. *Zero.* The all-zero sequence has $t_n = 0$, and $0 \leq 0 \leq 10^{-n}$; it is an expansion of 0.
2. *One half, twice.* The sequence $5, 0, 0, \dots$ has $t_n = \frac{1}{2}$ for every $n \geq 1$, and $\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2} + 10^{-n}$. The sequence $4, 9, 9, \dots$ has $t_n = \frac{4}{10} + \sum_{k=2}^n 9 \cdot 10^{-k} = \frac{1}{2} - 10^{-n}$, by Proposition 2.3.6, so $t_n \leq \frac{1}{2} \leq t_n + 10^{-n} = \frac{1}{2}$. Both are expansions of $\frac{1}{2}$.
3. *One.* The all-nine sequence has $t_n = \sum_{k=1}^n 9 \cdot 10^{-k} = 1 - 10^{-n}$ (Proposition 2.3.6), so $t_n \leq 1 \leq t_n + 10^{-n} = 1$: it is an expansion of 1.

A non-example by minimal change. Keep the sequence $5, 0, 0, \dots$ and move x from $\frac{1}{2}$ to 0.51. The inequalities still hold for $n = 1$ ($0.5 \leq 0.51 \leq 0.6$) and for $n = 2$ ($0.50 \leq 0.51 \leq 0.51$), and they fail at $n = 3$: $0.51 > 0.500 + 0.001$. So $5, 0, 0, \dots$ is not an expansion of 0.51; the clause “for every $n \geq 1$ ” is what fails.

Why this definition. Both inequalities are non-strict. With a strict right inequality, $x < t_n + 10^{-n}$, the all-nine sequence would be an expansion of no number at all: it would need $x < 1$ together with $1 - 10^{-n} \leq x$ for every n , and no real number satisfies both, by the Archimedean property (a school fact; *Analysis on the Real Line*, Chapter 1 proves it). The non-strict form is what makes $0.999\dots = 1$ true, and true **by the definition**: the inequalities for $x = 1$ are checked in Example 3, and no other number satisfies them, since $1 - 10^{-n} \leq x \leq 1$ for every n forces $x = 1$, again by the Archimedean property (this is the uniqueness half of (D2) below). Example 2 shows that the price is two expansions for some numbers.

What this book does not prove are the facts that make expansions behave as school says. We state them once and use them in this section and the next, always by name.

- (D1) Every $x \in [0, 1]$ has a decimal expansion that does **not** end in all 9s.
- (D2) Every digit sequence is a decimal expansion of **exactly one** $x \in [0, 1]$.
- (D3) If two **different** digit sequences are expansions of the same x , then one of them ends in all 9s and the other ends in all 0s.

Here a digit sequence **ends in all 9s** if there is $N \geq 1$ with $d_k = 9$ for every $k \geq N$, and likewise for 0s. By the uniqueness in (D2), we may write $x = 0.d_1 d_2 d_3 \dots$ for “ x is the number of which

$d_1, d_2, d_3, \dots$ is an expansion". So $0.4999\dots = 0.5000\dots = \frac{1}{2}$, as in (D3).

Neither (D1) nor (D2) follows from what we have so far. (D3) does, by a finite computation at the first place where the two sequences differ (Theorem 2.1.3, Proposition 2.3.6); we state it with the others to quote it by name. The **existence** half of (D2), that some x satisfies all the inequalities at once, is where the completeness of $\mathbb{R}$ hides, and *Analysis on the Real Line*, Chapter 1 takes completeness as an axiom. The **uniqueness** half of (D2) and the "not ending in all 9s" of (D1) rest on the Archimedean property: two numbers x, x' with $|x - x'| \leq 10^{-n}$ for every n must be equal, which *Analysis on the Real Line*, Chapter 1 also proves. Until then we use them as school facts, and each use names the fact it uses.

One consequence is used below. *A digit sequence with no digit 0 and no digit 9 is the **only** expansion of its number.* Indeed, by (D3) a second expansion of the same number would force one of the two to end in all 9s and the other in all 0s; the given sequence ends in neither, since it has no 9s and no 0s at all.

3.4.4 The real numbers are uncountable

With the facts in hand, the diagonal runs on digits. Two precautions are built in. Each number on the list gets its expansion by a **rule**, so that no choosing is needed. And the new digits avoid 0 and 9, so that the new number has only one expansion.

Theorem 3.4.4. The Reals Are Uncountable, by Decimals

The interval $[0, 1)$ is uncountable. Hence $\mathbb{R}$ is uncountable.

Idea. Take an arbitrary list of numbers in $[0, 1)$, write each as its expansion with no tail of 9s, and change the n -th digit of the n -th number to 4 or 5. The resulting digit sequence is an expansion of some y , by (D2). If y were the n -th number, it would have two different expansions, one with no 0s or 9s, against (D3).

Proof. Let $f: \mathbb{N} \rightarrow [0, 1)$ be any function. For $n \geq 1$ let $x_n := f(n - 1)$, so that the list $x_1, x_2, x_3, \dots$ has the same values as f .

For each $n \geq 1$, x_n has an expansion not ending in all 9s, by (D1), and only one: by (D3), two different such expansions would include one ending in all 9s. Let $(d_{n,k})_{k \geq 1}$ be this expansion. It is named by a rule, "the expansion of x_n not ending in all 9s", so no choice is involved. Define a digit sequence by

$$e_n := \begin{cases} 5 & \text{if } d_{n,n} \neq 5, \\ 4 & \text{if } d_{n,n} = 5, \end{cases} \quad (n \geq 1).$$

Then $e_n \neq d_{n,n}$ for every $n \geq 1$. By (D2), there is $y \in [0, 1]$ of which $(e_k)_{k \geq 1}$ is an expansion. Its first truncation is $t_1 = e_1/10 \in \{0.4, 0.5\}$, so $0.4 \leq t_1 \leq y \leq t_1 + 0.1 \leq 0.6$, and in particular $y \in [0, 1)$.

Let $n \geq 1$, and suppose $y = x_n$. Then (e_k) and $(d_{n,k})$ are both expansions of y , and they are different, since they differ in place n . But (e_k) has no digit 0 and no digit 9, so by the consequence of (D3) stated above it is the **only** expansion of y , a contradiction. So $y \neq x_n$ for every $n \geq 1$, and y is not a value of f . Since f was arbitrary, no function $\mathbb{N} \rightarrow [0, 1)$

is surjective; and $[0, 1)$ is non-empty, since it contains 0. By [Definition 3.2.1](#), $[0, 1)$ is not countable.

Finally, if $\mathbb{R}$ were countable, its subset $[0, 1)$ would be countable by [Proposition 3.2.3](#). So $\mathbb{R}$ is uncountable. ■

Why the digits 4 and 5. The first rule that comes to mind is $e_n := 9 - d_{n,n}$, which also changes every diagonal digit. Run it on the list $x_1 = \frac{1}{2}$, $x_n = 0$ for $n \geq 2$ (that is, $f(0) = \frac{1}{2}$ and $f(n) = 0$ for $n \geq 1$). The expansions are $0.5000\dots$ and $0.000\dots$, so $d_{1,1} = 5$ and $d_{n,n} = 0$ for $n \geq 2$. The naive rule produces $e_1 = 4$ and $e_n = 9$ for $n \geq 2$: the sequence $0.4999\dots$, which is an expansion of $\frac{1}{2} = x_1$ by [Example 2](#). The new digit sequence differs from every row, and its number is on the list. The rule with 4 and 5, on the same list, gives $0.4555\dots = \frac{41}{90}$, which is on no row. Avoiding 0 and 9 is exactly what the last paragraph of the proof uses.

Analysis on the Real Line, Chapter 1 proves that $\mathbb{R}$ is uncountable again, with no decimals at all, directly from its least upper bound axiom; that proof needs none of (D1)–(D3).

3.4.5 Transcendental numbers

A real number that is not algebraic is called **transcendental**. Section 3.3 showed that the set Alg of algebraic numbers is countable ([Theorem 3.3.4](#)).

Corollary 3.4.5. Transcendental Numbers Exist

The set $\mathbb{R} \setminus \text{Alg}$ of transcendental numbers is uncountable. In particular, there is a real number that is not algebraic.

Proof. By excluded middle ([Logic and Proof](#), [Proposition 2.5.3](#)), every real number is in Alg or not, so $\mathbb{R} = \text{Alg} \cup (\mathbb{R} \setminus \text{Alg})$. Suppose $\mathbb{R} \setminus \text{Alg}$ is countable. Since Alg is countable by [Theorem 3.3.4](#), $\mathbb{R}$ is countable by [Proposition 3.2.5](#), contradicting [Theorem 3.4.4](#). So $\mathbb{R} \setminus \text{Alg}$ is uncountable. In particular $\mathbb{R} \setminus \text{Alg} \neq \emptyset$, since $\emptyset$ is countable ([Definition 3.2.1](#)). So not every x satisfies $x \notin \mathbb{R} \setminus \text{Alg}$: otherwise $\mathbb{R} \setminus \text{Alg}$ would be an empty set ([Definition 1.1.4](#)), hence equal to $\emptyset$ ([Proposition 1.1.3](#)). By the classical law [Logic and Proof](#), [Proposition 6.2.1](#) (a) and $\neg\neg E$, some x lies in $\mathbb{R} \setminus \text{Alg}$; this x is a real number that is not algebraic. This last passage, from “not empty” to “has an element”, is a classical step (*Logic and Proof*, [Section 5.7](#)). ■

Remark. The proof names no transcendental number: it shows that the set of them is too large to be countable, hence not empty. Its classical steps are the split of $\mathbb{R}$ into Alg and its complement (excluded middle) and the last one, from “not empty” to “has an element”. Compare the classical proof of [Logic and Proof](#), [Theorem 8.7.1](#) in *Logic and Proof*, [Section 5.7](#), which also names no witness (unlike [Logic and Proof](#), [Proposition 8.7.2](#), which does). Proving that a particular number, such as π , is transcendental is a different and much harder matter.

3.4.6 Exercises

3.4.6.1 A. Check your understanding

Exercise. A1

These parts are about the proof of [Theorem 3.4.2](#).

- (a) Where in the proof is the surjectivity of f used, and what does it supply?
- (b) Run the proof on $X = \{1, 2, 3\}$ and $f(1) = \{1, 2\}$, $f(2) = \emptyset$, $f(3) = \{1, 3\}$. What is D ? Is D a value of f ?
- (c) State the idea of the proof in two sentences.
- (d) Why does the same argument not show that there is no surjection $X \rightarrow X$?

Solution.

- (a) It is used once, right after D is defined: surjectivity supplies an element $d \in X$ with $f(d) = D$. The two cases $d \in D$ and $d \notin D$ are then both refuted. Everything before that line holds for every function f .
- (b) Test each x against $f(x)$: $1 \in f(1)$, so $1 \notin D$; $2 \notin f(2) = \emptyset$, so $2 \in D$; $3 \in f(3)$, so $3 \notin D$. Hence $D = \{2\}$. It is not a value of f : the values are $\{1, 2\}$, $\emptyset$, $\{1, 3\}$, and D differs from $f(x)$ about x in each case ($1 \in f(1) \setminus D$, $2 \in D \setminus f(2)$, $3 \in f(3) \setminus D$).
- (c) Build the set D of those x that are not in their own image $f(x)$. Then D disagrees with every $f(x)$ about the element x , so it is not a value of f .
- (d) The set D is a **subset** of X , so it is a candidate value only for a function into $\mathcal{P}(X)$. For $f: X \rightarrow X$, the argument (where the test $x \in f(x)$ makes sense at all) shows that D is not a value of f , but D need not be an element of X , so nothing contradicts surjectivity. Indeed id_X is a surjection $X \rightarrow X$; with $f = \text{id}_X$ the argument is Russell's, and shows only that $D = R_X$ is not an element of X ([Theorem 1.1.4](#)).

3.4.6.2 B. Practice

Exercise. B1: A faded diagonal

Let $\mathbb{N}^{\mathbb{N}}$ be the set of all functions $\mathbb{N} \rightarrow \mathbb{N}$. Here is a proof that $\mathbb{N}^{\mathbb{N}}$ is uncountable, with its last step left blank. *Name the move, then complete it.*

Let $f: \mathbb{N} \rightarrow \mathbb{N}^{\mathbb{N}}$ be any function, and write $f_n := f(n)$, so that $f_0, f_1, f_2, \dots$ is a list of elements of $\mathbb{N}^{\mathbb{N}}$. Define $g: \mathbb{N} \rightarrow \mathbb{N}$ by $g(n) := f_n(n) + 1$. (Last step: _____.) Since f was arbitrary, no function $\mathbb{N} \rightarrow \mathbb{N}^{\mathbb{N}}$ is surjective, and as $\mathbb{N}^{\mathbb{N}}$ is non-empty (it contains the zero function), $\mathbb{N}^{\mathbb{N}}$ is not countable.

Solution. The move is **the diagonal**: g is built to disagree with the list's n -th entry in place n . The last step: for every $n \in \mathbb{N}$, $g(n) = f_n(n) + 1 \neq f_n(n)$, so $g \neq f_n$ by [Proposition 1.4.1](#) (two equal functions agree at n). Hence g is not on the list: g is an element of $\mathbb{N}^{\mathbb{N}}$ (its values $f_n(n) + 1$ are natural numbers) that is not a value of f , and f is not surjective.

Exercise. B2

Prove that the set of **infinite** subsets of $\mathbb{N}$ is uncountable.

Solution. Let $\mathcal{F}$ be the set of finite subsets of $\mathbb{N}$ and $\mathcal{I}$ the set of infinite ones, both subsets of $\mathcal{P}(\mathbb{N})$ defined by specification. By excluded middle (Logic and Proof, Proposition 2.5.3), every subset of $\mathbb{N}$ is finite or not finite, that is, finite or infinite (Definition 3.1.3), so $\mathcal{P}(\mathbb{N}) = \mathcal{F} \cup \mathcal{I}$. The set $\mathcal{F}$ is countable by Exercise (C1: Finite subsets of the natural numbers). If $\mathcal{I}$ were countable, $\mathcal{P}(\mathbb{N})$ would be countable by Proposition 3.2.5, contradicting Corollary 3.4.3. So $\mathcal{I}$ is uncountable.

Exercise. B3: Expansions of three numbers

For each of $x = \frac{1}{3}, \frac{1}{2}, 1$, find **all** decimal expansions of x , checking the inequalities of Definition 3.4.1 for each one you give, and say which of (D1)–(D3) each case illustrates.

Solution. We use Proposition 2.3.6 with $r = \frac{1}{10}$, after taking the constant factor $\frac{1}{10}$ out of the sum: $\sum_{k=1}^n 10^{-k} = \frac{1}{10} \sum_{k=1}^n (\frac{1}{10})^{k-1} = \frac{1}{10} \cdot \frac{1-10^{-n}}{1-\frac{1}{10}} = \frac{1}{9}(1-10^{-n})$.

$x = \frac{1}{3}$. The all-three sequence has $t_n = \frac{3}{9}(1-10^{-n}) = \frac{1}{3} - \frac{1}{3}10^{-n}$, so $t_n \leq \frac{1}{3} \leq t_n + 10^{-n}$. It is an expansion, and it does not end in all 9s, as (D1) promises. It ends in neither 9s nor 0s, so by (D3) no other expansion of $\frac{1}{3}$ exists. Answer: 0.333... only.

$x = \frac{1}{2}$. Example 2 after Definition 3.4.1 checks that 0.5000... and 0.4999... are expansions. There are no others: if three different expansions existed, two of them would both end in all 9s or both not; but by (D3), of two different expansions of one number, one ends in all 9s and the other in all 0s, and a sequence ending in all 0s does not end in all 9s (from some place on, its digits would be both 0 and 9). Answer: exactly these two, which illustrate (D3); the first is the one (D1) promises.

$x = 1$. The all-nine sequence is an expansion (Example 3). Conversely, let (d_k) be an expansion of 1, and suppose $d_m \leq 8$ for some m . Since every digit is at most 9, $t_m \leq \sum_{k=1}^m 9 \cdot 10^{-k} - 10^{-m} = 1 - 2 \cdot 10^{-m}$, so $t_m + 10^{-m} \leq 1 - 10^{-m} < 1$, against the definition. So every digit is 9. Answer: 0.999... only. The all-nine sequence is an expansion of 1, an instance of the existence half of (D2). Here (D3) has no second expansion to compare. The case also shows why (D1) is stated for $[0, 1)$ only: the one expansion of 1 ends in all 9s.

3.4.6.3 C. Going deeper

Exercise. C1: Exit check: Cantor

Prove Theorem 3.4.2: for every set X , there is no surjection $X \rightarrow \mathcal{P}(X)$. Deduce that $\mathcal{P}(\mathbb{N})$ is uncountable.

Solution. Let $f: X \rightarrow \mathcal{P}(X)$ be a function, and let $D := \{x \in X : x \notin f(x)\}$, a subset of X , so $D \in \mathcal{P}(X)$. Suppose f is surjective. Then $D = f(d)$ for some $d \in X$. If $d \in D$, then $d \notin f(d) = D$ by the definition of D , a contradiction; so $d \notin D$. But then $d \notin f(d)$ and

$d \in X$, so $d \in D$ by the definition of D , a contradiction. Hence f is not surjective, and since f was arbitrary, no surjection $X \rightarrow \mathcal{P}(X)$ exists.

For $X = \mathbb{N}$: by [Definition 3.2.1](#), $\mathcal{P}(\mathbb{N})$ is countable only if it is empty or there is a surjection $\mathbb{N} \rightarrow \mathcal{P}(\mathbb{N})$. It is not empty, since $\emptyset \in \mathcal{P}(\mathbb{N})$, and the theorem excludes the surjection. So $\mathcal{P}(\mathbb{N})$ is uncountable.

Marking scheme. (10 points) - 2: takes an arbitrary $f: X \rightarrow \mathcal{P}(X)$ and either assumes it onto, towards a contradiction, or sets out to show that it misses a set. - 3: defines $D = \{x \in X : x \notin f(x)\}$ and notes that it is a subset of X . - 2: uses surjectivity to obtain d with $f(d) = D$ (or, in the direct form, fixes an arbitrary x). - 2: refutes both cases, $d \in D$ and $d \notin D$ (or shows that D and $f(x)$ disagree about x). - 1: the deduction for $\mathcal{P}(\mathbb{N})$, handling both clauses of “countable”: non-empty, and no surjection from $\mathbb{N}$. The direct form (“ $D \neq f(x)$ for every x , since they disagree about x ”) earns the same points: 2 for the arbitrary f , 3 for D , 4 for the disagreement at every x , 1 for the deduction. Fatal errors (0 for the step): $D = \{x \in X : x \in f(x)\}$ (it may well be a value of f); checking the claim for one particular f and concluding for all, as in (N1) of Logic and Proof, Example (What is not allowed) (0 for the first step).

Exercise. C2: Permutations of the naturals

Prove that the set of all bijections $\mathbb{N} \rightarrow \mathbb{N}$ is uncountable.

Solution (Hint 1). Consider the bijections that only swap $2n$ and $2n + 1$, for some of the $n \in \mathbb{N}$, and fix every other number.

Solution (Hint 2). Such bijections correspond to $\{0, 1\}^{\mathbb{N}}$: the n -th entry says whether the pair $2n, 2n + 1$ is swapped.

Solution. Every $m \in \mathbb{N}$ is $2q + r$ for exactly one pair (q, r) with $q \in \mathbb{N}$ and $r \in \{0, 1\}$ ([Theorem 2.2.3](#) with divisor 2, the quotient lying in $\mathbb{N}$ as in the proof of [Proposition 3.2.5](#)). For $b \in \{0, 1\}^{\mathbb{N}}$ define $\pi_b: \mathbb{N} \rightarrow \mathbb{N}$ by

$$\pi_b(2q + r) := \begin{cases} 2q + r & \text{if } b(q) = 0, \\ 2q + 1 - r & \text{if } b(q) = 1, \end{cases}$$

so π_b swaps $2q$ and $2q + 1$ when $b(q) = 1$ and fixes them otherwise. It maps each pair $\{2q, 2q + 1\}$ to itself, and applying it twice returns every number to its place, so $\pi_b \circ \pi_b = \text{id}_{\mathbb{N}}$, and π_b is a bijection by [Theorem 1.5.4](#).

Let B be the set of bijections $\mathbb{N} \rightarrow \mathbb{N}$; it is non-empty, since $\text{id}_{\mathbb{N}} \in B$. Let $f: \mathbb{N} \rightarrow B$ be any function, and write $f_n := f(n)$. Define $b \in \{0, 1\}^{\mathbb{N}}$ by $b(n) := 1$ if $f_n(2n) = 2n$, and $b(n) := 0$ otherwise, and let $\pi := \pi_b \in B$. For every $n \in \mathbb{N}$, $\pi(2n) \neq f_n(2n)$: if $f_n(2n) = 2n$, then $b(n) = 1$ and $\pi(2n) = 2n + 1$; if $f_n(2n) \neq 2n$, then $b(n) = 0$ and $\pi(2n) = 2n$. So $\pi \neq f_n$ for every n ([Proposition 1.4.1](#)), and f is not surjective. Since f was arbitrary and $B \neq \emptyset$, B is uncountable ([Definition 3.2.1](#)).

This is the diagonal again, with the n -th entry tested at the place $2n$. Alternatively, $b \mapsto \pi_b$ is an injection $\{0, 1\}^{\mathbb{N}} \rightarrow B$, since $\pi_b(2q)$ is $2q$ or $2q+1$ according to $b(q)$; if B were countable, composing with an injection $B \rightarrow \mathbb{N}$ (Theorem 3.2.2) would make $\{0, 1\}^{\mathbb{N}}$ countable, against Theorem 3.4.1.

Marking scheme. (10 points) - 3: defines, for each $b \in \{0, 1\}^{\mathbb{N}}$ (or each set of pairs), the bijection that swaps $2n$ and $2n+1$ exactly when $b(n) = 1$, and proves it is a bijection. - 4: given an arbitrary list of bijections, defines b from the n -th entry's value at a place belonging to the n -th pair, and shows that the resulting bijection differs from the n -th entry; or shows that $b \mapsto \pi_b$ is injective. - 3: concludes, with the non-emptiness of the set of bijections or through Theorem 3.4.1 and the closure properties of countable sets. Fatal errors (0 for the step): the plain diagonal $g(n) := f_n(n) + 1$, which is not a bijection; a “bijection” that moves numbers between different pairs without checking that it stays one-to-one and onto.

3.5 Comparing Sizes: the Schröder–Bernstein Theorem

We have $(0, 1) \subseteq [0, 1] \subseteq \mathbb{R}$. Are the three sets the same size? Section 3.1 made “same size” mean “there is a bijection”, and for these sets a bijection is fiddly to write down: $[0, 1]$ has two endpoints that $(0, 1)$ lacks, and $\mathbb{R}$ is unbounded. An **injection**, on the other hand, is easy in both directions: the inclusion goes one way, and squeezing $[0, 1]$ into the middle of $(0, 1)$ goes the other. This section turns injections into a way of comparing sizes, and proves the theorem that makes the comparison useful: injections both ways give a bijection.

3.5.1 Comparing sizes with injections

An injection $A \rightarrow B$ places a copy of A inside B : distinct elements of A go to distinct elements of B , so B has room for all of A . For finite sets this is the pigeonhole principle read the other way round (Corollary 3.1.3). It deserves a name for all sets.

A is at most as big as B when A fits into B without collisions.

Definition 3.5.1. Comparing Sizes

Let A and B be sets. We write $|A| \leq |B|$, and say that A is **at most as big as** B , if **there exists** an injection $A \rightarrow B$. We write $|A| < |B|$, and say that A is **strictly smaller than** B , if $|A| \leq |B|$ and **not** $|A| = |B|$: there is an injection $A \rightarrow B$, and there is **no** bijection $A \rightarrow B$.

In words: “ $\leq$ ” asks for one injection, of our choosing. “ $<$ ” asks for an injection **and** rules out every bijection, which is a statement about all functions $A \rightarrow B$ at once. As with $|A| = |B|$ in Section 3.1 (Definition 3.1.1), the symbols $|A| \leq |B|$ and $|A| < |B|$ are **relations between the sets** A and B . No object $|A|$ is defined for an infinite set, and nothing here compares two numbers.

Examples.

- *Subsets.* If $A \subseteq B$, the inclusion $a \mapsto a$ is an injection $A \rightarrow B$, so $|A| \leq |B|$. In particular $|(0, 1)| \leq |[0, 1]| \leq |\mathbb{R}|$.
- *The degenerate case.* For every set B , $|\emptyset| \leq |B|$: the empty function $\emptyset \rightarrow B$ (Proposi-

tion 1.4.2 (a)) is injective, vacuously, since there are no $x_1, x_2 \in \emptyset$ to test.

- *Finite sets.* Let A have m elements and B have n elements (Definition 3.1.2). Then $|A| \leq |B|$ exactly when $m \leq n$. If $m \leq n$, take bijections $\varphi: \{1, \dots, m\} \rightarrow A$ and $\psi: \{1, \dots, n\} \rightarrow B$; the inclusion $\iota: \{1, \dots, m\} \rightarrow \{1, \dots, n\}$ is injective, so $\psi \circ \iota \circ \varphi^{-1}: A \rightarrow B$ is injective by Theorem 1.4.4 (a), φ^{-1} being a bijection (Theorem 1.5.4). If $m > n$, there is no injection $A \rightarrow B$ by Corollary 3.1.3 (a). So the relation agrees with the count of elements, which is why it is allowed to borrow the symbol $\leq$.
- *Countable sets.* By Theorem 3.2.2, (i) $\iff$ (ii), a set A is countable if and only if $|A| \leq |\mathbb{N}|$. “Countable” means “at most as big as $\mathbb{N}$ ”.
- *Power sets.* For every set X , $|X| < |\mathcal{P}(X)|$. The map $x \mapsto \{x\}$ is an injection $X \rightarrow \mathcal{P}(X)$, since $\{x_1\} = \{x_2\}$ gives $x_1 \in \{x_2\}$, so $x_1 = x_2$. And there is no bijection $X \rightarrow \mathcal{P}(X)$, because by Theorem 3.4.2 there is not even a surjection. In particular $|\mathbb{N}| < |\mathcal{P}(\mathbb{N})|$.

Non-example by minimal change. Replace “injection” by “function” in the definition. Then any set would be “at most as big as” any non-empty set, by a constant map: $\mathbb{R} \rightarrow \{0\}$ would put $\mathbb{R}$ below a one-element set. The clause that carries the meaning is “no two elements collide”.

Warning. $|A| < |B|$ does **not** mean “there is an injection $A \rightarrow B$ that is not a bijection”. The map $n \mapsto n + 1$ is an injection $\mathbb{N} \rightarrow \mathbb{N}$ that misses 0, and yet $|\mathbb{N}| = |\mathbb{N}|$. For infinite sets one injection that fails to be onto proves nothing about strictness; $|A| < |B|$ needs **every** map $A \rightarrow B$ to fail to be a bijection, which is why Cantor’s theorem, a statement about every $f: X \rightarrow \mathcal{P}(X)$, is the source of our strict inequalities.

The relation behaves like an order in two of the three ways. The third is the content of this section’s theorem.

Proposition 3.5.1. Comparing Sizes Is Reflexive and Transitive

For all sets A, B, C :

- (a) $|A| \leq |A|$;
- (b) if $|A| \leq |B|$ and $|B| \leq |C|$, then $|A| \leq |C|$.

Proof.

- (a) The identity id_A is injective, since $\text{id}_A(x_1) = \text{id}_A(x_2)$ says $x_1 = x_2$.
- (b) Let $f: A \rightarrow B$ and $g: B \rightarrow C$ be injections. Then $g \circ f: A \rightarrow C$ is injective by Theorem 1.4.4 (a). This proves the proposition. ■

As with equinumerosity (Section 3.1, after Proposition 3.1.1), we do not call this a partial order: there is no set of all sets for it to be a relation on (Theorem 1.1.4). And antisymmetry, the clause (PO2) of Definition 2.6.1, would say “if $|A| \leq |B|$ and $|B| \leq |A|$, then $A = B$ ”, which is false: $|\mathbb{N}| \leq |2\mathbb{N}|$ and $|2\mathbb{N}| \leq |\mathbb{N}|$, while $2\mathbb{N} \neq \mathbb{N}$. The right replacement for “ $A = B$ ” is “ $|A| = |B|$ ”. That is the Schröder–Bernstein Theorem below.

Before it, one more way to read $\leq$. A countable set was defined by a surjection from $\mathbb{N}$, and Theorem 3.2.2 turned that into an injection into $\mathbb{N}$. The same exchange works for all sets, with

one exception and one price.

Proposition 3.5.2. Injections and Surjections

Let A and B be sets with $A \neq \emptyset$. Then $|A| \leq |B|$ if and only if there is a surjection $B \rightarrow A$.

Proof. ($\Rightarrow$) Let $f: A \rightarrow B$ be an injection. Since $A \neq \emptyset$, [Theorem 1.5.1](#) (b) gives a left inverse $g: B \rightarrow A$, so $g \circ f = \text{id}_A$. This says that f is a right inverse of g ([Definition 1.5.1](#)), so g is surjective by [Proposition 1.5.2](#).

($\Leftarrow$) Let $g: B \rightarrow A$ be a surjection. By [Theorem 3.3.2](#), g has a right inverse $h: A \rightarrow B$, so $g \circ h = \text{id}_A$; **this step uses [Axiom 3.3.1](#)**, through [Theorem 3.3.2](#). Then g is a left inverse of h , so h is injective by [Theorem 1.5.1](#) (a). Hence $|A| \leq |B|$. This proves the proposition. ■

The hypothesis $A \neq \emptyset$ is needed for ($\Rightarrow$) only. For $A = \emptyset$ and $B = \{b\}$, the empty function is an injection $\emptyset \rightarrow \{b\}$, but there is no function $\{b\} \rightarrow \emptyset$ at all ([Proposition 1.4.2](#) (b)), let alone a surjection. The proof shows where the hypothesis enters: a left inverse $g: B \rightarrow A$ must send b somewhere in A . The choice in ($\Leftarrow$) is genuine in general, but not for $B = \mathbb{N}$: a surjection from $\mathbb{N}$ has least preimages, a rule, which is how [Theorem 3.2.2](#) went from (i) to (ii) without the axiom.

Quick check. Is $|\{0, 1\}^{\mathbb{N}}| \leq |\mathcal{P}(\mathbb{N})|$? Give an injection, and say whether your answer needed [Proposition 3.5.2](#).

Solution. Yes. Send a sequence $a \in \{0, 1\}^{\mathbb{N}}$ to the set $\{n \in \mathbb{N} : a(n) = 1\}$. If two sequences give the same set, they take the value 1 at the same places, and so the value 0 at the same places too; hence they are equal ([Proposition 1.4.1](#)), and the map is injective. No appeal to [Proposition 3.5.2](#) is needed: the definition asks for an injection, and we wrote one down. (The map is in fact a bijection, with inverse $S \mapsto \mathbf{1}_S$; this is the correspondence of [Section 1.4](#), recalled after [Corollary 3.4.3](#).)

3.5.2 The Schröder–Bernstein Theorem

Here is the replacement for antisymmetry. To show two sets are the same size, it is enough to fit each into the other.

Theorem 3.5.3. Schröder–Bernstein Theorem

Let A and B be sets. If $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$.

Idea. Let $f: A \rightarrow B$ and $g: B \rightarrow A$ be the two injections. We want a bijection $h: A \rightarrow B$, and at each $a \in A$ there are two candidates: go forward, $h(a) = f(a)$, or go back along g : let $h(a)$ be the unique b with $g(b) = a$. Going back is possible only for $a \in g(B)$. So a point

of A **not hit by** g must be sent forward by f . Then the point $f(a) \in B$ is used up, and $g(f(a))$ must not go back to it, or h would collide; so $g(f(a))$ must also be sent forward, and then $g(f(g(f(a))))$, and so on along the chain. Everything reached this way goes forward; everything else goes back.

Step roadmap.

① Build the set $C \subseteq A$ of points that are forced forward: $C_0 := A \setminus g(B)$, $C_{n+1} := g(f(C_n))$, $C := \bigcup_{n \in \mathbb{N}} C_n$.

② Define $h(a) := f(a)$ on C , and let $h(a)$ be the unique b with $g(b) = a$ off C .

③ Show h is injective; the only case with content is one point of each kind.

④ Show h is onto: for $b \in B$, look at $g(b)$ and ask whether it lies in C .

Before reading on: in step ④, why is $g(b) \notin C_0$, and what does that buy?

Proof. Let $f: A \rightarrow B$ and $g: B \rightarrow A$ be injections.

Step 1. By [Theorem 2.3.1](#), applied with the set $\mathcal{P}(A)$, the starting value $A \setminus g(B) \in \mathcal{P}(A)$ and the function $(n, S) \mapsto g(f(S))$ from $\mathbb{N} \times \mathcal{P}(A)$ to $\mathcal{P}(A)$, there are sets $C_n \subseteq A$, for $n \in \mathbb{N}$, with

$$C_0 = A \setminus g(B), \quad C_{n+1} = g(f(C_n)) \text{ for every } n \in \mathbb{N}.$$

Here $f(C_n) \subseteq B$ and $g(f(C_n)) \subseteq A$ are images ([Definition 1.6.1](#)). Let $C := \bigcup_{n \in \mathbb{N}} C_n \subseteq A$ ([Definition 1.2.4](#)). Two facts about C will be used.

(F1) If $a \in C$, then $g(f(a)) \in C$. Indeed $a \in C_n$ for some $n \in \mathbb{N}$; then $f(a) \in f(C_n)$, so $g(f(a)) \in g(f(C_n)) = C_{n+1} \subseteq C$.

(F2) If $a \in A \setminus C$, there is **exactly one** $b \in B$ with $g(b) = a$. It exists because $a \notin C_0 = A \setminus g(B)$ and $a \in A$, so $a \in g(B)$. It is unique because g is injective.

Step 2. Define $h: A \rightarrow B$ by

$$h(a) := \begin{cases} f(a) & \text{if } a \in C, \\ \text{the unique } b \in B \text{ with } g(b) = a & \text{if } a \notin C. \end{cases}$$

Each $a \in A$ falls under exactly one case, and in each case the value exists and is unique (in the second case by (F2)), so h is a function. Note that $g(h(a)) = a$ whenever $a \notin C$.

Step 3.

Claim. h is injective.

Proof of Claim. Let $a, a' \in A$ with $h(a) = h(a')$.

Case 1: $a, a' \in C$. Then $f(a) = f(a')$, and $a = a'$ because f is injective.

Case 2: $a, a' \notin C$. Then $a = g(h(a)) = g(h(a')) = a'$.

Case 3: $a \in C$ and $a' \notin C$. Then $h(a') = h(a) = f(a)$, so $a' = g(h(a')) = g(f(a))$. By (F1), $g(f(a)) \in C$, so $a' \in C$, a contradiction. So this case does not occur.

The case $a \notin C$, $a' \in C$ is Case 3 with the roles of a and a' swapped. Hence $a = a'$ in every case that occurs. ■

Step 4.

Claim. h is surjective.

Proof of Claim. Let $b \in B$, and consider $g(b) \in A$.

Case 1: $g(b) \notin C$. Then $h(g(b))$ is the unique element of B whose image under g is $g(b)$, and b is such an element. So $h(g(b)) = b$.

Case 2: $g(b) \in C$. Then $g(b) \in C_n$ for some $n \in \mathbb{N}$. Since $g(b) \in g(B)$, we have $g(b) \notin A \setminus g(B) = C_0$, so $n \neq 0$. By [Lemma 2.1.1](#) (c), $k := n - 1 \in \mathbb{N}$, and $g(b) \in C_{k+1} = g(f(C_k))$. By [Definition 1.6.1](#) there is $b' \in f(C_k)$ with $g(b') = g(b)$, and then $a \in C_k$ with $f(a) = b'$. As g is injective, $b = b' = f(a)$. Since $a \in C_k \subseteq C$, $h(a) = f(a) = b$.

In both cases b is a value of h . ■

By the two claims, $h: A \rightarrow B$ is a bijection, so $|A| = |B|$. This proves the theorem. ■

The answer to the prompt: $g(b)$ is a value of g , and C_0 consists exactly of the points of A that are **not** values of g . That is what pushes the index n of Case 2 up to at least 1, so that $g(b)$ has a predecessor $a \in C_{n-1}$ in its chain, and b is reached by f from that predecessor.

Every value of h is given by a rule, so the proof makes **no use of Axiom 3.3.1**: Step 1 is a recursion with an explicit step, and Step 2 picks the unique preimage, where there is nothing to choose. The case split “ $a \in C$ or $a \notin C$ ” is an instance of excluded middle, the classical step discussed in *Logic and Proof*, Section 5.7; nothing in the proof decides which case holds for a given a , and nothing needs to.

Each hypothesis is used. Without the injection g , the conclusion fails: $\mathbb{N} \rightarrow \mathcal{P}(\mathbb{N})$, $n \mapsto \{n\}$, is an injection, and $|\mathbb{N}| \neq |\mathcal{P}(\mathbb{N})|$ by [Theorem 3.4.2](#). In the proof, injectivity of f is used once, in Case 1 of Step 3, and injectivity of g twice, in (F2) and in Case 2 of Step 4.

Remark. The theorem needs both $|A| \leq |B|$ and $|B| \leq |A|$; are any two sets comparable, that is, is it true that $|A| \leq |B|$ or $|B| \leq |A|$ for all A and B ? The answer is yes, but the proof needs Zorn’s lemma, a principle equivalent to [Axiom 3.3.1](#); a later chapter devoted to Zorn’s lemma states it, and the companion volume on set theory proves it from the axiom. This chapter does not prove comparability, and nothing in it needs it.

3.5.3 Intervals of real numbers

The theorem’s first use is the question of the opening frame.

Example. The closed and the open unit interval

Show that $|[0, 1]| = |(0, 1)|$: first with [Theorem 3.5.3](#), then by writing down a bijection.

Solution. With the theorem. The inclusion $(0, 1) \rightarrow [0, 1]$ is an injection. In the other direction let $u(x) := x/2 + 1/4$ for $x \in [0, 1]$. If $0 \leq x \leq 1$, then $1/4 \leq u(x) \leq 3/4$, so $u(x) \in (0, 1)$, and u is a function $[0, 1] \rightarrow (0, 1)$. It is injective: $x/2 + 1/4 = y/2 + 1/4$ gives $x = y$. By [Theorem 3.5.3](#), $|[0, 1]| = |(0, 1)|$.

A bijection. The two extra points 0 and 1 are absorbed by a Hilbert's hotel ([Example \(Hilbert's Hotel\)](#)) run on the points $1, \frac{1}{2}, \frac{1}{3}, \dots$. Define $\varphi: [0, 1] \rightarrow (0, 1)$ by

$$\varphi(x) := \begin{cases} \frac{1}{2} & \text{if } x = 0, \\ \frac{1}{n+2} & \text{if } x = \frac{1}{n} \text{ for some } n \in \mathbb{Z}_{>0}, \\ x & \text{otherwise.} \end{cases}$$

The cases do not overlap, since $\frac{1}{n} \neq 0$, and in the second case $n = 1/x$ is determined by x ; so φ is a function, with values in $(0, 1)$ because $0 < \frac{1}{n+2} \leq \frac{1}{3}$. Let $H := \{0\} \cup \{\frac{1}{n} : n \in \mathbb{Z}_{>0}\} = \{0, 1, \frac{1}{2}, \frac{1}{3}, \dots\}$ and $H' := \{\frac{1}{m} : m \in \mathbb{Z}_{>0}, m \geq 2\} = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\}$. Then φ maps H onto H' : $0 \mapsto \frac{1}{2}$, and $\frac{1}{n} \mapsto \frac{1}{n+2}$ runs through $\frac{1}{3}, \frac{1}{4}, \dots$. It is injective on H , since $\frac{1}{n+2} = \frac{1}{n'+2}$ forces $n = n'$, and $\frac{1}{n+2} \leq \frac{1}{3} < \frac{1}{2}$. Off H it is the identity of $[0, 1] \setminus H = (0, 1) \setminus H'$. Since

$$[0, 1] = H \cup ([0, 1] \setminus H), \quad (0, 1) = H' \cup ((0, 1) \setminus H'),$$

both unions disjoint, φ is a bijection, and $|[0, 1]| = |(0, 1)|$. For instance $\varphi(1) = \frac{1}{3}$, $\varphi(\frac{1}{2}) = \frac{1}{4}$, $\varphi(\frac{3}{4}) = \frac{3}{4}$.

The first route took two lines and the second took a paragraph. That is the point of the theorem: we never had to find φ .

The second interval computation turns an unbounded set into a bounded one.

Example. The real line and its intervals

Show that $x \mapsto x/(1 + |x|)$ is a bijection $\mathbb{R} \rightarrow (-1, 1)$. Deduce that every open interval (c, d) with $c < d$ is equinumerous with $\mathbb{R}$, and then that every interval with more than one point is.

Solution. The bijection. Let $s(x) := x/(1 + |x|)$ for $x \in \mathbb{R}$; the denominator is at least 1, so $s(x)$ is defined, and $|s(x)| = |x|/(1 + |x|) < 1$, so $s(x) \in (-1, 1)$. For $y \in (-1, 1)$ let $t(y) := y/(1 - |y|)$; here $1 - |y| > 0$. We check that t is a two-sided inverse. For $x \in \mathbb{R}$, $|s(x)| = |x|/(1 + |x|)$, so

$$1 - |s(x)| = \frac{1}{1 + |x|}, \quad t(s(x)) = \frac{x}{1 + |x|} \cdot (1 + |x|) = x.$$

For $y \in (-1, 1)$, $|t(y)| = |y|/(1 - |y|)$, so

$$1 + |t(y)| = \frac{1}{1 - |y|}, \quad s(t(y)) = \frac{y}{1 - |y|} \cdot (1 - |y|) = y.$$

By [Theorem 1.5.4](#), s is a bijection $\mathbb{R} \rightarrow (-1, 1)$. For example $s(1) = \frac{1}{2}$ and $s(-3) = -\frac{3}{4}$.

Open intervals. The map $y \mapsto (y + 1)/2$ is a bijection $(-1, 1) \rightarrow (0, 1)$, with inverse $u \mapsto 2u - 1$. For $c < d$, the map $u \mapsto c + (d - c)u$ is a bijection $(0, 1) \rightarrow (c, d)$, with inverse $v \mapsto (v - c)/(d - c)$, since $d - c > 0$. Composing the three bijections (Theorem 1.4.4 (c)) gives a bijection $\mathbb{R} \rightarrow (c, d)$, so $|(c, d)| = |\mathbb{R}|$.

Every interval with more than one point. Let I be an interval (bounded or not, with or without its endpoints) containing two points $c < d$. An interval contains every number between two of its points, so $(c, d) \subseteq I \subseteq \mathbb{R}$. The inclusion is an injection $(c, d) \rightarrow I$. In the other direction, let $\beta: \mathbb{R} \rightarrow (c, d)$ be the bijection just built; then β restricted to I is an injection $I \rightarrow (c, d)$. By Theorem 3.5.3, $|I| = |(c, d)|$, and $|(c, d)| = |\mathbb{R}|$, so $|I| = |\mathbb{R}|$ by Proposition 3.1.1. This covers $[0, 1]$, $[0, 1)$, $(0, \infty)$, $[5, \infty)$ and $\mathbb{R}$ itself at once.

In particular every interval with more than one point is uncountable: it is equinumerous with $\mathbb{R}$, which is uncountable by Theorem 3.4.4, and a set equinumerous with a countable set is countable (compose a surjection from $\mathbb{N}$ with the bijection, Proposition 3.2.4).

Warning. Injections both ways do not make either injection a bijection. On $\mathbb{N}$, $n \mapsto 2n$ is an injection $\mathbb{N} \rightarrow \mathbb{N}$ in both directions, and it is not onto: 1 is never a value. Theorem 3.5.3 builds a **new** map h out of f and g ; it says nothing about f or g themselves. In Example (The closed and the open unit interval), $u(x) = x/2 + 1/4$ misses $\frac{9}{10} \in (0, 1)$, and the bijection is a different map.

Quick check. Give an injection $\mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$ and an injection $\mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$. What does Theorem 3.5.3 conclude?

Solution. $n \mapsto (n, 1)$ is injective, by the characteristic property of pairs (Definition 1.3.1). $(m, n) \mapsto 2^m 3^n$ is injective, as shown in the proof of Theorem 3.2.6. So $|\mathbb{N}| = |\mathbb{N} \times \mathbb{N}|$, and $\mathbb{N} \times \mathbb{N}$ is countably infinite (Definition 3.2.2). Section 3.2 showed only that $\mathbb{N} \times \mathbb{N}$ is countable; the theorem gives a bijection with $\mathbb{N}$ directly, with no listing at all.

3.5.4 Exercises

3.5.4.1 A. Check your understanding

Exercise. A1

These parts are about the proof of Theorem 3.5.3.

- In Step 4, Case 2, why is $g(b) \notin C_0$, and what would go wrong without it?
- Run the construction for $A = B = \mathbb{N}$ and $f(n) = g(n) = 2n$. Compute C_0 , C_1 and C , and the values $h(0), h(1), h(2), h(3), h(4)$.
- Where is the injectivity of f used?
- Give the idea of the proof in two sentences.

Solution.

- (a) $C_0 = A \setminus g(B)$ consists of the points that are not values of g , and $g(b)$ is a value of g . Without it the index n with $g(b) \in C_n$ could be 0, and then there would be no C_{n-1} to find a predecessor a in, so no $a \in C$ with $f(a) = b$.
- (b) $g(\mathbb{N}) = 2\mathbb{N}$, so $C_0 = \mathbb{N} \setminus 2\mathbb{N}$ is the set of odd numbers. $C_1 = g(f(C_0)) = \{4q : q \text{ odd}\} = \{4, 12, 20, \dots\}$. In general $C_{n+1} = \{4c : c \in C_n\}$, so $C_n = \{4^n q : q \text{ odd}\}$ by induction on n , and C is the set of numbers $4^k q$ with $k \in \mathbb{N}$ and q odd. Now:
- $0 \notin C$, since $4^k q \neq 0$ for odd q ($4^k > 0$ and $q \geq 1$). So $h(0)$ is the b with $2b = 0$: $h(0) = 0$, a fixed point.
 - $1 \in C_0$, so $h(1) = f(1) = 2$.
 - $2 \notin C$: for $k = 0$, 2 is not odd; for $k \geq 1$, $4^k q \geq 4 > 2$. So $h(2)$ is the b with $2b = 2$: $h(2) = 1$.
 - $3 \in C_0$, so $h(3) = 6$.
 - $4 = 4^1 \cdot 1 \in C_1$, so $h(4) = f(4) = 8$.
- (c) In Step 3, Case 1: from $f(a) = f(a')$ with $a, a' \in C$ it concludes $a = a'$.
- (d) Points of A that g does not reach must be sent forward by f , and so must everything reached from them by repeating $g \circ f$; this is the set C . Every other point of A is a value of g and is sent back to its unique preimage, and the two halves fit together into a bijection.

3.5.4.2 B. Practice

Exercise. B1: Injections both ways

Prove each of the following by giving an injection in each direction and applying [Theorem 3.5.3](#).

- (a) $|[0, 1]| = |\mathbb{R}|$.
- (b) $|\mathbb{N}| = |\mathbb{N} \times \mathbb{Z}|$.
- (c) $|(0, 1) \cup (2, 3)| = |(0, 1)|$.

Solution.

- (a) The inclusion $[0, 1] \rightarrow \mathbb{R}$ is injective. In the other direction, $x \mapsto (x/(1 + |x|) + 1)/2$ is a bijection $\mathbb{R} \rightarrow (0, 1)$ by [Example \(The real line and its intervals\)](#), so it is an injection $\mathbb{R} \rightarrow [0, 1]$, as $(0, 1) \subseteq [0, 1]$.
- (b) $n \mapsto (n, 0)$ is an injection $\mathbb{N} \rightarrow \mathbb{N} \times \mathbb{Z}$, by the characteristic property of pairs ([Definition 1.3.1](#)). In the other direction, let $j: \mathbb{Z} \rightarrow \mathbb{N}$ be $j(k) := 2k$ for $k \geq 0$ and $j(k) := -2k - 1$ for $k < 0$. Its values are in $\mathbb{N}$, even for $k \geq 0$ and odd for $k < 0$; since no natural number is both even and odd ([Theorem 2.2.3](#), uniqueness), $j(k) = j(k')$ forces k and k' to have the same sign, and then $k = k'$. So j is injective, and so is $(m, k) \mapsto (m, j(k))$. Following it by the injection $(m, n) \mapsto 2^m 3^n$ from the proof of [Theorem 3.2.6](#) gives an injection $\mathbb{N} \times \mathbb{Z} \rightarrow \mathbb{N}$ ([Theorem 1.4.4 \(a\)](#)).
- (c) The inclusion $(0, 1) \rightarrow (0, 1) \cup (2, 3)$ is injective. The map $x \mapsto x/4$ is injective and sends $(0, 1) \cup (2, 3)$ into $(0, \frac{1}{4}) \cup (\frac{1}{2}, \frac{3}{4}) \subseteq (0, 1)$.

Hence, by [Theorem 3.5.3](#), the three equalities hold.

Exercise. B2: A proper subset of the same size

Give an example of sets $A \subsetneq B \subseteq \mathbb{R}$ with $|A| = |B|$, such that A is uncountable and $B \setminus A$ is uncountable. Check each property, and list the properties any answer must have.

Solution. An answer must have: (1) $A \subseteq B \subseteq \mathbb{R}$; (2) $A \neq B$; (3) a bijection $A \rightarrow B$, or injections both ways; (4) A uncountable; (5) $B \setminus A$ uncountable. (Then B is uncountable too, since it contains A , by [Proposition 3.2.3](#).)

Example: $A = (0, 1)$, $B = (0, 3)$. (1) and (2): $(0, 1) \subseteq (0, 3)$ and $2 \in B \setminus A$. (3): $x \mapsto 3x$ is a bijection $(0, 1) \rightarrow (0, 3)$, with inverse $y \mapsto y/3$. (4) and (5): A and $B \setminus A = [1, 3)$ are intervals with more than one point, hence equinumerous with $\mathbb{R}$ by [Example \(The real line and its intervals\)](#), hence uncountable, as shown after it.

Another correct answer: $A = (0, 1)$, $B = \mathbb{R}$, with $B \setminus A \supseteq (1, 2)$ uncountable. Any answer can be checked against (1)–(5).

Exercise. B3: Interleaving two sequences

Prove that $|\{0, 1\}^{\mathbb{N}} \times \{0, 1\}^{\mathbb{N}}| = |\{0, 1\}^{\mathbb{N}}|$ by an explicit bijection.

Hint: put one sequence in the even places and the other in the odd places.

Solution. Every $m \in \mathbb{N}$ is $2n$ or $2n + 1$ for exactly one $n \in \mathbb{N}$, and not both ([Theorem 2.2.3](#) with divisor 2, the quotient in $\mathbb{N}$, as in the proof of [Proposition 3.2.5](#)). So for $a, b \in \{0, 1\}^{\mathbb{N}}$ we may define $c = \Phi(a, b) \in \{0, 1\}^{\mathbb{N}}$ by

$$c(2n) := a(n), \quad c(2n + 1) := b(n) \quad (n \in \mathbb{N}).$$

Define $\Psi: \{0, 1\}^{\mathbb{N}} \rightarrow \{0, 1\}^{\mathbb{N}} \times \{0, 1\}^{\mathbb{N}}$ by $\Psi(c) := (a, b)$ with $a(n) := c(2n)$ and $b(n) := c(2n + 1)$. Then $\Psi(\Phi(a, b)) = (a, b)$: its coordinates $n \mapsto c(2n) = a(n)$ and $n \mapsto c(2n + 1) = b(n)$ are a and b by [Proposition 1.4.1](#), so the pairs are equal ([Definition 1.3.1](#)). And $\Phi(\Psi(c))$ agrees with c at every even place $2n$ and every odd place $2n + 1$, which are all the places, so $\Phi(\Psi(c)) = c$ ([Proposition 1.4.1](#)). By [Theorem 1.5.4](#), Φ is a bijection.

3.5.4.3 C. Going deeper
Exercise. C1: The real line and the subsets of $\mathbb{N}$

Prove that $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$, using the decimal facts (D1)–(D3) of Section 3.4.

- For $x \in [0, 1)$ let $d_1 d_2 \dots$ be its decimal expansion that does not end in all 9s, and let $E(x) := \{10k + d_k : k \in \mathbb{Z}_{>0}\}$. Show that E is a well-defined injection $[0, 1) \rightarrow \mathcal{P}(\mathbb{N})$.
- For $S \subseteq \mathbb{N}$ let y_S be the number with decimal expansion $0.e_1 e_2 e_3 \dots$, where $e_{k+1} := 1$ if $k \in S$ and $e_{k+1} := 2$ if $k \notin S$, for $k \in \mathbb{N}$. Show that $S \mapsto y_S$ is a well-defined injection $\mathcal{P}(\mathbb{N}) \rightarrow [0, 1)$.
- Conclude.

Solution (Hint 1). The tools are the decimal facts (D1)–(D3), and [Theorem 2.2.3](#) with divisor 10 for reading digits back out of $E(x)$.

Solution (Hint 2). Begin (a) with: “By (D1), x has an expansion that does not end in all 9s, and by (D3) it has only one.”

Solution.

- (a) *Well defined.* By (D1), x has an expansion not ending in all 9s. It has only one: if two different expansions of x both failed to end in all 9s, (D3) would make one of them end in all 9s. So $d_1d_2\dots$ is determined by x , and $E(x) \subseteq \mathbb{N}$ because $10k + d_k \in \mathbb{N}$.
Injective. By [Theorem 2.2.3](#) with divisor 10, the number $10j + d_j$, with $0 \leq d_j \leq 9$, has quotient j and remainder d_j , and these are unique. So for each $k \geq 1$, the only element of $E(x)$ with quotient k is $10k + d_k$, and $E(x)$ determines every digit d_k . If $E(x) = E(x')$, then x and x' have the same expansion $d_1d_2\dots$, and by the uniqueness half of (D2) a digit sequence is an expansion of exactly one number; so $x = x'$.
- (b) *Well defined.* By (D2), the digit sequence $e_1e_2\dots$ is an expansion of exactly one $y_S \in [0, 1]$. With $t_1 = e_1/10$ as in [Definition 3.4.1](#), $y_S \leq t_1 + \frac{1}{10} \leq \frac{2}{10} + \frac{1}{10} = \frac{3}{10}$, so $y_S \in [0, 1)$.
Injective. Let $S \neq S'$, say $k \in S \setminus S'$ or $k \in S' \setminus S$. Then the two digit sequences differ at place $k + 1$. Neither contains a 0 or a 9, so neither ends in all 9s or in all 0s. If $y_S = y_{S'}$, the number would have two different expansions, and by (D3) one would end in all 9s, which is false. So $y_S \neq y_{S'}$. The index shift e_{k+1} is there so that whether $0 \in S$ is recorded too, in the first digit.
- (c) By (a), (b) and [Theorem 3.5.3](#), $|[0, 1]| = |\mathcal{P}(\mathbb{N})|$. By [Example \(The real line and its intervals\)](#), $[0, 1)$, an interval with more than one point, has $|[0, 1)| = |\mathbb{R}|$. By [Proposition 3.1.1](#), $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$.

Marking scheme. (10 points) - 2: shows the expansion in (a) is unique, citing (D1) and (D3), so E is a function. - 2: shows E is injective: the set determines each digit, and one digit sequence names one number (D2). - 2: shows y_S exists, is unique and lies in $[0, 1)$. - 2: shows $S \mapsto y_S$ is injective, using that the digits avoid 0 and 9 and (D3). - 2: applies [Theorem 3.5.3](#) and passes from $[0, 1)$ to $\mathbb{R}$ with a cited bijection or injections both ways. Fatal errors (0 for the step): treating “the expansion of x ” as unique without (D1) and (D3); concluding $y_S \neq y_{S'}$ from different digit sequences without ruling out the $0.4999\dots = 0.5$ case.

Exercise. C2: Find the error

Here is a proof that $|[0, 1]| = |(0, 1)|$.

Proof. (1) Let $u: [0, 1] \rightarrow (0, 1)$, $u(x) := x/2 + 1/4$. For $0 \leq x \leq 1$ we have $1/4 \leq u(x) \leq 3/4$, so u does map into $(0, 1)$. (2) If $u(x) = u(y)$, then $x/2 = y/2$, so $x = y$; hence u is injective. (3) The inclusion $\iota: (0, 1) \rightarrow [0, 1]$ is injective. (4) So there are injections in both directions between $[0, 1]$ and $(0, 1)$. (5) Therefore u is a bijection, and $|[0, 1]| = |(0, 1)|$.

Is the conclusion true? Does the argument prove it? If not, name the line that fails, say why, and give a counterexample to the step it takes.

Solution. The conclusion is true: [Example \(The closed and the open unit interval\)](#) proves it twice.

The argument does not prove it. Lines (1)–(4) are correct. Line (5) fails: it infers that a particular injection is a bijection from the existence of injections both ways. The step cites a general “fact”, that injections both ways make a given injection a bijection. That fact is neither proved nor a school fact, and it is false, so the line cites nothing: entry (N9) of Logic and Proof, [Example \(What is not allowed\)](#). And u is in fact not onto: $\frac{9}{10} \in (0, 1)$ is not a value, since every value lies in $[\frac{1}{4}, \frac{3}{4}]$.

Counterexample to the step. Take $A = B = \mathbb{N}$ and $f = g = (n \mapsto 2n)$. These are injections both ways, and the step would conclude that $n \mapsto 2n$ is a bijection $\mathbb{N} \rightarrow \mathbb{N}$; but 1 is not a value.

Repair. Replace line (5) by: “By [Theorem 3.5.3](#), $|[0, 1]| = |(0, 1)|$.” The theorem produces a bijection, but a different one, built from both u and ι .

3.6 Infinite Sets

Section 3.5 compared sizes with injections. Here we ask where $\mathbb{N}$ sits in that comparison. Every infinite set we have met contains a copy of $\mathbb{N}$, and it is tempting to say that this must always happen by picking new elements one at a time; this section makes that picking exact. It turns out to need [Axiom 3.3.1](#) again, this time applied to a family with one set for each non-empty subset of the given set, and it gives two results: $\mathbb{N}$ is the smallest infinite size, and a set is infinite exactly when it is the same size as a proper part of itself.

3.6.1 Every infinite set contains a copy of the natural numbers

The question has a goal theorem: we want to show that an infinite set X has a countably infinite subset, that is, a subset equinumerous with $\mathbb{N}$ ([Definition 3.2.2](#)). Equivalently, we want an injection $\mathbb{N} \rightarrow X$: its image is then such a subset.

Theorem 3.6.1. Infinite Sets Contain a Countably Infinite Subset

Every infinite set X has a countably infinite subset. Equivalently, there is an injection $\mathbb{N} \rightarrow X$.

Idea. The natural attempt picks $x_0 \in X$, then $x_1 \in X \setminus \{x_0\}$, then $x_2 \in X \setminus \{x_0, x_1\}$, and so on. Each pick is possible, because a finite set cannot exhaust the infinite X , and each single pick is one $\exists E$. But “and so on” is a definition by recursion, and [Theorem 2.3.1](#) needs a **function** g that says, once and for all, which new element to take after any finite stage. That function must pick an element from every non-empty set of the form $X \setminus F$ at once, with no rule for doing so when X is arbitrary. So we let [Axiom 3.3.1](#) supply a choice function c on all non-empty subsets of X first, and then the recursion is an ordinary one, with c as

its rule.

Proof. Let X be infinite. The family $(S)_{S \in \mathcal{P}(X) \setminus \{\emptyset\}}$, indexed by the non-empty subsets of X with each index as its own member, consists of non-empty sets, and its union is X : every index is a subset of X , and each $x \in X$ lies in the index $\{x\}$. By **Axiom 3.3.1** there is a function $c: \mathcal{P}(X) \setminus \{\emptyset\} \rightarrow X$ with

$$c(S) \in S \quad \text{for every non-empty } S \subseteq X.$$

This is the one line of the proof that uses the axiom.

Define $g: \mathbb{N} \times \mathcal{P}(X) \rightarrow \mathcal{P}(X)$ by $g(n, F) := F \cup \{c(X \setminus F)\}$ if $F \neq X$, and $g(n, F) := F$ if $F = X$. In the first case $X \setminus F$ is non-empty, so $c(X \setminus F)$ is defined; so g is a function. By **Theorem 2.3.1**, applied with the set $\mathcal{P}(X)$, the starting value $\emptyset$ and this g , there are sets $F_n \subseteq X$, for $n \in \mathbb{N}$, with $F_0 = \emptyset$ and $F_{n+1} = g(n, F_n)$.

We show by induction (Logic and Proof, Axiom 4.2.1) that F_n **has exactly n elements, for every $n \in \mathbb{N}$** . For $n = 0$, $F_0 = \emptyset$ has 0 elements (**Definition 3.1.2**). Suppose F_n has n elements. Then F_n is finite and X is not, so $F_n \neq X$, and $F_{n+1} = F_n \cup \{c(X \setminus F_n)\}$ with $c(X \setminus F_n) \in X \setminus F_n$. This is a union of disjoint sets with n elements and 1 element, so F_{n+1} has $n + 1$ elements by **Proposition 3.1.5**.

In particular $F_n \neq X$ for every $n \in \mathbb{N}$, so we may define

$$x_n := c(X \setminus F_n) \in X \setminus F_n, \quad \text{and then} \quad F_{n+1} = F_n \cup \{x_n\}.$$

Hence $F_n \subseteq F_{n+1}$ for every n , and by induction on k , $F_m \subseteq F_{m+k}$ for all $m, k \in \mathbb{N}$.

Now let $m, n \in \mathbb{N}$ with $m < n$. Then $x_m \in F_{m+1} \subseteq F_n$, because $m + 1 \leq n$ (**Lemma 2.1.1** (e)), so $k := n - (m + 1) \in \mathbb{N}$ (**Lemma 2.1.1** (d)) and $F_{m+1} \subseteq F_{(m+1)+k} = F_n$; and $x_n \notin F_n$. So $x_m \neq x_n$. Hence $n \mapsto x_n$ is an injection $\mathbb{N} \rightarrow X$. Its image $\{x_n : n \in \mathbb{N}\} \subseteq X$ is equinumerous with $\mathbb{N}$, since $n \mapsto x_n$ is a bijection from $\mathbb{N}$ onto it. This proves the theorem. ■

The proof shows where the axiom enters and why one $\exists E$ per step is not enough: $\exists E$ names one element for one set, while the recursion needs g as a single function defined on every stage at once. Each F_n is then determined by c ; the only arbitrary thing is c itself.

A non-example by a minimal change. Run the same construction on the finite set $X = \{a, b\}$, with $a \neq b$. The choice function c and the recursion still make sense, since g is defined on every $F \subseteq X$. But $F_1 = \{c(X)\}$ has one element, F_2 adds the other, so $F_2 = X$, and from then on g leaves F_n unchanged. So $x_2 = c(X \setminus F_2) = c(\emptyset)$ is not defined. The step that fails is “ F_n is finite and X is not, so $F_n \neq X$ ”, exactly where the hypothesis that X is infinite is used.

Remark. Compare *Analysis on the Real Line*, Chapter 1’s proof that the real numbers cannot be listed. There, a given list is used to build nested closed intervals, and each interval is obtained from the previous one by an explicit rule (keep an outer third that misses the next point of the list), so no choice principle is involved. Here X is an arbitrary infinite set, and there is no rule available for picking an element of an arbitrary non-empty subset of it; the choice function **is** the rule.

For a particular X a rule may exist: for $X \subseteq \mathbb{N}$, “the least element” is one (Theorem 2.1.3), and then the proof needs no axiom. That the theorem cannot be proved in general from the other rules of set formation alone is a result of set theory; we state it without proof, and nothing depends on it.

3.6.2 Consequences: the smallest infinite size

The theorem says $|\mathbb{N}| \leq |X|$ for every infinite X (Definition 3.5.1). Three consequences follow at once.

- **$\mathbb{N}$ is the smallest infinite size.** If X is infinite and $|X| \leq |\mathbb{N}|$, then $|X| = |\mathbb{N}|$, by Theorem 3.5.3 and the theorem. So no infinite set is strictly smaller than $\mathbb{N}$.
- **Countable sets are exactly the sets of size at most $|\mathbb{N}|$,** by Theorem 3.2.2, (i) $\iff$ (ii). Combining with the first point: a countable set is finite or equinumerous with $\mathbb{N}$, which Theorem 3.2.2 (iii) already said.
- **An uncountable set is strictly bigger than $\mathbb{N}$.** If X is uncountable, then X is infinite (a finite set is countable, by Theorem 3.2.2), so $|\mathbb{N}| \leq |X|$; and $|\mathbb{N}| = |X|$ would make X countable. So $|\mathbb{N}| < |X|$. For example $|\mathbb{N}| < |\mathbb{R}|$, since $\mathbb{R}$ is uncountable (Theorem 3.4.4).

Warning. “Infinite” and “uncountable” are different words. $\mathbb{Q}$ is infinite (it contains $\mathbb{N}$, which is infinite by Example (Hilbert’s Hotel), and a subset of a finite set is finite, by Proposition 3.1.6) and countable (Theorem 3.2.8). Likewise “ X contains a copy of $\mathbb{N}$ ” is not “ X is equinumerous with $\mathbb{N}$ ”: $\mathbb{R}$ contains $\mathbb{N}$, and $|\mathbb{N}| < |\mathbb{R}|$ (as above).

Quick check. Let X be uncountable. Is $X \setminus \{x\}$ uncountable for every $x \in X$? Why must X have an element at all?

Solution. $X \neq \emptyset$, because $\emptyset$ is countable by Definition 3.2.1. For $x \in X$, if $X \setminus \{x\}$ were countable, then $X = (X \setminus \{x\}) \cup \{x\}$ would be a union of two countable sets, countable by Proposition 3.2.5. So $X \setminus \{x\}$ is uncountable.

3.6.3 Infinite sets are the same size as a proper part

Section 3.1 proved that a finite set is never equinumerous with a proper subset (Theorem 3.1.9), and Example (Hilbert’s Hotel) showed $\mathbb{N}$ doing exactly that. With Theorem 3.6.1, the hotel can be built inside any infinite set: run it on the copy of $\mathbb{N}$, and leave everything else alone.

Theorem 3.6.2. Infinite Sets and Proper Subsets

A set X is infinite if and only if $|X| = |Y|$ for some proper subset $Y \subsetneq X$.

Idea. One direction is already done: Section 3.1 showed that finite sets have no proper subset of their own size. For the other, take the copy $x_0, x_1, x_2, \dots$ of $\mathbb{N}$ inside X and move every guest one room up, $x_n \mapsto x_{n+1}$, as $n \mapsto n + 1$ does in the Quick check after Example

(Hilbert's Hotel). Room x_0 is vacated, and every point outside the copy stays where it is.

Proof. ($\Leftarrow$) This is the contrapositive of [Theorem 3.1.9](#): a finite set is not equinumerous with any proper subset, so a set that is equinumerous with one is not finite.

($\Rightarrow$) Let X be infinite. By [Theorem 3.6.1](#) there is an injection $n \mapsto x_n$ from $\mathbb{N}$ to X ; let $S := \{x_n : n \in \mathbb{N}\}$. Each element of S is x_n for exactly one $n \in \mathbb{N}$, by injectivity. Define $h: X \rightarrow X \setminus \{x_0\}$ by

$$h(x) := \begin{cases} x_{n+1} & \text{if } x = x_n \text{ with } n \in \mathbb{N}, \\ x & \text{if } x \notin S. \end{cases}$$

This is a function into $X \setminus \{x_0\}$: $x_{n+1} \neq x_0$ because $n+1 \neq 0$, and an $x \notin S$ is not x_0 .

h is injective. Two points of S : $x_{m+1} = x_{n+1}$ gives $m+1 = n+1$, so $m = n$. Two points outside S : h is the identity there. One of each: $h(x_n) = x_{n+1} \in S$, while $h(x) = x \notin S$, so the values differ.

h is onto $X \setminus \{x_0\}$. Let $y \in X \setminus \{x_0\}$. If $y \notin S$, then $y = h(y)$. If $y \in S$, then $y = x_m$ with $m \neq 0$, so $n := m-1 \in \mathbb{N}$ ([Lemma 2.1.1](#) (c)) and $y = x_{n+1} = h(x_n)$.

So h is a bijection, and $Y := X \setminus \{x_0\}$ is a proper subset of X , as $x_0 \in X \setminus Y$, with $|X| = |Y|$. This proves the theorem. ■

Together with Section 3.1, this gives a characterization of “infinite” that never mentions $\mathbb{N}$ or counting: a set is infinite precisely when some proper part of it is as big as the whole. The finite intuition “a part is smaller than the whole” is exactly the property that separates finite sets from infinite ones. The direction ($\Rightarrow$) used [Axiom 3.3.1](#), through [Theorem 3.6.1](#); the direction ($\Leftarrow$) did not.

Quick check. Name a proper subset of $\mathbb{R}$ that is equinumerous with $\mathbb{R}$, and give a bijection.

Solution. $(-1, 1)$, with the bijection $x \mapsto x/(1 + |x|)$ of [Example \(The real line and its intervals\)](#). (Any interval with more than one point other than $\mathbb{R}$ itself works, by the same example.) No choice is needed for $\mathbb{R}$: the bijection is a formula.

3.6.4 Exercises

3.6.4.1 A. Check your understanding

Exercise. A1

- Where does the proof of [Theorem 3.6.1](#) use [Axiom 3.3.1](#)? Why is one $\exists E$ per step not enough?
- True or false, with a reason: every infinite subset of $\mathbb{R}$ contains a countably infinite

subset; every uncountable set contains an uncountable proper subset.

- (c) Write down the bijection $X \rightarrow X \setminus \{x_0\}$ of the proof of [Theorem 3.6.2](#) for $X = \mathbb{Z}$ and $x_n = n$ for $n \in \mathbb{N}$.
- (d) For $X = \mathbb{N}$, is choice needed to find a countably infinite subset?

Solution.

- (a) At the line “By [Axiom 3.3.1](#) there is a function c with $c(S) \in S$ for every non-empty $S \subseteq X$.” One $\exists E$ names one element of one non-empty set; that would give x_0 , then x_1 , and any finite number of steps. But [Theorem 2.3.1](#) needs the step as a **function** g , defined once for every stage F at once, before the sequence exists. Defining g means picking an element of $X \setminus F$ for every such F simultaneously, and that is what the axiom provides.
- (b) Both are true. The first is [Theorem 3.6.1](#) applied to the infinite subset. For the second, let X be uncountable; then $X \neq \emptyset$, and for $x \in X$, $X \setminus \{x\}$ is an uncountable proper subset, by the Quick check after the warning above.
- (c) $h(n) = n + 1$ for $n \in \mathbb{N}$, and $h(k) = k$ for $k < 0$. It maps $\mathbb{Z}$ onto $\mathbb{Z} \setminus \{0\}$: a positive m is $h(m - 1)$, and a negative k is $h(k)$.
- (d) No. $\mathbb{N}$ itself is countably infinite, via $\text{id}_{\mathbb{N}}$. (More generally, an infinite $X \subseteq \mathbb{N}$ can be listed in increasing order by [Lemma 3.2.1](#), a rule.)

3.6.4.2 B. Practice

Exercise. B1: Removing finitely many points

Let X be infinite and let $F \subseteq X$ be finite. Prove that $|X \setminus F| = |X|$.

Solution. First, $X \setminus F$ is infinite: otherwise $X = (X \setminus F) \cup F$ would be a union of two disjoint finite sets, finite by [Proposition 3.1.5](#).

By [Theorem 3.6.1](#) applied to $X \setminus F$ (this uses [Axiom 3.3.1](#)), there is an injection $n \mapsto s_n$ from $\mathbb{N}$ to $X \setminus F$; let $S := \{s_n : n \in \mathbb{N}\}$. Let F have k elements, and let $\varphi : \{1, \dots, k\} \rightarrow F$ be a bijection ([Definition 3.1.2](#)). No two of $X \setminus (F \cup S)$, F and S meet, and together they make up X ; and $X \setminus F$ is the union of the disjoint sets $X \setminus (F \cup S)$ and S . Define $h : X \rightarrow X \setminus F$ by

$$h(x) := \begin{cases} x & \text{if } x \notin F \cup S, \\ s_{i-1} & \text{if } x = \varphi(i), i \in \{1, \dots, k\}, \\ s_{n+k} & \text{if } x = s_n, n \in \mathbb{N}. \end{cases}$$

Each x falls under one case, with one i or one n , since φ and $n \mapsto s_n$ are injective. The second case sends F bijectively onto $\{s_0, \dots, s_{k-1}\}$, and the third sends S bijectively onto $\{s_m : m \geq k\}$; these two image sets are disjoint and together make up S . The first case is the identity on $X \setminus (F \cup S)$. So h is a bijection, and $|X \setminus F| = |X|$. (For $k = 0$, $F = \emptyset$ and h is the identity.)

Exercise. B2: Absorbing a countable set

Let X be infinite, and let C be countable with $C \cap X = \emptyset$. Prove that $|X \cup C| = |X|$.

Hint: absorb C into a countably infinite subset of X .

Solution. By [Theorem 3.6.1](#) (using [Axiom 3.3.1](#)) there is a countably infinite $S \subseteq X$. The set $S \cup C$ is countable by [Proposition 3.2.5](#), and infinite, because it contains S (a subset of a finite set is finite, [Proposition 3.1.6](#)), which is infinite: if $\varphi: \{1, \dots, k\} \rightarrow S$ were a bijection and $\beta: \mathbb{N} \rightarrow S$ is a bijection given by the theorem, then $\beta^{-1} \circ \varphi$ would be a bijection $\{1, \dots, k\} \rightarrow \mathbb{N}$ ([Proposition 3.1.1](#), [Theorem 1.4.4](#)), against [Example \(Hilbert's Hotel\)](#). By [Theorem 3.2.2](#) (iii), $S \cup C$ is countably infinite. So $|S \cup C| = |\mathbb{N}| = |S|$, and [Proposition 3.1.1](#) gives a bijection $\psi: S \cup C \rightarrow S$.

Now $X \cup C$ is the union of the disjoint sets $X \setminus S$ and $S \cup C$ (disjoint because $C \cap X = \emptyset$), and X is the union of the disjoint sets $X \setminus S$ and S . Define $h: X \cup C \rightarrow X$ by $h(x) := x$ for $x \in X \setminus S$ and $h(x) := \psi(x)$ for $x \in S \cup C$. It is a bijection on each piece, onto the corresponding piece, so it is a bijection, and $|X \cup C| = |X|$.

Exercise. B3: A big hole

Give an example of an infinite set X and a proper subset $Y \subsetneq X$ with $|Y| = |X|$ and $X \setminus Y$ infinite. Check each property, and list the properties any answer must have.

Solution. An answer must have: (1) $Y \subseteq X$; (2) $Y \neq X$; (3) a bijection $X \rightarrow Y$, or injections both ways; (4) X infinite; (5) $X \setminus Y$ infinite. (The hole $X \setminus Y$ in the proof of [Theorem 3.6.2](#) was the single point x_0 ; this asks for more.)

Example: $X = \mathbb{N}$, $Y = 2\mathbb{N}$. (1), (2): $2\mathbb{N} \subseteq \mathbb{N}$ and $1 \notin 2\mathbb{N}$. (3): $n \mapsto 2n$ is a bijection $\mathbb{N} \rightarrow 2\mathbb{N}$ ([Example \(Hilbert's Hotel\)](#)). (4): $\mathbb{N}$ is infinite ([Example \(Hilbert's Hotel\)](#)). (5): $X \setminus Y$ is the set of odd numbers (every $m \in \mathbb{N}$ is even or odd, not both, by [Theorem 2.2.3](#)), and $n \mapsto 2n + 1$ is a bijection from $\mathbb{N}$ onto it; if it were finite, $\mathbb{N}$ would be finite too (compose with a bijection from $\{1, \dots, k\}$, [Theorem 1.4.4](#)), which it is not.

Another correct answer: $X = \mathbb{R}$, $Y = (0, \infty)$. (1), (2): $(0, \infty) \subseteq \mathbb{R}$ and $0 \notin (0, \infty)$. (3): [Example \(The real line and its intervals\)](#). (4), (5): $n \mapsto -n$ is a bijection from $\mathbb{N}$ onto $\{-n : n \in \mathbb{N}\}$, so this set is infinite, as in (5) above. It lies inside $X \setminus Y = (-\infty, 0] \subseteq \mathbb{R}$, so by [Proposition 3.1.6](#) neither $(-\infty, 0]$ nor $\mathbb{R}$ is finite.

3.6.4.3 C. Going deeper**Exercise. C1: Increasing functions**

A function $f: \mathbb{N} \rightarrow \mathbb{N}$ is **strictly increasing** if $f(m) < f(n)$ whenever $m < n$. Let I be the set of strictly increasing functions $\mathbb{N} \rightarrow \mathbb{N}$. Prove that I is uncountable.

Below, the proof is started; the second half is blank. *Name the move, then complete it.*

Let $n \mapsto f_n$ be any function from $\mathbb{N}$ to I . By [Theorem 2.3.1](#), applied with the set

$\mathbb{N}$, the starting value $f_0(0) + 1$ and the function $(n, x) \mapsto \max\{x, f_{n+1}(n+1)\} + 1$ from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ (a step that depends on n), there is a function $g: \mathbb{N} \rightarrow \mathbb{N}$ with

$$g(0) = f_0(0) + 1, \quad g(n+1) = \max\{g(n), f_{n+1}(n+1)\} + 1.$$

Solution (Hint 1). Two things must be shown about g : that it lies in I , and that it differs from each f_n somewhere. Where?

Solution (Hint 2). The second half begins: “For every $n \in \mathbb{N}$, $g(n+1) \geq g(n) + 1 > g(n)$.”

Solution. The move is **the diagonal**: given any list, build an object that differs from the n -th entry at the n -th place. Here the object must also stay inside I , which is why $g(n+1)$ is pushed above $g(n)$ as well as above $f_{n+1}(n+1)$.

Completion. For every $n \in \mathbb{N}$, $g(n+1) = \max\{g(n), f_{n+1}(n+1)\} + 1 \geq g(n) + 1 > g(n)$. From this, $g(m) < g(n)$ whenever $m < n$, by induction on n ([Theorem 2.2.1](#) with base $m+1$): $g(m) < g(m+1)$, and $g(m) < g(n) < g(n+1)$. So $g \in I$.

Next, $g(n) > f_n(n)$ for every $n \in \mathbb{N}$: for $n = 0$, $g(0) = f_0(0) + 1$; for $n = k+1$, $g(k+1) \geq f_{k+1}(k+1) + 1$. So $g(n) \neq f_n(n)$, and hence $g \neq f_n$, for every $n \in \mathbb{N}$. So $g \in I$ is not a value of $n \mapsto f_n$, and no function $\mathbb{N} \rightarrow I$ is onto. Since $I \neq \emptyset$ (it contains $\text{id}_{\mathbb{N}}$), I is not countable.

Marking scheme. (10 points) - 1: names the diagonal move. - 4: shows g is strictly increasing: 2 for the step $g(n) < g(n+1)$ for every n , and 2 for the passage to all $m < n$. - 3: shows $g(n) > f_n(n)$ for every n , treating $n = 0$ and $n = k+1$ separately, and concludes $g \neq f_n$. - 2: concludes that no function $\mathbb{N} \rightarrow I$ is onto and, noting $I \neq \emptyset$ ($\text{id}_{\mathbb{N}} \in I$), that I is not countable. Fatal errors (0 for the part concerned): taking $g(n) < g(n+1)$ alone as strict increase, with no passage to all $m < n$; inferring $g \neq f_n$ without exhibiting a point where they differ, or exhibiting one only for $n \geq 1$.

The max in g is needed: the plain diagonal $g(n) := f_n(n) + 1$ need not be strictly increasing (for $f_0(n) = 5 + n$ and $f_1(n) = n$, it gives $g(0) = 6 > g(1) = 2$).

Exercise. C2: The irrationals

Deduce from [Exercise \(B2: Absorbing a countable set\)](#) that $|\mathbb{R} \setminus \mathbb{Q}| = |\mathbb{R}|$.

Solution. The set $\mathbb{R} \setminus \mathbb{Q}$ is uncountable: otherwise $\mathbb{R} = (\mathbb{R} \setminus \mathbb{Q}) \cup \mathbb{Q}$ would be a union of two countable sets ([Theorem 3.2.8](#)), countable by [Proposition 3.2.5](#), against [Theorem 3.4.4](#). An uncountable set is infinite, since finite sets are countable ([Theorem 3.2.2](#)). Now apply [Exercise \(B2: Absorbing a countable set\)](#) with $X := \mathbb{R} \setminus \mathbb{Q}$ and $C := \mathbb{Q}$, which is countable and disjoint from X : $|(\mathbb{R} \setminus \mathbb{Q}) \cup \mathbb{Q}| = |\mathbb{R} \setminus \mathbb{Q}|$, and the left side is $|\mathbb{R}|$.

3.6.4.4 R. Review

Exercise. R1

Prove that every integer $n \geq 1$ is a sum of distinct powers of 2: there are an integer $r \geq 1$ and natural numbers $k_1 > k_2 > \cdots > k_r$ with $n = 2^{k_1} + 2^{k_2} + \cdots + 2^{k_r}$.

Solution. We use [Theorem 2.2.2](#) with $n_0 = 1$. Let $n \geq 1$, and suppose every integer m with $1 \leq m < n$ is such a sum.

The largest power of 2 not above n . The set $T := \{k \in \mathbb{N} : 2^k > n\}$ is non-empty, since $2^n \geq 1 + n > n$ by [Theorem 2.3.7](#) with $x = 1$. By [Theorem 2.1.3](#) it has a least element m . Since $2^0 = 1 \leq n$, $0 \notin T$, so $m \geq 1$, and $K := m - 1 \in \mathbb{N}$ ([Lemma 2.1.1](#) (c)). As $K < m$, $K \notin T$; so

$$2^K \leq n < 2^{K+1}.$$

The step. Let $s := n - 2^K$, so $0 \leq s < 2^{K+1} - 2^K = 2^K$. If $s = 0$, then $n = 2^K$, a sum with $r = 1$. If $s \geq 1$, then $s < n$, because $2^K \geq 1$. By the hypothesis, $s = 2^{k_1} + \cdots + 2^{k_r}$ with $k_1 > \cdots > k_r$. Each $2^{k_i} \leq s < 2^K$, so $k_i < K$ (if $k_i \geq K$, then $2^{k_i} \geq 2^K$ by [Proposition 2.3.4](#) (e)). Hence

$$n = 2^K + 2^{k_1} + \cdots + 2^{k_r}, \quad K > k_1 > \cdots > k_r,$$

a sum of distinct powers of 2. By [Theorem 2.2.2](#), every $n \geq 1$ is such a sum. For example $13 = 2^3 + 2^2 + 2^0$.

The base 1 is the right one: 0 is not a sum of one or more powers of 2, since each power is positive.

Exercise. R2

Show that $\forall x \exists y R(x, y) \not\vdash \exists y \forall x R(x, y)$ by a counter-model whose domain is $\mathbb{N}$. Say in words what the two sentences say when $R(x, y)$ is read as “ $x < y$ ”.

Solution. Let $\mathfrak{M}$ have domain $\mathbb{N}$ and $R^{\mathfrak{M}} := \{(d, e) \in \mathbb{N} \times \mathbb{N} : d < e\}$. Both formulas are sentences, so no assignment matters.

The premise is true. $\forall x \exists y (x < y)$ says “every natural number has a larger one”. For each $d \in \mathbb{N}$, $e := d + 1 \in \mathbb{N}$ has $d < e$.

The conclusion is false. $\exists y \forall x (x < y)$ says “some natural number is larger than every natural number”. For any $e \in \mathbb{N}$, the choice $d := e$ gives $d < e$ false, so no e works.

So $\mathfrak{M}$ is a counter-model, and $\forall x \exists y R(x, y) \not\vdash \exists y \forall x R(x, y)$. By the soundness statement of *Logic and Proof*, Section 4.11 (the subsection “From counter-models to non-derivability”), a counter-model shows non-derivability: $\forall x \exists y R(x, y) \not\vdash \exists y \forall x R(x, y)$.

This is the converse of *Logic and Proof*, Proposition 5.6.2 with the variables renamed and the two places of R exchanged (read $R(x, y)$ as $R(y, x)$), and the counter-model is the one of *Logic and Proof*, Example (The swap of $\exists$ and $\forall$): there the domain was $\mathbb{Z}$ and R was read as $>$; with the roles of the variables exchanged, R read as $<$ and the domain cut down to $\mathbb{N}$, the same witnesses work, as checked above. In the premise the larger number y is chosen

after x and depends on it; the conclusion asks for one y in advance.

Exercise. R3

For an integer $n \geq 1$, how many truth functions with n inputs are there? Prove your answer.

Solution. There are 2^{2^n} . A truth function with n inputs (Logic and Proof, Definition 4.4.4) assigns a truth value to each row $(a_1, \dots, a_n)$; read a row as the function $i \mapsto a_i$ on $\{1, \dots, n\}$, and let L be the set of rows.

There are 2^n rows. The map sending a row to $\{i \in \{1, \dots, n\} : a_i = \text{T}\}$ is a bijection $L \rightarrow \mathcal{P}(\{1, \dots, n\})$: a row is determined by the places where it is T, and every set of places arises. Since $\{1, \dots, n\}$ has n elements, Proposition 3.1.8 gives 2^n subsets, so L has 2^n elements (compose the bijections).

There are 2^{2^n} truth functions. In the same way, a truth function f is determined by the set $\{\rho \in L : f(\rho) = \text{T}\} \subseteq L$, and every subset of L arises; so the truth functions are in bijection with $\mathcal{P}(L)$, which has 2^{2^n} elements by Proposition 3.1.8 again. For $n = 1, 2, 3$ this gives 4, 16, 256.

Exercise. R4

A function $f: \mathbb{N} \rightarrow \mathbb{N}$ is **eventually zero** if there is $N \in \mathbb{N}$ with $f(n) = 0$ for every $n \geq N$. Is the set of eventually zero functions countable?

Solution. Yes. The tool is a **countable union of countable sets**. For $N \in \mathbb{N}$ let Z_N be the set of $f: \mathbb{N} \rightarrow \mathbb{N}$ with $f(n) = 0$ for every $n \geq N$; the set of eventually zero functions is $Z := \bigcup_{N \in \mathbb{N}} Z_N$. Each Z_N is countable:

- Z_0 has one element, the zero function, so it is countable.
- For $N = 1$, $f \mapsto f(0)$ is a bijection $Z_1 \rightarrow \mathbb{N}$: a function in Z_1 is determined by its value at 0, and every value occurs.
- For $N \geq 2$, $f \mapsto (f(0), \dots, f(N-1))$ is a bijection $Z_N \rightarrow \mathbb{N}^N$, by the same reasoning and the characteristic property of tuples; $\mathbb{N}^N$ is countable by Corollary 3.2.7, and so is Z_N (compose a surjection $\mathbb{N} \rightarrow \mathbb{N}^N$ with the inverse bijection, Proposition 3.2.4).

By Theorem 3.3.3, Z is countable. (That theorem uses Axiom 3.3.1; here it could be avoided, since each Z_N comes with an explicit listing, but the question did not ask for that.)

Why the look-alike fails. A diagonal argument, as in Theorem 3.4.1, builds from a list $f_0, f_1, \dots$ of eventually zero functions the function $g(n) := f_n(n) + 1$. It differs from every f_n , but $g(n) \geq 1$ for every n , so g is **not** eventually zero: it is not in Z , and nothing is contradicted. A diagonal argument proves uncountability only when the diagonal object stays inside the set.

3.7 Gate 0: Logic and Sets Examination

The five chapters of *Logic and Proof* and the first three chapters of this volume taught one skill in many forms: reading a mathematical statement exactly, and proving it in sentences that a reader can check line by line. This section is an examination on all of it, and the test of whether *Analysis on the Real Line* will be readable: from its Chapter 1 on, every proof assumes that negating a quantified statement, naming a witness, running an induction and chasing an element through a preimage are routine. A reader who comes to the book by a reading path that begins with *Analysis on the Real Line*, such as a first course in real analysis, takes this Gate first, as an entry test, and goes back only to the sections that the remediation map at the end names. Grading your own work is a skill too; the front matter's page on getting feedback without a teacher describes the loop of attempting, writing, grading against the scheme and only then reading the solution.

3.7.1 Instructions

Algorithm.

- **Time: 2 hours.** Closed book, no notes; paper and pen. Attempt all six problems. The times in brackets are a guide; they add up to 120 minutes.
- Write complete proofs in plain prose, at level 3 of Logic and Proof, Definition 8.3.1. Rule tags are allowed where a problem asks for them.
- Each problem is worth 10 points, 60 in all. Its solution is folded. Open a solution only after the two hours, and **grade before you read it**: apply the marking scheme at the end of the solution to your own script, point by point, and only then read the book's proof.
- **Pass line: 70% overall (42 of 60 points), and no core problem at 0.** The core problems are marked **core**: P1, P2, P4 and P5. They test what *Analysis on the Real Line*, Chapter 1 uses from its first page: quantifiers and their negation, turning a derivation into prose, induction and recursion, sets and functions. P3 and P6 are not core.
- If you fail, follow the remediation map, then wait a week and retake the whole Gate (the same problems) before starting *Analysis on the Real Line*.

The definition drill comes first and is not scored. It is the one retrieval list for those eight chapters: every major definition in them, to be written from memory before the clock starts.

Warning. A proof that reads well is not thereby correct, and grading your own script is where a reader working alone is most easily fooled. Grade the step, not the impression: a marking-scheme line such as “ y given as a function of δ ” is either on your page or it is not. A script that says “choose y close to 1” in Problem P1 has not given y as a function of δ , however clear the idea, and loses those points.

3.7.2 Definition drill

Quick check. State each of the following **in symbols**, with every quantifier and every bound written out, as the book defines it. Allow about fifteen minutes, before the two hours start.

- (Q1) $\Gamma \vdash \psi$.
- (Q2) $\Gamma \models \psi$, for formulas of propositional logic.
- (Q3) φ is a tautology.
- (Q4) The side conditions of $\forall I$ and of $\exists E$ on the flagged name x_0 .
- (Q5) $\exists!x \varphi$.
- (Q6) $A \subseteq B$.
- (Q7) $\bigcap_{i \in I} A_i$, for a family $(A_i)_{i \in I}$ of subsets of a set X .
- (Q8) A function $f: X \rightarrow Y$.
- (Q9) $f: X \rightarrow Y$ is injective; f is surjective.
- (Q10) The image $f(A)$ and the preimage $f^{-1}(B)$.
- (Q11) An equivalence relation on X ; a partition of X .
- (Q12) A greatest element and a maximal element of $A \subseteq X$, for a partial order $\preceq$ on X .
- (Q13) A least upper bound of A in a partially ordered set $(X, \preceq)$.
- (Q14) The principle of induction; the well-ordering of $\mathbb{N}$.
- (Q15) A is countable.
- (Q16) $|A| = |B|$; $|A| \leq |B|$.
- (Q17) The axiom of choice.

Solution. The answers are the definitions as the book states them; the label after each answer is where to look when an answer is wrong.

- (Q1) For a list $\Gamma = \varphi_1, \dots, \varphi_n$ ($n \geq 0$): $\Gamma \vdash \psi$ if **there is** a derivation whose premises all belong to Γ and whose last line is ψ (Logic and Proof, Definition 1.2.3).
- (Q2) $\Gamma \models \psi$ if for **every** valuation v : if $v(\varphi) = T$ for every φ in Γ , then $v(\psi) = T$ (Logic and Proof, Definition 4.5.3).
- (Q3) $v(\varphi) = T$ for **every** valuation v (Logic and Proof, Definition 4.5.1).
- (Q4) $\forall I$, closing a flagged box with name x_0 and writing $\forall x \varphi$ on line k : at line k , x_0 appears in no premise, in no assumption of a box still open at line k , and not in $\forall x \varphi$ (Logic and Proof, Definition 5.4.1). $\exists E$, from $\exists x \varphi$ and a flagged box with name x_0 assuming $\varphi[x_0/x]$ and ending in χ , writing χ on line k : x_0 does not appear in χ , in $\exists x \varphi$, in any premise, or in any assumption of a box still open at line k (Logic and Proof, Definition 5.6.1). “Appears” means written anywhere, after a quantifier symbol included.
- (Q5) $\exists x (\varphi \wedge \forall y (\varphi[y/x] \rightarrow y = x))$, where y is a variable other than x that does not appear in φ (Logic and Proof, Definition 5.8.1).
- (Q6) $\forall x (x \in A \rightarrow x \in B)$ (Definition 1.1.2).
- (Q7) $\bigcap_{i \in I} A_i := \{x \in X : \forall i \in I (x \in A_i)\}$; for $I = \emptyset$ it is X (Definition 1.2.4).
- (Q8) A relation $G \subseteq X \times Y$, the graph, with $\forall x \in X \exists! y \in Y (x, y) \in G$, together with the codomain Y : formally $f = (G, Y)$ (Definition 1.4.1).
- (Q9) Injective: $\forall x_1 \in X \forall x_2 \in X (f(x_1) = f(x_2) \rightarrow x_1 = x_2)$ (Definition 1.4.3). Surjective: $\forall y \in Y \exists x \in X f(x) = y$ (Definition 1.4.4).

- (Q10) For $A \subseteq X$: $y \in f(A) \iff y \in Y \wedge \exists a \in A f(a) = y$ (Definition 1.6.1). For $B \subseteq Y$: $x \in f^{-1}(B) \iff x \in X \wedge f(x) \in B$ (Definition 1.6.2).
- (Q11) A relation $\sim$ on X with (Eq1) $\forall x \in X x \sim x$; (Eq2) $\forall x, y \in X (x \sim y \rightarrow y \sim x)$; (Eq3) $\forall x, y, z \in X (x \sim y \wedge y \sim z \rightarrow x \sim z)$ (Definition 2.5.1). A set $\mathcal{P}$ of subsets of X with (P1) $\forall P \in \mathcal{P} P \neq \emptyset$; (P2) $\forall P, Q \in \mathcal{P} (P \neq Q \rightarrow P \cap Q = \emptyset)$; (P3) $\bigcup \mathcal{P} = X$ (Definition 2.5.3).
- (Q12) g is a greatest element: $g \in A \wedge \forall a \in A a \preceq g$ (Definition 2.6.3). m is a maximal element: $m \in A \wedge \neg \exists a \in A m \prec a$, where $m \prec a$ means $m \preceq a \wedge m \neq a$ (Definition 2.6.4).
- (Q13) $s \in X$ with (S1) $\forall a \in A a \preceq s$, and (S2) $\forall u \in X ((\forall a \in A a \preceq u) \rightarrow s \preceq u)$ (Definition 2.6.5).
- (Q14) For every property P of natural numbers: if $P(0)$ and $\forall n \in \mathbb{N} (P(n) \rightarrow P(n+1))$, then $\forall n \in \mathbb{N} P(n)$ (Logic and Proof, Axiom 4.2.1). Every non-empty $S \subseteq \mathbb{N}$ has a least element: $S \neq \emptyset \rightarrow \exists m \in S \forall s \in S m \leq s$ (Theorem 2.1.3).
- (Q15) $A = \emptyset$, or there is $f: \mathbb{N} \rightarrow A$ with $\forall a \in A \exists n \in \mathbb{N} f(n) = a$ (Definition 3.2.1).
- (Q16) $|A| = |B|$ if there is a bijection $A \rightarrow B$ (Definition 3.1.1); $|A| \leq |B|$ if there is an injection $A \rightarrow B$ (Definition 3.5.1).
- (Q17) For every set I and every family $(A_i)_{i \in I}$ of sets with $\forall i \in I A_i \neq \emptyset$, there is $c: I \rightarrow \bigcup_{i \in I} A_i$ with $\forall i \in I c(i) \in A_i$ (Axiom 3.3.1).

Count an item right only if every quantifier, every bound and the order of the quantifiers match. The commonest slips are a missing “ $\in X$ ” bound in (Q9), “ $\exists y$ ” for “ $\exists! y$ ” in (Q8), and, in (Q12), “ $\forall a \in A a \preceq m$ ” for maximal, which is the definition of greatest.

3.7.3 The problems

Exercise. P1: core

(20 minutes.) Say that a function $f: \mathbb{R} \rightarrow \mathbb{R}$ has **property L** if for every $x \in \mathbb{R}$ there is $\delta > 0$ such that for every $y \in \mathbb{R}$ with $|y - x| < \delta$, we have $f(y) \geq f(x)$.

- Write property L in symbols, with bounded quantifiers.
- Write its negation with no $\neg$ in front of a quantifier, a connective or a parenthesis.
- Prove that $f(x) = x^2$ does not have property L.
- Prove that the function $f: \mathbb{R} \rightarrow \mathbb{R}$ with $f(x) = 1$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$ (the indicator of $(0, \infty)$, with codomain $\mathbb{R}$) has property L, and that it does **not** have the property obtained from L by moving “there is $\delta > 0$ ” to the front.

Solution. (a) Property L is

$$\forall x \in \mathbb{R} \exists \delta > 0 \forall y \in \mathbb{R} (|y - x| < \delta \rightarrow f(y) \geq f(x)).$$

(b) Push the negation inward one symbol at a time, as in the procedure of *Logic and Proof*, Section 4.9 (“Negating a definition, mechanically”): *Logic and Proof*, Proposition 6.2.3 three times, the bound of each quantifier kept, and *Logic and Proof*, Proposition 6.2.2 once. Write

α for $|y - x| < \delta$ and β for $f(y) \geq f(x)$.

$$\begin{aligned} & \neg \forall x \in \mathbb{R} \exists \delta > 0 \forall y \in \mathbb{R} (\alpha \rightarrow \beta) \\ & \dashv \vdash \exists x \in \mathbb{R} \forall \delta > 0 \neg \forall y \in \mathbb{R} (\alpha \rightarrow \beta) \\ & \dashv \vdash \exists x \in \mathbb{R} \forall \delta > 0 \exists y \in \mathbb{R} \neg(\alpha \rightarrow \beta) \\ & \dashv \vdash \exists x \in \mathbb{R} \forall \delta > 0 \exists y \in \mathbb{R} (\alpha \wedge \neg \beta). \end{aligned}$$

The first step does two moves at once ($\neg \forall$ to $\exists \neg$, then $\neg \exists$ to $\forall \neg$). After the chain, $\neg \beta$, that is $\neg(f(y) \geq f(x))$, is replaced by $f(y) < f(x)$, a fact about the order of $\mathbb{R}$ and not a law of logic. The negation of L is therefore

$$\exists x \in \mathbb{R} \forall \delta > 0 \exists y \in \mathbb{R} (|y - x| < \delta \wedge f(y) < f(x)).$$

Note what stays: the bound $\delta > 0$ is kept, and only the quantifier in front of it changes; “ $\exists \delta \leq 0$ ” would be the error of negating the bound. The implication becomes a conjunction, never another implication.

(c) We prove the negation (b) for $f(x) = x^2$, with the witness $x = 1$. Let $\delta > 0$ be arbitrary, and put

$$y := 1 - \frac{1}{2} \min\{\delta, 1\}.$$

Since $0 < \min\{\delta, 1\} \leq 1$, we have $\frac{1}{2} \leq y < 1$. Then $|y - 1| = \frac{1}{2} \min\{\delta, 1\} \leq \frac{1}{2} \delta < \delta$, and, as $0 < y < 1$, $y^2 = y \cdot y < y \cdot 1 < 1 = f(1)$. So $|y - 1| < \delta$ and $f(y) < f(1)$. Since $\delta > 0$ was arbitrary, the negation of L holds at $x = 1$, and x^2 does not have property L.

The order of the choices is the order of the quantifiers in (b): x is fixed before δ , and y is a function of δ . The minimum with 1 is needed: $y := 1 - \delta/2$ fails for large δ , since $\delta = 4$ gives $y = -1$ and $y^2 = 1 \not< 1$.

(d) Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be the function of (d): $f(x) = 1$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. Let $x \in \mathbb{R}$.

Case 1: $x > 0$. Take $\delta := x$, which is positive. If $|y - x| < x$, then $y > x - x = 0$, so $f(y) = 1 = f(x)$, and $f(y) \geq f(x)$.

Case 2: $x \leq 0$. Take $\delta := 1$. For every y , $f(y) \in \{0, 1\}$, so $f(y) \geq 0 = f(x)$.

So f has property L. In Case 1, δ **depends on** x : points close to 0 need small δ .

The swapped property is $\exists \delta > 0 \forall x \in \mathbb{R} \forall y \in \mathbb{R} (|y - x| < \delta \rightarrow f(y) \geq f(x))$. We prove its negation, $\forall \delta > 0 \exists x \in \mathbb{R} \exists y \in \mathbb{R} (|y - x| < \delta \wedge f(y) < f(x))$. Let $\delta > 0$, and take $x := \delta/4$ and $y := -\delta/4$. Then $|y - x| = \delta/2 < \delta$, while $f(y) = 0 < 1 = f(x)$, since $y < 0 < x$. So no single δ serves every x . Here x depends on δ , which is legitimate, because δ is now fixed first.

Marking scheme. (10 points) - 2: (a) correct, with $\forall x \in \mathbb{R}$, $\exists \delta > 0$, $\forall y \in \mathbb{R}$ in this order and the condition $|y - x| < \delta$ as the hypothesis of an implication (or as a bound $\forall y \in \mathbb{R}$ with $|y - x| < \delta$). - 3: (b) correct: the bounds kept ($\forall \delta > 0$, not $\exists \delta \leq 0$), every quantifier switched, and the implication negated to a conjunction. - 2: (c) a witness x fixed **before** δ , and a y given as a function of δ that works for **every** $\delta > 0$ (any correct y , such as $1 - \frac{1}{2} \min\{\delta, 1\}$ or $1 - \min\{\delta/2, \frac{1}{2}\}$, earns the points). - 3: (d) 2 for the case split proving L, with a δ for each x ; 1 for refuting the swap, with x and y depending on δ . Fatal errors (0 for the step): a negation that keeps “ $\rightarrow$ ”; in (c), y chosen before δ or for one particular δ ; in the first half of (d), a single δ claimed for all x .

Exercise. P2: core

(20 minutes.) Here is a derivation of

$$\begin{aligned} &\forall z \exists y G(y, z), \quad \forall y \exists x F(x, y), \\ &\forall x \forall y \forall z ((F(x, y) \wedge G(y, z)) \rightarrow H(x, z)) \vdash \forall z \exists x H(x, z), \end{aligned}$$

with four justifications left blank, marked (i)–(iv).

1		$\forall z \exists y G(y, z)$	
2		$\forall y \exists x F(x, y)$	
3		$\forall x \forall y \forall z ((F(x, y) \wedge G(y, z)) \rightarrow H(x, z))$	
4		z_0 $\exists y G(y, z_0)$	$\forall E$ 1
5		y_0 $G(y_0, z_0)$	
6		$\exists x F(x, y_0)$	(i)
7		x_0 $F(x_0, y_0)$	
8		$\forall y \forall z ((F(x_0, y) \wedge G(y, z)) \rightarrow H(x_0, z))$	$\forall E$ 3
9		$\forall z ((F(x_0, y_0) \wedge G(y_0, z)) \rightarrow H(x_0, z))$	$\forall E$ 8
10		$(F(x_0, y_0) \wedge G(y_0, z_0)) \rightarrow H(x_0, z_0)$	$\forall E$ 9
11		$F(x_0, y_0) \wedge G(y_0, z_0)$	$\wedge I$ 7, 5
12		$H(x_0, z_0)$	$\rightarrow E$ 10, 11
13		$\exists x H(x, z_0)$	(ii)
14		$\exists x H(x, z_0)$	(iii)
15		$\exists x H(x, z_0)$	$\exists E$ 4, 5–14
16		$\forall z \exists x H(x, z)$	(iv)

- (a) Fill in the justifications (i)–(iv): the rule and the cited lines.
- (b) Explain why line 13 sits inside both of the boxes that begin at lines 5 and 7, and why each of these two boxes may close where it does.
- (c) Read $F(x, y)$ as “ $f(x) = y$ ”, $G(y, z)$ as “ $g(y) = z$ ” and $H(x, z)$ as “ $g(f(x)) = z$ ”, for functions $f: X \rightarrow Y$ and $g: Y \rightarrow Z$. Write the plain-prose proof that this derivation encodes, that the composite of two surjections is surjective, with one sentence for each box.

Solution. (a) (i) $\forall E$ 2, with the term y_0 for y . (ii) $\exists I$ 12, with the witness x_0 for x in $H(x, z_0)$. (iii) $\exists E$ 6, 7–13: the existential of line 6 and the box of x_0 . (iv) $\forall I$ 4–15: the flagged box of z_0 , whose range runs from its first line, line 4, to its last line, line 15 (Logic and Proof, Definition 5.4.1).

(b) Line 12, $H(x_0, z_0)$, contains the witness x_0 . A box of $\exists E$ may close only on a last line in which its own name does not appear (Logic and Proof, Definition 5.6.1). So the box of x_0 cannot close on line 12; the $\exists I$ of line 13 replaces x_0 by a bound variable first, and it

must come **before** that box closes, that is, inside it, and therefore inside the box of y_0 that contains it.

The box of x_0 closes at line 14. Its name x_0 does not appear in the last line $\exists x H(x, z_0)$, nor in the existential $\exists x F(x, y_0)$ of line 6, nor in the premises 1–3 (where only the bound variable x is written), nor in the one open assumption at line 14, $G(y_0, z_0)$ on line 5. The box of y_0 closes at line 15. Its name y_0 does not appear in $\exists x H(x, z_0)$, nor in $\exists y G(y, z_0)$ of line 4, nor in the premises; the only box still open at line 15 is the flagged box of z_0 , which has no assumption. Of the four clauses, the last-line clause is the one that shapes this derivation: the other three hold for any fresh names, while the last line had to be prepared by line 13. The names x_0 and y_0 may not leave their boxes because each names one witness, about which only its box’s assumption is known; outside, it would be treated as arbitrary, which is error (N3) of Logic and Proof, Example (What is not allowed). The name z_0 is different: it may appear in lines 14 and 15, because its box is closed by $\forall I$ at line 16, where it no longer appears.

(c) Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be surjective, and let $z \in Z$ be arbitrary (the box of z_0). Since g is surjective, there is $y \in Y$ with $g(y) = z$; fix one (the box of y_0 , lines 5–14, eliminating line 4). Since f is surjective, there is $x \in X$ with $f(x) = y$; fix one (the box of x_0 , lines 7–13, eliminating line 6). Then $g(f(x)) = g(y) = z$, so z is a value of $g \circ f$ (lines 8–13; premise 3 is the substitution of equals that the equation chain performs). Since z was arbitrary, $g \circ f$ is surjective, which is the surjectivity half of [Theorem 1.4.4](#). The derivation leaves the bounds $z \in Z$, $y \in Y$, $x \in X$ implicit; the prose restores them.

Marking scheme. (10 points) - 4: (a), one point per blank: the $\forall E$ with its line and instance term; the $\exists I$ with its line; the inner $\exists E$ citing line 6 and the box 7–13; the final $\forall I$ citing the box 4–15. - 2: (b) the fresh-name condition of $\exists E$ stated with all its clauses, in terms of where the name **appears**, for each box’s own name; and why x_0 and y_0 may not leave their boxes (so the $\exists I$ must come first). - 4: (c) “let $z \in Z$ ”, “choose y with $g(y) = z$ ”, “choose x with $f(x) = y$ ”, “then $g(f(x)) = z$ ”, in this order, each sentence carrying its box. Fatal errors (0 for the step): “let x be arbitrary” for an $\exists E$; prose in which y is chosen before z is fixed.

Exercise. P3

(20 minutes.) For each argument below, answer three questions. Is the conclusion true? Does the argument prove it? If not, which line fails, which entry of Logic and Proof, Example (What is not allowed) does it fall under, and what is a counterexample **to the illegal step** (not merely to the conclusion)?

- (a) *Claim:* $n^2 + n$ is even for every $n \in \mathbb{N}$. *Proof:* $0^2 + 0 = 0$ is even. Assume that $n^2 + n$ is even for all $n \in \mathbb{N}$. Then $(n+1)^2 + (n+1) = (n^2 + n) + 2(n+1)$ is even. By induction, the claim holds.
- (b) *Claim:* if $A_n \subseteq \mathbb{R}$ is bounded above for every $n \in \mathbb{N}$, then $\bigcup_{n \in \mathbb{N}} A_n$ is bounded above. *Proof:* For each n , let M_n be an upper bound of A_n . Let M be a number larger than every M_n . Then M is an upper bound of the union.

Solution. (a) **The conclusion is true.** Let $n \in \mathbb{N}$. By [Theorem 2.2.3](#) with divisor 2, $n = 2q + r$ with $q \in \mathbb{Z}$ and $r \in \{0, 1\}$. If $r = 0$, then $n^2 + n = n(n+1) = 2(q(n+1))$; if $r = 1$, then $n+1 = 2(q+1)$ and $n^2 + n = 2(n(q+1))$. In both cases the factor in brackets

is an integer (Proposition 2.1.2), so $2 \mid n^2 + n$ (Definition 2.2.1).

The argument does not prove it. The failing line is “Assume that $n^2 + n$ is even for all $n \in \mathbb{N}$ ”. That assumption is the claim itself, so everything after it is circular: entry (N9) of Logic and Proof, Example (What is not allowed), citing a line that is not available, here the goal. The step of an induction is a $\rightarrow$ I box inside a $\forall$ I box: fix one arbitrary $n \in \mathbb{N}$, assume that $n^2 + n$ is even **for this** n , and derive that $(n+1)^2 + (n+1)$ is even. With that one line changed, the rest of the argument is correct: the identity holds, since both sides equal $n^2 + 3n + 2$, and a sum of two even integers is even ($2a + 2b = 2(a+b)$).

A counterexample to the illegal step. The argument infers $\forall n P(n)$ from $P(0)$ and from “if $P(n)$ for all n , then $P(n+1)$ for all n ”. Take $P(n)$: “ $n = 0$ ”. Then $P(0)$ holds, and the second premise holds because its hypothesis “ $P(n)$ for all n ” is false ($P(1)$ fails), so the implication is vacuously true. Yet $\forall n P(n)$ is false. The conclusion of (a) is true, so it has no counterexample; the step has one.

(b) **The conclusion is false.** Take $A_n := \{n\}$. Each A_n is bounded above, by n . The union is $\mathbb{N}$, and $\mathbb{N}$ is not bounded above in $\mathbb{R}$: for every real M there is a natural number larger than M (a school fact here; *Analysis on the Real Line*, Chapter 1 proves it from its axiom).

The failing line is “Let M be a number larger than every M_n ”. From “for every n there is a number larger than M_n ” it passes to “there is a number larger than M_n for every n ”: from $\forall n \exists M$ to $\exists M \forall n$. This is entry (N2) of Logic and Proof, Example (What is not allowed), a witness that depends on n presented as uniform. It is the converse of Logic and Proof, Proposition 5.6.2, and the converse is not even valid: Logic and Proof, Example (The swap of $\exists$ and $\forall$) gives a counter-model. *A counterexample to the step:* $M_n := n$. For each n there is a number larger than M_n , namely $n+1$, and no real number is larger than every n .

The line before it is not the error. Choosing an upper bound M_n for every n at once, with no rule, is a use of the axiom of choice (Axiom 3.3.1), and the book allows it, flagged. And for **finitely** many indices the bad line becomes correct: $\max\{M_0, \dots, M_k\} + 1$ exists by Proposition 2.3.8, so a union of finitely many sets bounded above is bounded above.

Marking scheme. (10 points) - (a), 5 points. 1: says the conclusion is true, with a reason (a proof or a correct induction). 2: points to the line “assume ... for all n ”. 1: names the error, circularity, (N9). 1: gives a counterexample to the step, not to the conclusion. - (b), 5 points. 1: says the conclusion is false, with a witness such as $A_n = \{n\}$. 2: points to the line “let M be larger than every M_n ”. 1: names the quantifier swap $\forall n \exists M$ to $\exists M \forall n$, or (N2). 1: gives a counterexample to the step, not to the conclusion, such as $M_n = n$. Fatal errors (0 for the part): in (a), calling the argument correct; in (b), locating the error in the choice of the M_n .

Exercise. P4: core

(20 minutes.) Let $a_0 = 0$, $a_1 = 1$ and $a_{n+2} = 3a_{n+1} - 2a_n$ for $n \in \mathbb{N}$.

- Say which result guarantees that these equations define a sequence.
- Prove that $a_n = 2^n - 1$ for every $n \in \mathbb{N}$.
- Suppose instead that $a_1 = 2$, with a_0 and the rule unchanged. Compute a_2 and a_3 , and say exactly which line of your proof in (b) fails.

Solution. (a) Two-step recursion, [Corollary 2.3.3](#), with $X = \mathbb{R}$, $a = 0$, $b = 1$ and $h(n, x, y) := 3y - 2x$: there is exactly one sequence $(a_n)_{n \in \mathbb{N}}$ of reals with $a_0 = 0$, $a_1 = 1$ and $a_{n+2} = 3a_{n+1} - 2a_n$ for every $n \in \mathbb{N}$.

(b) The rule looks two steps back, so the induction carries two values at once. For $n \in \mathbb{N}$ let $Q(n)$ be the statement

$$a_n = 2^n - 1 \quad \text{and} \quad a_{n+1} = 2^{n+1} - 1.$$

Base case. $Q(0)$ says $a_0 = 2^0 - 1 = 0$ and $a_1 = 2^1 - 1 = 1$, and both hold.

Step. Let $n \in \mathbb{N}$ and suppose $Q(n)$. The second half of $Q(n)$ is the first half of $Q(n+1)$. For the second half, by the rule and both halves of $Q(n)$,

$$\begin{aligned} a_{n+2} &= 3a_{n+1} - 2a_n = 3(2^{n+1} - 1) - 2(2^n - 1) \\ &= 3 \cdot 2^{n+1} - 2^{n+1} - 1 = 2 \cdot 2^{n+1} - 1 = 2^{n+2} - 1, \end{aligned}$$

where $2 \cdot 2^n = 2^{n+1}$ and $2 \cdot 2^{n+1} = 2^{n+2}$ by [Definition 2.3.2](#). So $Q(n+1)$ holds. By Logic and Proof, Axiom 4.2.1, $Q(n)$ holds for every $n \in \mathbb{N}$, and in particular $a_n = 2^n - 1$ for every $n \in \mathbb{N}$. This is the pair induction of the two-step example in Section 2.3, after [Definition 2.3.2](#).

Alternatively, by strong induction (Logic and Proof, Theorem 4.2.2): let $n \in \mathbb{N}$ and suppose $a_k = 2^k - 1$ for every $k < n$. If $n = 0$ or $n = 1$, check the value directly. If $n \geq 2$, then $n-1$ and $n-2$ are natural numbers ([Lemma 2.1.1](#) (c), twice) smaller than n , so the hypothesis applies to both, and the same computation with $n-2$ in place of n gives $a_n = 2^n - 1$. The two values $n = 0, 1$ are exactly where the hypothesis would be used out of range.

(c) Now $a_2 = 3 \cdot 2 - 2 \cdot 0 = 6$ and $a_3 = 3 \cdot 6 - 2 \cdot 2 = 14$, while $2^2 - 1 = 3$ and $2^3 - 1 = 7$. The failing line is the base case at $n = 1$: in the proof of (b), the half “ $a_1 = 2^1 - 1$ ” of $Q(0)$, since now $a_1 = 2 \neq 1$. The check $a_0 = 0$ still holds, and the step is still a valid implication; but every value from a_2 on is computed from a_1 , so the failure propagates. (The same induction proves $a_n = 2^{n+1} - 2$ for the new sequence.)

Marking scheme. (10 points) - 1: (a) names two-step recursion ([Corollary 2.3.3](#)), or [Theorem 2.3.1](#) applied to pairs. - 2: (b) states a form of induction under which both a_0 and a_1 are checked: two base cases for strong induction, or the base $Q(0)$ of a pair statement. - 3: (b) the step’s algebra, $3(2^{n+1} - 1) - 2(2^n - 1) = 2^{n+2} - 1$ (or the same with shifted indices). - 1: (b) the induction hypothesis used only for indices where it is available (no a_{-1} ; in strong induction, $n \geq 2$ in the step). - 1: (b) the conclusion, citing the induction principle used. - 2: (c) $a_2 = 6$, $a_3 = 14$ (1), and the failing line identified as the check of a_1 (1). Fatal errors (0 for the step): only a_0 checked, so a_1 is never verified; an induction hypothesis “for all n ”, entry (N9) of Logic and Proof, Example (What is not allowed).

Exercise. P5: core

(20 minutes.) Let $f: X \rightarrow Y$ be a function.

(a) For a family $(B_i)_{i \in I}$ of subsets of Y , prove that

$$f^{-1}\left(Y \setminus \bigcup_{i \in I} B_i\right) = \bigcap_{i \in I} (X \setminus f^{-1}(B_i)),$$

with the intersection taken in X .

(b) Prove that $f(X \setminus A) \subseteq Y \setminus f(A)$ for **every** $A \subseteq X$ if and only if f is injective.

Solution. (a) Put $B := \bigcup_{i \in I} B_i$, a subset of Y since every B_i is. Then

$$\begin{aligned} f^{-1}(Y \setminus B) &= X \setminus f^{-1}(B) = X \setminus \bigcup_{i \in I} f^{-1}(B_i) \\ &= \bigcap_{i \in I} (X \setminus f^{-1}(B_i)). \end{aligned}$$

The first equality is [Theorem 1.6.1](#) (c), the second is [Theorem 1.6.1](#) (a), and the third is [Theorem 1.2.3](#), applied to the family $(f^{-1}(B_i))_{i \in I}$ of subsets of X , with the intersection taken in X . Nothing excludes $I = \emptyset$: then both sides are X .

Alternatively, element by element. For every x : $x \in f^{-1}(Y \setminus B)$ iff $x \in X$ and $f(x) \in Y \setminus B$, iff $x \in X$ and $f(x) \notin B$ (as $f(x) \in Y$ anyway), iff $x \in X$ and there is no $i \in I$ with $f(x) \in B_i$, iff $x \in X$ and for every $i \in I$, $x \notin f^{-1}(B_i)$, iff $x \in X$ and for every $i \in I$, $x \in X \setminus f^{-1}(B_i)$, iff $x \in \bigcap_{i \in I} (X \setminus f^{-1}(B_i))$. The steps use [Definition 1.6.2](#), [Definition 1.2.4](#), [Definition 1.2.1](#) and Logic and Proof, Proposition 6.2.3 (b).

(b) ($\Leftarrow$) Suppose f is injective, and let $A \subseteq X$ and $y \in f(X \setminus A)$. By [Definition 1.6.1](#) there is $x \in X \setminus A$ with $f(x) = y$; in particular $y \in Y$. Suppose, for contradiction, that $y \in f(A)$. Then there is $a \in A$ with $f(a) = y$. So $f(x) = f(a)$, and injectivity gives $x = a \in A$, contradicting $x \notin A$. Hence $y \notin f(A)$, so $y \in Y \setminus f(A)$. As y was arbitrary, $f(X \setminus A) \subseteq Y \setminus f(A)$, and A was arbitrary.

($\Rightarrow$) By contrapositive, as in Logic and Proof, Example (Even square, even number, by contraposition). Suppose f is not injective: by Logic and Proof, Proposition 6.2.3 (a), twice, and Logic and Proof, Proposition 6.2.2 there are $x_1, x_2 \in X$ with $f(x_1) = f(x_2)$ and $x_1 \neq x_2$. Take $A := \{x_1\}$. Then $x_2 \in X \setminus A$, so $f(x_2) \in f(X \setminus A)$. But $f(x_2) = f(x_1) \in f(A)$, so $f(x_2) \notin Y \setminus f(A)$. So the inclusion fails for this A , and it does not hold for every $A \subseteq X$.

The two inclusions hide one witness each: $y \in f(X \setminus A)$ hands us an x , and $y \in f(A)$ hands us an a . The proof works by comparing the two, which is what injectivity is for. Compare [Theorem 1.6.4](#) (b), which characterizes injectivity through intersections in the same way.

Marking scheme. (10 points) - 4: (a) by either route. For the first, each cited result must match its use (which part of [Theorem 1.6.1](#), and De Morgan in X); for the second, each equivalence justified by a definition. A proof of the two inclusions separately earns the same marks. - 3: (b) ($\Leftarrow$), with the witness $x \in X \setminus A$ of $y \in f(X \setminus A)$ and the witness $a \in A$ of $y \in f(A)$ both named, and injectivity applied to them. - 3: (b) ($\Rightarrow$), by contrapositive or contradiction, with a correct set A built from the two points with the same value. Fatal errors (0 for the step): treating $y \in f(A)$ as “ $x \in A$ ” for the x already in hand (a skipped $\exists E$); proving ($\Rightarrow$) from the inclusion for one given A only (the hypothesis “for every A ” dropped; choosing one A where it fails, as above, is correct), or with “for every A ” read as “for some A ”.

Exercise. P6

(20 minutes.) Let S be the set of sequences $(a_n)_{n \in \mathbb{N}}$ of natural numbers, indexed from 0, with $a_n \leq n$ for every $n \in \mathbb{N}$; a sequence is a function $\mathbb{N} \rightarrow \mathbb{N}$, so $S \subseteq \mathbb{N}^{\mathbb{N}}$.

- (a) Prove that S is uncountable.
- (b) A sequence (a_n) is **eventually constant** if there is $N \in \mathbb{N}$ with $a_n = a_N$ for every $n \geq N$. Let $E \subseteq S$ consist of the sequences in S that are eventually constant. Prove that E is countable, and explain why running your construction from (a) on a list of E produces no contradiction.

Solution. (a) The obvious diagonal, $g(n) := f_n(n) + 1$, leaves S : already $g(0) \geq 1 > 0$. Worse, place 0 offers no room at all, since every sequence in S has $a_0 = 0$. So the diagonal is shifted by one place: the n -th sequence of the list is made to differ at place $n + 1$, where there are at least two values to choose from.

Let $f: \mathbb{N} \rightarrow S$ be an arbitrary function, and write $f_n := f(n)$. Define $g: \mathbb{N} \rightarrow \mathbb{N}$ by $g(0) := 0$ and, for $n \in \mathbb{N}$,

$$g(n+1) := \begin{cases} 0 & \text{if } f_n(n+1) \neq 0, \\ 1 & \text{if } f_n(n+1) = 0. \end{cases}$$

This defines g on all of $\mathbb{N}$, since every natural number is 0 or $n + 1$ for exactly one $n \in \mathbb{N}$ (Lemma 2.1.1 (c)). Then $g \in S$: $g(0) = 0 \leq 0$, and $g(n+1) \leq 1 \leq n+1$. For every $n \in \mathbb{N}$, $g(n+1) \neq f_n(n+1)$ by the two cases, so $g \neq f_n$. Hence g is not a value of f , and f is not surjective. Since f was arbitrary, there is no surjection $\mathbb{N} \rightarrow S$. And $S \neq \emptyset$, since the zero sequence lies in S . By Definition 3.2.1, S is not countable.

(b) For $N \in \mathbb{N}$ let E_N be the set of sequences $(a_n) \in S$ with $a_n = a_N$ for every $n \geq N$. By the definition of eventually constant, $E = \bigcup_{N \in \mathbb{N}} E_N$.

Each E_N is countable. For $N = 0$, a sequence in E_0 is constant with value $a_0 \leq 0$, so E_0 has one element, the zero sequence; a constant map $\mathbb{N} \rightarrow E_0$ is onto, and E_0 is countable. Let $N \geq 1$, and define $\pi: E_N \rightarrow \mathbb{N}^{N+1}$ by $\pi(a) := (a_0, a_1, \dots, a_N)$, an $(N+1)$ -tuple. It is injective: if $\pi(a) = \pi(b)$, then $a_k = b_k$ for $k \leq N$, by the characteristic property of tuples (Section 1.3), and for $n \geq N$, $a_n = a_N = b_N = b_n$; so $a = b$ (Proposition 1.4.1). The set $\mathbb{N}^{N+1}$ is countable by Corollary 3.2.7, as $N+1 \geq 2$, so by Theorem 3.2.2, (i) $\Rightarrow$ (ii), there is an injection $j: \mathbb{N}^{N+1} \rightarrow \mathbb{N}$. Then $j \circ \pi: E_N \rightarrow \mathbb{N}$ is injective (Theorem 1.4.4), and E_N is countable by (ii) $\Rightarrow$ (i).

So E is countable, as a union of countably many countable sets, by Theorem 3.3.3. That theorem uses the axiom of choice once, to pick a list of each E_N . Here the choice can be avoided (compare Example (Unions that come with their lists); see the marking scheme). The solution does not need this; citing Theorem 3.3.3 is a complete proof.

In fact each E_N is finite: $a_0, a_1, \dots, a_N$ can take $1, 2, \dots, N+1$ values, so E_N has $(N+1)!$ elements (Definition 2.3.1), which is $1, 2, 6, 24, 120$ for $N = 0, \dots, 4$. This count rests on Exercise (C2: Counting products and functions) and an induction on the number of factors, and it earns the same marks.

Why there is no contradiction. The set E is countable and non-empty, so it has a list: a surjection $e: \mathbb{N} \rightarrow E$. The construction of (a) applies to e , since e is in particular a function $\mathbb{N} \rightarrow S$, and produces $g \in S$ with $g \neq e(n)$ for every n . Since every element of E is some $e(n)$, this shows $g \notin E$: the diagonal sequence is not eventually constant. That is all it shows. A diagonal argument proves that a given list misses something in the set the diagonal element lives in, here S ; it contradicts the countability of E only if the diagonal element lies in E , and it does not.

Marking scheme. (10 points) - 1: (a) starts from an arbitrary list, a function $\mathbb{N} \rightarrow S$, not

a particular one. - 2: (a) defines a diagonal element that lies in S , with the problem at place 0 handled. - 2: (a) shows that it differs from every f_n , at a named place. - 1: (a) concludes, with the form “countable means empty or the image of $\mathbb{N}$ ” handled ($S \neq \emptyset$). - 2: (b) writes E as a countable union of countable sets, with a reason why each piece is countable (an injection into $\mathbb{N}$, a surjection from $\mathbb{N}$, or finiteness via [Exercise \(C2: Counting products and functions\)](#)). Alternative route, same 2 points: an injection of E into the finite sequences of [Exercise \(B2: Finite sequences\)](#), $a \mapsto (k \mapsto a_{k-1})$ on $\{1, \dots, N+1\}$ with N least such that $a \in E_N$, with injectivity checked and [Theorem 3.2.2](#) cited. - 2: (b) explains that the diagonal element of a list of E lies in S but not in E , so it contradicts nothing. Fatal errors (0 for the step): a diagonal element not in S , such as $g(n) := f_n(n) + 1$, which already has $g(0) \geq 1 > 0$; proving (a) for one particular list.

3.7.4 Remediation map

Add up your points. A pass is 42 of 60 with no core problem (P1, P2, P4, P5) at 0. For each problem where you lost more than 3 points, reread the sections in its row, redo their A exercises and one C exercise. Then retake that problem if you passed, or the whole Gate after a week if you failed. A wrong answer in the definition drill sends you to the label printed after its answer.

Problem	What it tests	Sections to reread
P1	quantifiers, bounds and their negation; the order of choices	<i>Logic and Proof</i> , Section 4.1 (Logic and Proof, Definition 5.1.1); <i>Logic and Proof</i> , Section 4.9 (Logic and Proof, Proposition 6.2.1, Logic and Proof, Proposition 6.2.2, Logic and Proof, Proposition 6.2.3); <i>Logic and Proof</i> , Section 4.10 (Logic and Proof, Example (Uniformly continuous implies continuous)); <i>Logic and Proof</i> , Section 5.2 (Logic and Proof, Example (A constant sequence))
P2	the quantifier rules and their side conditions; a derivation read as prose	<i>Logic and Proof</i> , Sections 4.3–4.6 (Logic and Proof, Definition 5.3.1, Logic and Proof, Definition 5.4.1, Logic and Proof, Definition 5.5.1, Logic and Proof, Definition 5.6.1); <i>Logic and Proof</i> , Sections 5.2–5.3 (Logic and Proof, Example (The reciprocals tend to zero), Logic and Proof, Definition 8.3.1); Section 1.4 (Theorem 1.4.4)

Problem	What it tests	Sections to reread
P3	checking a proof; is the conclusion true, and does the argument prove it	<i>Logic and Proof</i> , Section 5.6 (Logic and Proof, Example (What is not allowed)); <i>Logic and Proof</i> , Section 3.2 (Logic and Proof, Axiom 4.2.1); <i>Logic and Proof</i> , Section 4.6 (Logic and Proof, Proposition 5.6.2); <i>Logic and Proof</i> , Section 4.11 (Logic and Proof, Example (The swap of $\exists$ and $\forall$)); Section 2.2 (Theorem 2.2.1 , and “How induction proofs go wrong”, Example (All Numbers Are Equal))
P4	recursion and induction with two starting values	<i>Logic and Proof</i> , Section 3.2 (Logic and Proof, Theorem 4.2.2); Section 2.2 (Theorem 2.2.2); Section 2.3 (Corollary 2.3.3 , and the two-step example after Definition 2.3.2)
P5	images and preimages, element by element	Section 1.2 (Theorem 1.2.3); Section 1.6 (Theorem 1.6.1 , Theorem 1.6.4)
P6	countable sets and the diagonal argument	Section 3.2 (Definition 3.2.1); Section 3.3 (Theorem 3.3.3); Section 3.4 (Theorem 3.4.1 , Theorem 3.4.2)