

24.2 Elimination, Pivoting, and Cost

Chapter 3 solved $\mathbf{Ax} = \mathbf{b}$ by elimination and packaged the work as $\mathbf{A} = \mathbf{LU}$ or $\mathbf{PA} = \mathbf{LU}$. It said nothing about how much arithmetic that costs, and almost nothing about what rounding does to it — only a remark, about a two-by-two matrix with a tiny corner entry, ending: “Chapter 24 makes this precise.” Chapter 8 left a second note: “Chapter 24 reads Gaussian elimination itself as a sequence of Schur complements of 1×1 pivots.” Chapter 13 left a third, that Cholesky “does about half the arithmetic of a general LU factorization” and needs no pivoting.

This section pays all three. Section 1 supplied the model and the word *backward stable*; here that word is applied to the oldest algorithm in the book. Throughout, matrices are **real**, since the representable numbers of Definition 24.1.1 are real, and $|\mathbf{M}|$ denotes the entrywise absolute value ($|m_{ij}|$), with $|\mathbf{M}| \leq |\mathbf{N}|$ read entrywise.

24.2.1 Elimination is a sequence of Schur complements

Chapter 8 built a machine for one block elimination: if $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ with $\mathbf{A}$ invertible, then subtracting $\mathbf{CA}^{-1}$ times the first block row from the second leaves $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{M}/\mathbf{A} \end{pmatrix}$, where $\mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}$ is the Schur complement (Definition 8.3.1, Theorem 8.3.1). Gaussian elimination is that machine run with a 1×1 block, $n - 1$ times. Saying so exactly is the cleanest description of the algorithm there is, and it is what makes block versions of it possible.

Theorem 24.2.1. Elimination Produces Schur Complements

Let $\mathbf{A} \in M_n(F)$ and $1 \leq k \leq n - 1$, and suppose the first k steps of elimination without row swaps (Chapter 3 §06) are carried out with **nonzero pivots**, producing $\mathbf{A}^{(k)}$. Partition

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_k & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}, \quad \mathbf{A}^{(k)} = \begin{pmatrix} \mathbf{V} & \mathbf{B}' \\ \mathbf{O} & \mathbf{S} \end{pmatrix},$$

with $\mathbf{A}_k, \mathbf{V} \in M_k(F)$, the zero block being what the first k steps created. Then $\mathbf{A}_k$ is invertible and

$$\mathbf{S} = \mathbf{A}/\mathbf{A}_k = \mathbf{D} - \mathbf{CA}_k^{-1}\mathbf{B}.$$

In particular, for $k = 1$ and $a_{11} \neq 0$, one step of elimination replaces the trailing $(n - 1) \times (n - 1)$ block by the Schur complement of the 1×1 pivot block.

Idea. The first k steps multiply $\mathbf{A}$ on the left by a unit lower triangular matrix whose columns $k + 1, \dots, n$ are those of the identity — that is exactly what Lemma 3.6.3 (c) says, and it is the whole point. Such a matrix has block form $\begin{pmatrix} \mathbf{L}_{11} & \mathbf{O} \\ \mathbf{L}_{21} & \mathbf{I} \end{pmatrix}$. Multiply that by the partitioned $\mathbf{A}^{(k)}$, match the four blocks against $\mathbf{A}$, and the $\mathbf{L}_{11}$ cancels out of the bottom-right block, leaving the Schur complement.

Proof. Write ℓ_j for the j -th multiplier vector, whose first j entries are zero, and recall from Chapter 3 §06 that step j multiplies on the left by $\mathbf{G}_j(-\ell_j)$, with inverse $\mathbf{G}_j(\ell_j)$ by

Lemma 3.6.3 (b). Hence $\mathbf{A} = \mathbf{L}_k \mathbf{A}^{(k)}$ where, by Lemma 3.6.3 (c),

$$\mathbf{L}_k = \mathbf{G}_1(\ell_1) \cdots \mathbf{G}_k(\ell_k) = \mathbf{I} + \sum_{j=1}^k \ell_j \mathbf{e}_j^\top.$$

Only columns $1, \dots, k$ of $\mathbf{L}_k$ differ from those of $\mathbf{I}$, so partitioning $\mathbf{L}_k$ as $\mathbf{A}$ is partitioned gives

$$\mathbf{L}_k = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{O} \\ \mathbf{L}_{21} & \mathbf{I}_{n-k} \end{pmatrix},$$

with $\mathbf{L}_{11} \in M_k(F)$ unit lower triangular, hence invertible by Lemma 3.6.1 (b).

By Theorem 8.1.1,

$$\begin{pmatrix} \mathbf{A}_k & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{O} \\ \mathbf{L}_{21} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{V} & \mathbf{B}' \\ \mathbf{O} & \mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_{11}\mathbf{V} & \mathbf{L}_{11}\mathbf{B}' \\ \mathbf{L}_{21}\mathbf{V} & \mathbf{L}_{21}\mathbf{B}' + \mathbf{S} \end{pmatrix}.$$

Comparing blocks: $\mathbf{A}_k = \mathbf{L}_{11}\mathbf{V}$, $\mathbf{B} = \mathbf{L}_{11}\mathbf{B}'$, $\mathbf{C} = \mathbf{L}_{21}\mathbf{V}$, and $\mathbf{D} = \mathbf{L}_{21}\mathbf{B}' + \mathbf{S}$.

Now $\mathbf{V}$ is upper triangular with the first k pivots on its diagonal, and those are nonzero by hypothesis, so $\mathbf{V}$ is invertible by Lemma 3.6.2. Therefore $\mathbf{A}_k = \mathbf{L}_{11}\mathbf{V}$ is invertible, and $\mathbf{A}_k^{-1} = \mathbf{V}^{-1}\mathbf{L}_{11}^{-1}$. Hence

$$\mathbf{C} \mathbf{A}_k^{-1} \mathbf{B} = (\mathbf{L}_{21}\mathbf{V})(\mathbf{V}^{-1}\mathbf{L}_{11}^{-1})(\mathbf{L}_{11}\mathbf{B}') = \mathbf{L}_{21}\mathbf{B}',$$

and so $\mathbf{S} = \mathbf{D} - \mathbf{L}_{21}\mathbf{B}' = \mathbf{D} - \mathbf{C} \mathbf{A}_k^{-1} \mathbf{B} = \mathbf{A}/\mathbf{A}_k$. This proves the theorem. ■

Example. Two steps, two Schur complements

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 7 & 5 \\ -1 & 4 & 8 \end{pmatrix}.$$

Carry out two steps of elimination without row swaps, then compute $\mathbf{A}/\mathbf{A}_1$ and $\mathbf{A}/\mathbf{A}_2$ directly from Definition 8.3.1, and check [Theorem 24.2.1](#) in both cases.

Solution. The elimination. The first multipliers are $\ell_{21} = 2/1 = 2$ and $\ell_{31} = -1/1 = -1$; subtracting ℓ_{i1} times row 1 from row i gives $\mathbf{A}^{(1)}$. The second multiplier is $\ell_{32} = 6/3 = 2$, and row 3 becomes $(0, 0, 11 - 2(-1)) = (0, 0, 13)$:

$$\mathbf{A}^{(1)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 3 & -1 \\ 0 & 6 & 11 \end{pmatrix}, \quad \mathbf{A}^{(2)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 3 & -1 \\ 0 & 0 & 13 \end{pmatrix}.$$

The Schur complement of the 1×1 block. Here $\mathbf{A}_1 = (1)$, $\mathbf{B} = \begin{pmatrix} 2 & 3 \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} 2 & -1 \end{pmatrix}^\top$, and

$\mathbf{D} = \begin{pmatrix} 7 & 5 \\ 4 & 8 \end{pmatrix}$, so

$$\begin{aligned} \mathbf{A}/\mathbf{A}_1 &= \mathbf{D} - \mathbf{C}\mathbf{A}_1^{-1}\mathbf{B} = \begin{pmatrix} 7 & 5 \\ 4 & 8 \end{pmatrix} - \begin{pmatrix} 2 \\ -1 \end{pmatrix} \begin{pmatrix} 2 & 3 \end{pmatrix} \\ &= \begin{pmatrix} 7 & 5 \\ 4 & 8 \end{pmatrix} - \begin{pmatrix} 4 & 6 \\ -2 & -3 \end{pmatrix} = \begin{pmatrix} 3 & -1 \\ 6 & 11 \end{pmatrix}, \end{aligned}$$

which is exactly the trailing block of $\mathbf{A}^{(1)}$.

The Schur complement of the 2×2 block. Now $\mathbf{A}_2 = \begin{pmatrix} 1 & 2 \\ 2 & 7 \end{pmatrix}$, with $\det \mathbf{A}_2 = 3$ and $\mathbf{A}_2^{-1} = \frac{1}{3} \begin{pmatrix} 7 & -2 \\ -2 & 1 \end{pmatrix}$, while $\mathbf{C} = \begin{pmatrix} -1 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 3 & 5 \end{pmatrix}^\top$. Then

$$\begin{aligned} \mathbf{C}\mathbf{A}_2^{-1} &= \frac{1}{3} \begin{pmatrix} -15 & 6 \end{pmatrix} = \begin{pmatrix} -5 & 2 \end{pmatrix}, \\ \mathbf{A}/\mathbf{A}_2 &= 8 - \begin{pmatrix} -5 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 5 \end{pmatrix} = 8 - (-5) = 13, \end{aligned}$$

the trailing 1×1 block of $\mathbf{A}^{(2)}$. Both readings agree with [Theorem 24.2.1](#), and as an arithmetic check the pivots 1, 3, 13 multiply to 39, which is $\det \mathbf{A}$ — the corollary below says why that is no coincidence.

Two consequences come out for free. The first identifies the pivots with ratios of minors, which is how Chapter 13 §01 wrote them in Sylvester's criterion.

Corollary 24.2.2. Pivots Are Ratios of Leading Minors

Let $\mathbf{A} \in M_n(F)$, let $1 \leq k \leq n-1$, and write $\mathbf{A}_j$ for the leading principal $j \times j$ submatrix of $\mathbf{A}$, with $\det \mathbf{A}_0 := 1$.

- (a) If the first k pivots of elimination without row swaps are nonzero, then $\mathbf{A}_1, \dots, \mathbf{A}_k$ are invertible and the j -th pivot equals $\det \mathbf{A}_j / \det \mathbf{A}_{j-1}$ for $j \leq k$.
- (b) Conversely, if $\mathbf{A}_1, \dots, \mathbf{A}_{n-1}$ are invertible then elimination without row swaps succeeds on $\mathbf{A}$ with nonzero pivots, and $\mathbf{A} = \mathbf{L}\mathbf{U}$ by Theorem 3.6.4.

Proof.

- (a) Invertibility of $\mathbf{A}_1, \dots, \mathbf{A}_k$ is [Theorem 24.2.1](#) applied with $1, \dots, k$ in place of k . For $j = 1$ the pivot is $a_{11} = \det \mathbf{A}_1 / \det \mathbf{A}_0$, as claimed. For $2 \leq j \leq k$, the j -th pivot is the (j, j) entry of $\mathbf{A}^{(j-1)}$, which by the theorem (with $j-1$ in place of k) is the $(1, 1)$ entry of $\mathbf{A}/\mathbf{A}_{j-1}$. That entry is a_{jj} minus row j of $\mathbf{C}$ times $\mathbf{A}_{j-1}^{-1}$ times column j of $\mathbf{B}$, so it involves only rows and columns $1, \dots, j$ of $\mathbf{A}$: it equals the 1×1 matrix $\mathbf{A}_j/\mathbf{A}_{j-1}$. By Theorem 8.3.2 (a), $\det \mathbf{A}_j = \det \mathbf{A}_{j-1} \cdot \det(\mathbf{A}_j/\mathbf{A}_{j-1})$, which is the claim.
- (b) We show by induction on j that the j -th pivot is nonzero, for $j = 1, \dots, n-1$. For $j = 1$ the pivot is $a_{11} = \det \mathbf{A}_1 \neq 0$. Let $2 \leq j \leq n-1$ and suppose the first $j-1$ pivots are nonzero. Then [Theorem 24.2.1](#) applies with $j-1$ in place of k , and the calculation in (a) identifies the j -th pivot as $\det \mathbf{A}_j / \det \mathbf{A}_{j-1}$, which is nonzero because $\mathbf{A}_j$ and $\mathbf{A}_{j-1}$ are invertible (Theorem 7.5.2). So elimination never stops, and Theorem 3.6.4 gives $\mathbf{A} = \mathbf{L}\mathbf{U}$.

■

Quick check. The matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ has no LU factorization (Example (An invertible matrix with no LU factorization)). Which hypothesis of [Corollary 24.2.2](#) (b) fails, and what does [Theorem 24.2.1](#) say about it?

Solution. $\mathbf{A}_1 = \mathbf{O}_1$ is not invertible, so (b) does not apply. [Theorem 24.2.1](#) does not apply either: it assumes the first pivot is nonzero, and here it is 0, so the Schur complement $\mathbf{A}/\mathbf{A}_1$ is not even defined — Definition 8.3.1 requires the pivot block to be invertible. This is the whole content of Chapter 3’s “elimination stops”: there is no Schur complement to form.

24.2.2 What elimination costs

A **flop** is one arithmetic operation: one addition, subtraction, multiplication, or division, or one square root. Counting flops is crude — it ignores which operations are slow and how data moves — but it captures how the work grows with n . We quote leading terms, writing $\frac{2}{3}n^3 + O(n^2)$ and meaning it. **The constant matters and the exponent matters more:** halving the constant is worth having, but a smaller exponent overtakes it at some size and stays ahead forever.

Proposition 24.2.3. The Cost of Elimination

Counting each of the five operations as one flop, and counting only arithmetic on matrix and vector entries:

- (a) Elimination without row swaps on $\mathbf{A} \in M_n(\mathbb{R})$ costs

$$\frac{n(n-1)(4n+1)}{6} = \frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n$$

flops, that is $\frac{2}{3}n^3 + O(n^2)$.

- (b) Forward substitution with a unit lower triangular $\mathbf{L}$ costs $n^2 - n$ flops and back substitution with an upper triangular $\mathbf{U}$ costs n^2 , so solving $\mathbf{Ax} = \mathbf{b}$ from a factorization costs $2n^2 - n$ flops.
- (c) The Cholesky recursion of Chapter 13 §02 on a positive definite $\mathbf{A} \in M_n(\mathbb{R})$ costs

$$\frac{n(n+1)(2n+1)}{6} = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n$$

flops, that is $\frac{1}{3}n^3 + O(n^2)$, of which n are square roots.

Proof.

- (a) At step k the algorithm computes $n-k$ multipliers, one division each, and then replaces each of the $(n-k)^2$ entries of the trailing block by $a_{ij} - \ell_{ik}a_{kj}$, two flops each. The entries below the pivot are set to zero without arithmetic. So the cost is

$$\sum_{k=1}^{n-1} [(n-k) + 2(n-k)^2] = \sum_{j=1}^{n-1} (j + 2j^2) = \frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{3},$$

using $\sum_{j \leq m} j = m(m+1)/2$ and $\sum_{j \leq m} j^2 = m(m+1)(2m+1)/6$ with $m = n-1$.

Putting the two terms over 6 gives $n(n-1)(3+2(2n-1))/6 = n(n-1)(4n+1)/6$.

- (b) Forward substitution computes $x_i = b_i - \sum_{j<i} \ell_{ij}x_j$ with no division, since $\ell_{ii} = 1$: that is $2(i-1)$ flops, and $\sum_{i \leq n} 2(i-1) = n(n-1)$. Back substitution computes $x_i = (b_i - \sum_{j>i} u_{ij}x_j)/u_{ii}$, which is $2(n-i) + 1$ flops, and $\sum_{i \leq n} (2(n-i) + 1) = n(n-1) + n = n^2$.
- (c) Column j of the recursion computes $\ell_{jj} = (a_{jj} - \sum_{k<j} \ell_{jk}^2)^{1/2}$, costing $2(j-1)$ flops for the sum and one square root, and then $\ell_{ij} = (a_{ij} - \sum_{k<j} \ell_{ik}\ell_{jk})/\ell_{jj}$ for each of the $n-j$ rows below, costing $2(j-1) + 1$ flops each. The total is

$$\begin{aligned} \sum_{j=1}^n (n-j+1)(2j-1) &= \sum_{i=1}^n i(2n-2i+1) \\ &= (2n+1) \frac{n(n+1)}{2} - 2 \cdot \frac{n(n+1)(2n+1)}{6}, \end{aligned}$$

substituting $i = n-j+1$; the two terms combine to $n(n+1)(2n+1)(\frac{1}{2} - \frac{1}{3}) = n(n+1)(2n+1)/6$. ■

Comparing (a) and (c): $\frac{1}{3}n^3$ against $\frac{2}{3}n^3$. This is Chapter 13 §02's "Cholesky does about half the arithmetic of a general LU factorization", now a count rather than a claim, for the reason that section gave — the recursion produces one triangular factor instead of two. Part (b) says the factorization is the expensive part: solving m systems with the same $\mathbf{A}$ costs $\frac{2}{3}n^3 + 2mn^2$, not m times $\frac{2}{3}n^3$.

24.2.3 One row at a time

Every algorithm below computes its entries the same way: accumulate a sum of products into a running total, then possibly divide. Analyzing that pattern once saves doing it three times.

Lemma 24.2.4. One Row of Substitution

Let $m \geq 0$, let $t_1, \dots, t_m, y_1, \dots, y_m, c$, and $d \neq 0$ be representable, and suppose

$$s_0 = c, \quad s_p = \text{fl}(s_{p-1} - \text{fl}(t_p y_p)) \quad (p = 1, \dots, m), \quad z = \text{fl}(s_m/d)$$

are computed in the model of Definition 24.1.1, with $(m+1)u < 1$. Then there are $\theta_0, \dots, \theta_m$ with

$$c = dz(1 + \theta_0) + \sum_{p=1}^m t_p y_p (1 + \theta_p), \quad |\theta_p| \leq \gamma_{m+1} \text{ for all } p.$$

If the final division is omitted, so that $z = s_m$, the same identity holds with $d = 1$ and $|\theta_p| \leq \gamma_m$.

Idea. Unroll the recursion, then divide the whole identity by the product of the accumulation factors. Dividing is the trick: it moves those factors onto c 's neighbors as **inverse** factors, which is exactly what Lemma 24.1.1 was stated to allow.

Proof. By (FP2) there are $\delta_p, \varepsilon_p, \eta$, all of modulus at most u , with $\text{fl}(t_p y_p) = t_p y_p(1 + \delta_p)$, $s_p = (s_{p-1} - t_p y_p(1 + \delta_p))(1 + \varepsilon_p)$, and $z = (s_m/d)(1 + \eta)$. Induction on m gives

$$s_m = c \prod_{p=1}^m (1 + \varepsilon_p) - \sum_{p=1}^m t_p y_p (1 + \delta_p) \prod_{q=p}^m (1 + \varepsilon_q),$$

the case $m = 0$ being $s_0 = c$, and the step from $m - 1$ to m multiplying everything by $1 + \varepsilon_m$ and subtracting the new term.

Each $1 + \varepsilon_p$ is nonzero, since $|\varepsilon_p| \leq u < 1$, so we may divide the identity by $\prod_{p \leq m} (1 + \varepsilon_p)$ and rearrange:

$$c = s_m \prod_{p=1}^m (1 + \varepsilon_p)^{-1} + \sum_{p=1}^m t_p y_p (1 + \delta_p) \prod_{q=1}^{p-1} (1 + \varepsilon_q)^{-1}.$$

From $z = (s_m/d)(1 + \eta)$ we get $s_m = d z (1 + \eta)^{-1}$. So the first term is $d z$ times a product of $m + 1$ factors of the form $(1 + \delta)^{\pm 1}$ with $|\delta| \leq u$, which by Lemma 24.1.1 is $1 + \theta_0$ with $|\theta_0| \leq \gamma_{m+1}$. The p -th term of the sum carries $1 + (p - 1) = p \leq m$ such factors, so it is $t_p y_p (1 + \theta_p)$ with $|\theta_p| \leq \gamma_p \leq \gamma_{m+1}$.

If the division is omitted, the first term carries only the m inverse factors, so $|\theta_0| \leq \gamma_m$, and the rest of the argument is unchanged. This proves the lemma. ■

24.2.4 Solving a triangular system

Here is the cleanest backward error analysis in the subject, and the model for every other one in this chapter. Note in advance what it does **not** assume: nothing about the size of the diagonal entries of $\mathbf{T}$, nothing about its conditioning, and no pivoting of any kind.

Theorem 24.2.5. Backward Error of a Triangular Solve

Let $\mathbf{T} \in M_n(\mathbb{R})$ be lower triangular with representable entries and nonzero diagonal, let $\mathbf{b} \in \mathbb{R}^n$ be representable, and let $nu < 1$. Let $\hat{\mathbf{x}}$ be computed by forward substitution in the model: for each i , the products $t_{ij}\hat{x}_j$ are subtracted from b_i one at a time in increasing order of j , and the result is divided by t_{ii} — that is, exactly the computation of Lemma 24.2.4 with $m = i - 1$. Then there is a lower triangular $\Delta\mathbf{T}$ with

$$(\mathbf{T} + \Delta\mathbf{T})\hat{\mathbf{x}} = \mathbf{b} \quad \text{and} \quad |\Delta t_{ij}| \leq \gamma_i |t_{ij}| \leq \gamma_n |t_{ij}| \quad (j \leq i).$$

The same statement, with γ_{n+1-i} in place of γ_i , holds for back substitution with an upper triangular $\mathbf{T}$.

Idea. Row i is the pattern of Lemma 24.2.4, with $c = b_i$, the $m = i - 1$ products $t_{ij}\hat{x}_j$, and the final division by t_{ii} . The lemma hands back an identity in which every coefficient of $\hat{x}_j$ is t_{ij} times something within γ_i of 1; that “something minus one”, times t_{ij} , is Δt_{ij} .

Proof. Fix i . The quantities $\hat{x}_1, \dots, \hat{x}_{i-1}$ are already computed, hence representable, so Lemma 24.2.4 applies with $m = i - 1$, $c = b_i$, $t_p = t_{ip}$, $y_p = \hat{x}_p$, $d = t_{ii}$, and $z = \hat{x}_i$. Since $m + 1 = i \leq n$ and $nu < 1$, it gives $\theta_0, \dots, \theta_{i-1}$ of modulus at most γ_i with

$$b_i = t_{ii}\hat{x}_i(1 + \theta_0) + \sum_{j < i} t_{ij}\hat{x}_j(1 + \theta_j).$$

Define $\Delta t_{ii} = t_{ii}\theta_0$, $\Delta t_{ij} = t_{ij}\theta_j$ for $j < i$, and $\Delta t_{ij} = 0$ for $j > i$. Then $\Delta \mathbf{T}$ is lower triangular, $|\Delta t_{ij}| \leq \gamma_i |t_{ij}| \leq \gamma_n |t_{ij}|$ since $\gamma_i \leq \gamma_n$, and the display says precisely that row i of $(\mathbf{T} + \Delta \mathbf{T})\hat{\mathbf{x}}$ equals b_i . As i was arbitrary, $(\mathbf{T} + \Delta \mathbf{T})\hat{\mathbf{x}} = \mathbf{b}$.

For back substitution, row i has $n - i$ products instead of $i - 1$, so the lemma applies with $m = n - i$ and the bound is γ_{n+1-i} . This proves the theorem. ■

This is as strong as a backward error statement gets. The perturbation is bounded **entrywise**, and entrywise *relative to the entries of $\mathbf{T}$ itself*: a zero entry of $\mathbf{T}$ is not perturbed at all, and a tiny entry is perturbed only in its own last digits. So a triangular solve is backward stable in the sense of Definition 24.1.3 — with the constant $c = 1.02n$ in any model with $nu \leq 0.01$, since $\gamma_n \leq 1.02nu$ there — whatever $\kappa(\mathbf{T})$ may be.

Warning. Backward stable is not accurate, and this theorem is where the point bites hardest. Theorem 24.2.5 holds for every nonsingular triangular $\mathbf{T}$. It does not say $\hat{\mathbf{x}}$ is near $\mathbf{x}$. By Theorem 24.1.3 the forward error can be about $\kappa(\mathbf{T})\gamma_n$, and triangular matrices can be spectacularly ill-conditioned. Take the $n \times n$ matrix $\mathbf{T}$ with 1 on the diagonal and -1 everywhere above it. Its row sums of moduli are $n, n-1, \dots, 1$, so $\|\mathbf{T}\|_\infty = n$. Its inverse has 1 on the diagonal and 2^{j-i-1} in position (i, j) for $j > i$: back substitution on $\mathbf{T}\mathbf{x} = \mathbf{e}_j$ gives $x_j = 1$ and $x_i = \sum_{p>i} x_p$ for $i < j$, and each partial sum of $1, 1, 2, 4, \dots$ is the next term. Hence $\|\mathbf{T}^{-1}\|_\infty = 1 + (1 + 2 + \dots + 2^{n-2}) = 2^{n-1}$ and $\kappa_\infty(\mathbf{T}) = n2^{n-1}$, so for $n = 60$ a backward stable solve can still return nonsense. Stability is a promise about the *data*, and a promise about the data is worth only what the conditioning of the problem makes it worth.

24.2.5 The factorization itself

For the factorization we analyze the ordering that computes each entry of $\mathbf{L}$ and $\mathbf{U}$ by one accumulation. In exact arithmetic it produces the same factors as Chapter 3's row operations — for invertible $\mathbf{A}$ this is Theorem 3.6.5 — but it groups the arithmetic differently, and rounding notices the difference.

Algorithm. Elimination, one entry at a time. Input: $\mathbf{A} \in M_n(\mathbb{R})$. Set $\ell_{ii} = 1$. For $i = 1, 2, \dots, n$:

- for $j = i, \dots, n$, put $u_{ij} = a_{ij} - \sum_{k < i} \ell_{ik}u_{kj}$;
- if $u_{ii} = 0$, stop; otherwise for $r = i + 1, \dots, n$, put $\ell_{ri} = (a_{ri} - \sum_{k < i} \ell_{rk}u_{ki})/u_{ii}$.

Theorem 24.2.6. Backward Error of a Factorization

Let $\mathbf{A} \in M_n(\mathbb{R})$ have representable entries, let $nu < 1$, and suppose the algorithm above runs to completion in the model of Definition 24.1.1, producing $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{U}}$. Then

$$\widehat{\mathbf{L}}\widehat{\mathbf{U}} = \mathbf{A} + \Delta\mathbf{A} \quad \text{with} \quad |\Delta\mathbf{A}| \leq \gamma_n |\widehat{\mathbf{L}}| |\widehat{\mathbf{U}}|$$

entrywise.

Idea. Each entry is one application of [Lemma 24.2.4](#): the entries of $\mathbf{U}$ with no division, the entries of $\mathbf{L}$ with one. In both cases the lemma returns a_{ij} as a sum of the products $\widehat{\ell}_{ip}\widehat{u}_{pj}$, each with its own factor $1 + \theta_p$. Collecting the terms is exactly forming the (i, j) entry of $\widehat{\mathbf{L}}\widehat{\mathbf{U}}$, and the θ_p are what is left over.

Proof. Case $i \leq j$. The computation of $\widehat{u}_{ij}$ is [Lemma 24.2.4](#) with $m = i - 1$, $c = a_{ij}$, $t_p = \widehat{\ell}_{ip}$, $y_p = \widehat{u}_{pj}$, no division, and $z = \widehat{u}_{ij}$. Writing θ_i for the θ_0 of the lemma, and using $\widehat{\ell}_{ii} = 1$,

$$a_{ij} = \sum_{p=1}^i \widehat{\ell}_{ip}\widehat{u}_{pj}(1 + \theta_p), \quad |\theta_p| \leq \gamma_{i-1}.$$

(For $i = 1$ the sum has one term and $\theta_1 = 0$.) Since $\widehat{\ell}_{ip} = 0$ for $p > i$, the sum without the θ_p is $(\widehat{\mathbf{L}}\widehat{\mathbf{U}})_{ij}$. Hence

$$|a_{ij} - (\widehat{\mathbf{L}}\widehat{\mathbf{U}})_{ij}| = \left| \sum_{p \leq i} \widehat{\ell}_{ip}\widehat{u}_{pj}\theta_p \right| \leq \gamma_{i-1} \sum_{p \leq i} |\widehat{\ell}_{ip}||\widehat{u}_{pj}| = \gamma_{i-1} (|\widehat{\mathbf{L}}||\widehat{\mathbf{U}}|)_{ij},$$

the last equality again because $\widehat{\ell}_{ip} = 0$ for $p > i$.

Case $i > j$. The computation of $\widehat{\ell}_{ij}$ is [Lemma 24.2.4](#) with $m = j - 1$, $c = a_{ij}$, $t_p = \widehat{\ell}_{ip}$, $y_p = \widehat{u}_{pj}$, $d = \widehat{u}_{jj}$, and $z = \widehat{\ell}_{ij}$. Writing θ_j for θ_0 ,

$$a_{ij} = \sum_{p=1}^j \widehat{\ell}_{ip}\widehat{u}_{pj}(1 + \theta_p), \quad |\theta_p| \leq \gamma_j.$$

Since $\widehat{u}_{pj} = 0$ for $p > j$, the sum without the θ_p is again $(\widehat{\mathbf{L}}\widehat{\mathbf{U}})_{ij}$, and the same estimate gives $|a_{ij} - (\widehat{\mathbf{L}}\widehat{\mathbf{U}})_{ij}| \leq \gamma_j (|\widehat{\mathbf{L}}||\widehat{\mathbf{U}}|)_{ij}$.

In both cases the constant is at most γ_n . This proves the theorem. ■

Remark. What is proved and what is not. The theorem is about the ordering displayed above. Chapter 3's ordering, which updates the whole trailing block at each step, satisfies the same bound $|\Delta\mathbf{A}| \leq \gamma_n |\widehat{\mathbf{L}}| |\widehat{\mathbf{U}}|$; the proof is a different induction of comparable length, and **we do not give it here**. Nothing later in this chapter uses it. Also, the theorem assumes the algorithm completes, which in the model means every computed $\widehat{u}_{ii}$ is nonzero; it says nothing about what happens when one of them is zero, or about the case where a pivot is nonzero only by a rounding accident. One more seam will appear below: the growth bound for partial pivoting, proved in the next subsection, concerns the **exact** elimination

in Chapter 3's ordering, whereas the bound just proved concerns the factors this ordering computes in the model. In exact arithmetic, and for invertible $\mathbf{A}$, the two orderings produce the same $\mathbf{L}$ and $\mathbf{U}$, by Theorem 3.6.5; in the model they agree only to within rounding, and we do not quantify that here. Nothing below is allowed to cross that seam quietly. Eq. 24.2.1, proved next, measures growth in the **computed** factors, which is a quantity one can read off after the fact; the a priori bound 2^{n-1} of Theorem 24.2.8 is a statement about exact elimination. The three places that put the two together — the warning after Example (A matrix that really grows) and the solutions to A1 and B3 — say that they are doing so, and each of them uses the combination only to show that even its most favorable reading proves nothing.

The bound is useless until we know how big $|\widehat{\mathbf{L}}||\widehat{\mathbf{U}}|$ is, and *that* is not controlled by $\mathbf{A}$. It is controlled by what the algorithm does to $\mathbf{A}$, which is why elimination needs a pivoting rule at all. The quantity to watch has a name.

Definition 24.2.1. Growth factor

Let elimination — with whatever pivoting rule is in force — be run on $\mathbf{A} \in M_n(\mathbb{R})$, producing an upper factor $\mathbf{U}$. The **growth factor** of that run is

$$g_n := \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|},$$

the largest entry in modulus the algorithm produces, relative to the largest entry in modulus it was given. For a run in the model, $\mathbf{U}$ means the computed $\widehat{\mathbf{U}}$.

Remark. The classical definition takes the maximum over **every** intermediate matrix $\mathbf{A}^{(0)} = \mathbf{A}, \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(n-1)} = \mathbf{U}$ of the row-operation ordering, not only over $\mathbf{U}$. That quantity is at least g_n , so any upper bound proved for it bounds g_n too, and it is what the growth theorem of the next subsection actually estimates. The ratio above is the form the backward error bound needs.

Suppose a pivoting rule keeps every multiplier at most 1 in modulus — the next subsection exhibits one, for the algorithm above as well as for Chapter 3's ordering. Then $|\widehat{\ell}_{ip}| \leq 1$, so for all i, j

$$(|\widehat{\mathbf{L}}||\widehat{\mathbf{U}}|)_{ij} \leq \sum_p |\widehat{u}_{pj}| \leq n \max_{p,q} |\widehat{u}_{pq}| = n g_n \max_{p,q} |a_{pq}|,$$

and Theorem 24.2.6 becomes

$$\max_{i,j} |\Delta a_{ij}| \leq n \gamma_n g_n \max_{i,j} |a_{ij}|. \quad (24.2.1)$$

Since γ_n is about nu , the relative backward error is about $n^2 g_n u$. The dimensional factor n^2 is pessimistic and nobody minds it. **Everything therefore depends on g_n** — and on nothing else.

24.2.6 Pivoting

Proposition 24.2.7. Partial Pivoting Bounds the Multipliers

Run elimination with **partial pivoting**: at step k , before eliminating, exchange row k with a row $r \geq k$ for which $|a_{rk}^{(k-1)}|$ is largest. Then every multiplier satisfies $|\ell_{ik}| \leq 1$, in exact arithmetic; and in the model $|\widehat{\ell}_{ik}| \leq 1$ as well.

Proof. After the exchange, the pivot $a_{kk}^{(k-1)}$ has the largest modulus in its column among rows $k, \dots, n$. If it is zero then the whole column is zero, every multiplier is set to 0, and there is nothing to prove. Otherwise $|a_{ik}^{(k-1)}| \leq |a_{kk}^{(k-1)}|$ for $i > k$, so $|\ell_{ik}| = |a_{ik}^{(k-1)}|/|a_{kk}^{(k-1)}| \leq 1$.

In the model, $\widehat{\ell}_{ik} = \text{fl}(x)$ for a real x with $|x| \leq 1$. Both 1 and -1 lie in $\mathcal{R}$, by Definition 24.1.1, and the one with the sign of x is at distance $1 - |x|$ from x ; since $\text{fl}(x)$ is a **nearest** element of $\mathcal{R}$, also $|\text{fl}(x) - x| \leq 1 - |x|$. If $|\text{fl}(x)|$ were greater than 1 we would get $|\text{fl}(x) - x| \geq |\text{fl}(x)| - |x| > 1 - |x|$, a contradiction. So $|\widehat{\ell}_{ik}| \leq 1$. ■

Remark. **The same rule, one entry at a time.** The proposition is stated for Chapter 3's ordering, which has the column entries $a_{rk}^{(k-1)}$ in hand at step k . In the entry-at-a-time algorithm those entries appear as the candidate numerators $a_{ri} - \sum_{k < i} \widehat{\ell}_{rk} \widehat{u}_{ki}$ with $r \geq i$, and partial pivoting there means: form all of them, exchange rows i and r so that the largest of them in modulus becomes $\widehat{u}_{ii}$, and only **then** divide. The proof above applies word for word, with the candidate numerator in place of $a_{rk}^{(k-1)}$ — it uses nothing about those numbers except that the divisor is the largest of them in modulus — so every multiplier of that ordering satisfies $|\widehat{\ell}_{ri}| \leq 1$ too. That is what licenses the hypothesis under which Eq. 24.2.1 was derived for the computed factors of Theorem 24.2.6.

Theorem 24.2.8. Growth Under Partial Pivoting

In exact arithmetic, elimination with partial pivoting on any $\mathbf{A} \in M_n(\mathbb{R})$ has growth factor $g_n \leq 2^{n-1}$.

Idea. One step replaces an entry by itself minus a multiple, of modulus at most 1, of another entry. Two numbers of modulus at most M cannot make more than $2M$. Repeat $n - 1$ times.

Proof. Write $M_k = \max_{i,j} |a_{ij}^{(k)}|$, and let $\mathbf{B} = (b_{ij})$ be $\mathbf{A}^{(k-1)}$ after the row exchange of step k ; an exchange only permutes entries, so $\max_{i,j} |b_{ij}| = M_{k-1}$. Every entry of $\mathbf{A}^{(k)}$ is either an entry of $\mathbf{B}$ (the rows $1, \dots, k$ are left alone, and the entries below the pivot are set to 0) or of the form

$$a_{ij}^{(k)} = b_{ij} - \ell_{ik} b_{kj}, \quad |a_{ij}^{(k)}| \leq |b_{ij}| + |\ell_{ik}| |b_{kj}| \leq 2M_{k-1},$$

using $|\ell_{ik}| \leq 1$ from Proposition 24.2.7. Hence $M_k \leq 2M_{k-1}$ for every k , and by induction

$M_k \leq 2^k M_0$. The intermediate matrices are $\mathbf{A}^{(0)}, \dots, \mathbf{A}^{(n-1)}$, so

$$\max_{i,j,k} |a_{ij}^{(k)}| = \max_{0 \leq k \leq n-1} M_k \leq 2^{n-1} M_0,$$

which is $g_n \leq 2^{n-1}$. This proves the theorem. ■

Remark. One expects the same argument in the model to give $M_k \leq 2(1+u)^2 M_{k-1}$, the subtraction and the multiplication each contributing a factor $1+\delta$, and hence $g_n \leq (2(1+u)^2)^{n-1}$, which for any u worth using is 2^{n-1} in all but the last digits. **We have not written that induction out**, and nothing in this section cites it: [Theorem 24.2.8](#), the exact-arithmetic statement, is the only growth bound we have, and the passages that carry it over to a computed factorization say that they are doing so.

Quick check. Run elimination on $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ without row swaps, and again with partial pivoting, and compute g_2 each time. Does partial pivoting make the growth factor smaller?

Solution. Without a swap the multiplier is 3 and $u_{22} = 4 - 3 \cdot 2 = -2$, so $\mathbf{U} = \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix}$ and $g_2 = 2/4 = \frac{1}{2}$. With partial pivoting the rows are exchanged, the multiplier is $\frac{1}{3}$, and $u_{22} = 2 - \frac{1}{3} \cdot 4 = \frac{2}{3}$, so $\mathbf{U} = \begin{pmatrix} 3 & 4 \\ 0 & 2/3 \end{pmatrix}$ and $g_2 = 4/4 = 1$. Pivoting made the growth factor *larger*, and $g_n < 1$ is perfectly possible. Partial pivoting does not minimize growth; it bounds the multipliers, and [Theorem 24.2.8](#) turns that into a worst-case bound — here $2^1 = 2$, which both runs respect.

The bound 2^{n-1} is dreadful, and it is attained.

Example. A matrix that really grows

Let $\mathbf{Y}_n \in M_n(\mathbb{R})$ have 1 on the diagonal, -1 strictly below it, 0 strictly above it except in the last column, and 1 throughout the last column. For $n = 4$,

$$\mathbf{Y}_4 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}.$$

Run elimination with partial pivoting on $\mathbf{Y}_4$ and on $\mathbf{Y}_5$, and compute the growth factor of each.

Solution. The runs below are in exact arithmetic, so the factors are $\mathbf{L}$ and $\mathbf{U}$, with no hats. Every entry of $\mathbf{Y}_n$ has modulus 1, so g_n is simply the largest modulus of an entry of the resulting $\mathbf{U}$.

Step 1 on $\mathbf{Y}_4$. Column 1 is $(1, -1, -1, -1)$; all four entries have modulus 1, so the largest is already in place and no exchange is made. The multipliers are -1 , so each lower row has row 1 **added** to it. Rows 2, 3, 4 become $(0, 1, 0, 2)$, $(0, -1, 1, 2)$, $(0, -1, -1, 2)$: the last column has doubled.

Step 2. Column 2 below the diagonal is $(-1, -1)$ against the pivot 1; again no exchange, multipliers -1 . Rows 3, 4 become $(0, 0, 1, 4)$ and $(0, 0, -1, 4)$.

Step 3. Multiplier -1 again; row 4 becomes $(0, 0, 0, 8)$. So

$$\mathbf{U} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix}, \quad g_4 = \frac{8}{1} = 8 = 2^3.$$

For $n = 5$ the same three observations repeat once more: no exchange is ever made, every multiplier is -1 , and the last column runs $1, 2, 4, 8, 16$ down the diagonal of $\mathbf{U}$, giving $g_5 = 16 = 2^4$.

Every entry of $\mathbf{Y}_n$ is ± 1 , and partial pivoting does exactly what it is told; for these two sizes the bound of [Theorem 24.2.8](#) is attained, not approached. The same three observations run by induction for every n , and [Exercise \(C2: The growth bound is attained for every \$n\$ \)](#) carries that induction out.

Warning. Partial pivoting is not proved stable by anything in this chapter. [Eq. 24.2.1](#) bounds the backward error by $n\gamma_n g_n \max_{i,j} |a_{ij}|$, where g_n is the growth in the **computed** factors, and the only bound we have on growth is [Theorem 24.2.8](#)'s 2^{n-1} , for the **exact** elimination; the remark after [Theorem 24.2.6](#) records that gap and we have not closed it. Grant it anyway, and the resulting backward error bound is about $n^2 2^{n-1} u$, which for $n = 60$ exceeds 1 and says nothing at all — so even the optimistic reading proves nothing. What is true in practice is that g_n is almost always a small multiple of 1 — but “almost always” is a statement about the matrices people meet, not a theorem, and [Example \(A matrix that really grows\)](#), with [Exercise \(C2: The growth bound is attained for every \$n\$ \)](#) for general n , shows there is no theorem to be had. Partial pivoting is used because it is cheap and because its failures are rare and constructed, not because it is backward stable in the sense of [Definition 24.1.3](#). That is an honest summary of the state of the subject, and the reader should not accept a tidier one.

24.2.7 The example Chapter 3 promised

Chapter 3 §06 ended its discussion of pivoting with a two-by-two matrix and a promise: after describing how $\begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix}$ makes elimination produce a computed $\mathbf{L}\mathbf{U}$ whose $(2, 2)$ entry is 0 instead of 1, it wrote “Swapping the rows first gives the harmless multiplier ε . Chapter 24 makes this precise.” Here it is, made precise.

Example. The tiny pivot, in the model

Work in the six-digit decimal model of Example (The model we compute in), where $u =$

5×10^{-6} , and take

$$\mathbf{A} = \begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \varepsilon = 10^{-8}.$$

Factor and solve without an exchange, then with one. In each case compute $\hat{\mathbf{L}}\hat{\mathbf{U}}$ exactly and report the backward error.

Solution. The exact answer. Subtracting the first equation's ε -multiple, $(1 - \varepsilon)x_1 = 1$, so

$$\mathbf{x} = \left(\frac{1}{1 - \varepsilon}, \frac{1 - 2\varepsilon}{1 - \varepsilon} \right) = (1.00000001 \dots, 0.99999999 \dots).$$

Also $\kappa_\infty(\mathbf{A}) = 4/(1 - \varepsilon) < 5$, exactly as in Chapter 16 §08: the problem is well conditioned.

Without the exchange. The multiplier is $\text{fl}(1/\varepsilon) = 10^8$, and

$$\hat{u}_{22} = \text{fl}(1 - \text{fl}(10^8 \cdot 1)) = \text{fl}(1 - 10^8) = \text{fl}(-99999999) = -10^8,$$

the last step by rounding to six digits. Forward substitution gives $\hat{y}_1 = 1$ and $\hat{y}_2 = \text{fl}(2 - 10^8) = -10^8$; back substitution gives $\hat{x}_2 = \text{fl}(-10^8 / -10^8) = 1$ and then

$$\hat{x}_1 = \text{fl}((1 - 1)/\varepsilon) = 0.$$

The computed answer is $(0, 1)$ — Chapter 16 §08's $(0, 1)$, reproduced here in a model rather than by hand-waving about digits. Multiplying the computed factors *exactly*,

$$\begin{aligned} \hat{\mathbf{L}}\hat{\mathbf{U}} &= \begin{pmatrix} 1 & 0 \\ 10^8 & 1 \end{pmatrix} \begin{pmatrix} 10^{-8} & 1 \\ 0 & -10^8 \end{pmatrix} \\ &= \begin{pmatrix} 10^{-8} & 1 \\ 1 & 0 \end{pmatrix} = \mathbf{A} + \Delta\mathbf{A}, \quad \Delta\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}. \end{aligned}$$

So $\|\Delta\mathbf{A}\|_\infty / \|\mathbf{A}\|_\infty = \frac{1}{2}$. The backward error is not $O(u)$; it is $O(1)$. This factorization is **not** backward stable, and it is not even close.

With the exchange. Partial pivoting compares $|\varepsilon|$ with 1 and swaps, so we factor $\mathbf{PA} = \begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix}$ with $\mathbf{Pb} = (2, 1)$. The multiplier is $\text{fl}(\varepsilon/1) = \varepsilon$, and $\hat{u}_{22} = \text{fl}(1 - \text{fl}(\varepsilon \cdot 1)) = \text{fl}(1 - 10^{-8}) = 1$, since $1 - 10^{-8}$ lies between the representable neighbors 0.999999 and 1 and is far nearer the second. Forward substitution gives $\hat{y}_1 = 2$ and $\hat{y}_2 = \text{fl}(1 - \text{fl}(2\varepsilon)) = \text{fl}(1 - 2 \times 10^{-8}) = 1$; back substitution gives $\hat{x}_2 = 1$ and $\hat{x}_1 = \text{fl}((2 - 1)/1) = 1$. The computed answer is $(1, 1)$, whose relative error against the exact $\mathbf{x}$ is about 10^{-8} , far *better* than u . Multiplying the computed factors exactly,

$$\begin{aligned} \hat{\mathbf{L}}\hat{\mathbf{U}} &= \begin{pmatrix} 1 & 0 \\ 10^{-8} & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 \\ 10^{-8} & 1 + 10^{-8} \end{pmatrix} = \mathbf{PA} + \Delta\mathbf{A}, \quad \Delta\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & 10^{-8} \end{pmatrix}, \end{aligned}$$

so $\|\Delta\mathbf{A}\|_\infty / \|\mathbf{A}\|_\infty = 5 \times 10^{-9}$, comfortably below u . This factorization is backward stable.

What the exchange achieved. Exactly what Eq. 24.2.1 predicts. Without it the growth factor is $g_2 = 10^8$: an entry of modulus 10^8 is created out of a matrix whose largest entry is 1, and creating it destroys the information in $a_{22} = 1$, which lies beneath its last digit. With the

exchange, [Proposition 24.2.7](#) keeps the multiplier at $\varepsilon \leq 1$, $g_2 = 1$, and nothing large is ever made. Note what was **not** the problem: $\kappa_\infty(\mathbf{A}) < 5$ throughout. Chapter 16 §08 said so: “Conditioning bounds what *any* method can promise; whether a given method keeps that promise is stability.”

24.2.8 Cholesky needs no pivoting

For a positive definite matrix the whole difficulty disappears, and the reason is a one-line inequality about the entries of the factor. Chapter 13 §02 asserted it — “it needs no pivoting: the recursion above never divides by something that positive definiteness allows to be zero” — and left the arithmetic half here.

Theorem 24.2.9. Cholesky Does Not Grow

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite with representable entries, and let $\mathbf{L}$ be its Cholesky factor, $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$ with $\ell_{ii} > 0$ (Theorem 13.2.3).

- (a) $|\ell_{ij}| \leq \sqrt{a_{ii}}$ for all $j \leq i$, and

$$(|\mathbf{L}||\mathbf{L}^\top|)_{ik} \leq \sqrt{a_{ii}a_{kk}} \leq \max_{p,q} |a_{pq}| \quad \text{for all } i, k.$$

- (b) Suppose $(n+1)u < 1$ and the recursion of Chapter 13 §02 runs to completion in the model, producing $\hat{\mathbf{L}}$. Then

$$\hat{\mathbf{L}}\hat{\mathbf{L}}^\top = \mathbf{A} + \Delta\mathbf{A}, \quad |\Delta\mathbf{A}| \leq \gamma_{n+1} |\hat{\mathbf{L}}| |\hat{\mathbf{L}}^\top|.$$

- (c) If moreover $\gamma_{n+1} < 1$, then

$$\max_{i,j} |\Delta a_{ij}| \leq \frac{\gamma_{n+1}}{1 - \gamma_{n+1}} \max_{i,j} |a_{ij}|.$$

Idea. For (a): the rows of $\mathbf{L}$ have known lengths, because $(\mathbf{L}\mathbf{L}^\top)_{ii} = a_{ii}$ is the squared length of row i . An entry is no longer than its row, and Cauchy–Schwarz compares two rows. For (b): each off-diagonal entry is [Lemma 24.2.4](#) again, and the diagonal entry is the same lemma followed by a square root, which contributes two more factors. For (c): part (b) bounds the row lengths of $\hat{\mathbf{L}}$ in terms of $\mathbf{A}$, and then (a)’s argument runs again on $\hat{\mathbf{L}}$.

Proof.

- (a) Row i of $\mathbf{L}$ has $\sum_p \ell_{ip}^2 = (\mathbf{L}\mathbf{L}^\top)_{ii} = a_{ii}$, so every single term satisfies $\ell_{ij}^2 \leq a_{ii}$. By the Cauchy–Schwarz inequality (Theorem 11.1.5) applied to the rows i and k of $|\mathbf{L}|$,

$$(|\mathbf{L}||\mathbf{L}^\top|)_{ik} = \sum_p |\ell_{ip}| |\ell_{kp}| \leq \left(\sum_p \ell_{ip}^2 \right)^{1/2} \left(\sum_p \ell_{kp}^2 \right)^{1/2} = \sqrt{a_{ii}a_{kk}}.$$

Finally $\sqrt{a_{ii}a_{kk}} \leq \max_p a_{pp}$, and $\max_p a_{pp} = \max_{p,q} |a_{pq}|$: the diagonal entries are positive (take $\mathbf{x} = \mathbf{e}_p$ in Definition 13.1.1), and $|a_{pq}| = |(\mathbf{L}\mathbf{L}^\top)_{pq}| \leq (|\mathbf{L}||\mathbf{L}^\top|)_{pq} \leq \sqrt{a_{pp}a_{qq}} \leq \max_p a_{pp}$, the middle step by the display just proved.

- (b) Fix $i > j$. The computation $\hat{\ell}_{ij} = \text{fl}((a_{ij} - \sum_{k < j} \hat{\ell}_{ik} \hat{\ell}_{jk}) / \hat{\ell}_{jj})$ is [Lemma 24.2.4](#) with

$m = j - 1$ and $d = \widehat{\ell}_{jj}$, so, writing θ_j for θ_0 ,

$$a_{ij} = \sum_{p=1}^j \widehat{\ell}_{ip} \widehat{\ell}_{jp} (1 + \theta_p), \quad |\theta_p| \leq \gamma_j \leq \gamma_{n+1}.$$

Since $\widehat{\ell}_{jp} = 0$ for $p > j$, the sum without the θ_p is $(\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top)_{ij}$, so

$$|a_{ij} - (\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top)_{ij}| \leq \gamma_{n+1} \sum_{p \leq j} |\widehat{\ell}_{ip}| |\widehat{\ell}_{jp}| = \gamma_{n+1} (|\widehat{\mathbf{L}}| |\widehat{\mathbf{L}}^\top|)_{ij}.$$

For $i = j$, let $\widehat{s}$ be the value accumulated before the square root, so $\widehat{\ell}_{jj} = \text{fl}(\sqrt{\widehat{s}}) = \sqrt{\widehat{s}}(1 + \eta)$ with $|\eta| \leq u$, and $\widehat{s} = \widehat{\ell}_{jj}^2(1 + \eta)^{-2}$. By the **proof** of Lemma 24.2.4 with $m = j - 1$ and no division — we need not just the size of θ'_0 but its shape, and the proof produces $1 + \theta'_0$ as the product of the $j - 1$ factors $(1 + \varepsilon_p)^{-1}$ —

$$a_{jj} = \widehat{s}(1 + \theta'_0) + \sum_{p < j} \widehat{\ell}_{jp}^2 (1 + \theta_p), \quad |\theta'_0|, |\theta_p| \leq \gamma_{j-1}.$$

Substituting $\widehat{s}$, the first term is $\widehat{\ell}_{jj}^2$ times a product of $(j - 1) + 2 \leq n + 1$ factors of the form $(1 + \delta)^{\pm 1}$, which by Lemma 24.1.1 is $1 + \theta_j$ with $|\theta_j| \leq \gamma_{n+1}$. So again $a_{jj} = \sum_{p \leq j} \widehat{\ell}_{jp} \widehat{\ell}_{jp} (1 + \theta_p)$ with all $|\theta_p| \leq \gamma_{n+1}$, and the displayed estimate follows as before. Since $\widehat{\mathbf{L}}$ is lower triangular, the entries with $i < j$ are the transposes of these, and $\Delta \mathbf{A}$ is symmetric.

- (c) The diagonal case of (b) reads $|a_{ii} - (\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top)_{ii}| \leq \gamma_{n+1} (\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top)_{ii}$, because $(|\widehat{\mathbf{L}}| |\widehat{\mathbf{L}}^\top|)_{ii} = \sum_p \widehat{\ell}_{ip}^2 = (\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top)_{ii}$. Hence $(\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top)_{ii}(1 - \gamma_{n+1}) \leq a_{ii}$, that is

$$\sum_p \widehat{\ell}_{ip}^2 \leq \frac{a_{ii}}{1 - \gamma_{n+1}}.$$

Cauchy–Schwarz as in (a), now for $\widehat{\mathbf{L}}$, gives $(|\widehat{\mathbf{L}}| |\widehat{\mathbf{L}}^\top|)_{ik} \leq \sqrt{a_{ii} a_{kk}} / (1 - \gamma_{n+1}) \leq \max_{p,q} |a_{pq}| / (1 - \gamma_{n+1})$. Substituting into (b) gives (c). This proves the theorem. ■

Compare (c) with Eq. 24.2.1. For a general matrix the backward error carried the factor g_n , which partial pivoting bounds only by 2^{n-1} , sharply. Here the same quantity is bounded by the largest entry of $\mathbf{A}$ itself, with **no growth factor and no pivoting**: the Cholesky recursion is unconditionally backward stable. The reason sits entirely in part (a) — the rows of the factor cannot be longer than the diagonal of $\mathbf{A}$ allows — and that is a fact about positive definiteness, not about arithmetic. Chapter 13 §02 called positive definiteness “the one structural hypothesis under which elimination is unconditionally safe”; this is the proof.

Warning. Part (a) bounds $|\ell_{ij}|$ by $\sqrt{a_{ii}}$, the diagonal entry of its own row, not of its column. The column version is false. For

$$\mathbf{A} = \begin{pmatrix} 0.01 & 0.09 \\ 0.09 & 100 \end{pmatrix},$$

which is positive definite since $0.01 > 0$ and $\det \mathbf{A} = 1 - 0.0081 > 0$ (Theorem 13.1.5 (d)), the factor is $\ell_{11} = 0.1$, $\ell_{21} = 0.9$, and $|\ell_{21}| = 0.9$ is nine times $\sqrt{a_{11}} = 0.1$. It is at most $\sqrt{a_{22}} = 10$, as (a) requires. An entry of $\mathbf{L}$ can be far larger than the pivot it was divided by;

what it cannot exceed is the size of its own row.

24.2.9 Exercises

24.2.9.1 A. Check your understanding

Exercise. A1

- Define a flop and the growth factor g_n .
- State the partial pivoting rule, and say what [Proposition 24.2.7](#) gets from it.
- State the leading term of the flop count for an LU factorization, for two triangular solves and for Cholesky.
- Decide whether each statement is correct, with a reason: (i) “a triangular solve needs pivoting when the diagonal entries are small”; (ii) “elimination with partial pivoting is backward stable”; (iii) “a backward stable factorization gives an accurate solution”.
- In one sentence, what does [Theorem 24.2.1](#) say about the trailing block after k steps?

Solution.

- A flop is one arithmetic operation — an addition, subtraction, multiplication, division, or square root. The growth factor is $g_n = \max_{i,j} |u_{ij}| / \max_{i,j} |a_{ij}|$, the largest entry in modulus the elimination produces relative to the largest in modulus it was given ([Definition 24.2.1](#)); the classical version maximizes over every intermediate matrix as well, and is at least as large.
- At step k , exchange row k with the row $r \geq k$ maximizing $|a_{rk}^{(k-1)}|$. The consequence is $|\ell_{ik}| \leq 1$ for every multiplier, in exact arithmetic and in the model.
- $\frac{2}{3}n^3$, $2n^2$, and $\frac{1}{3}n^3$ respectively ([Proposition 24.2.3](#)); the third is half the first.
- (i) Incorrect. [Theorem 24.2.5](#) holds for every nonsingular triangular matrix, with no hypothesis on the diagonal and no pivoting. (ii) Incorrect as stated. The only growth bound available is [Theorem 24.2.8](#)’s $g_n \leq 2^{n-1}$, proved for the exact elimination, and [Example \(A matrix that really grows\)](#) shows it is attained; even granting its transfer to the computed factors — the gap recorded in the remark after [Theorem 24.2.6](#) — the resulting backward error bound $n^2 2^{n-1} u$ is vacuous for moderate n . (iii) Incorrect. It gives an accurate solution only when the problem is well conditioned; [Theorem 24.1.3](#) supplies the missing factor $\kappa(\mathbf{A})$.
- It is the Schur complement $\mathbf{A}/\mathbf{A}_k$ of the leading $k \times k$ block — so elimination is the repeated formation of Schur complements of 1×1 pivots.

24.2.9.2 B. Practice

Exercise. B1: A Schur complement by two routes

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & -1 \\ 4 & 5 & -1 \\ -2 & 8 & 2 \end{pmatrix}.$$

- (a) Carry out one step of elimination without row swaps and read off the trailing 2×2 block.
- (b) Compute $\mathbf{A}/\mathbf{A}_1 = \mathbf{D} - \mathbf{C}a_{11}^{-1}\mathbf{B}$ directly from Definition 8.3.1 and check that the two agree.
- (c) Carry out the second step and verify Corollary 24.2.2 (a) by computing $\det \mathbf{A}_1$, $\det \mathbf{A}_2$, and $\det \mathbf{A}_3$.

Solution.

- (a) The multipliers are $\ell_{21} = 4/2 = 2$ and $\ell_{31} = -2/2 = -1$. Subtracting gives rows $(0, 3, 1)$ and $(0, 9, 1)$, so the trailing block is $\begin{pmatrix} 3 & 1 \\ 9 & 1 \end{pmatrix}$.
- (b) Here $\mathbf{A}_1 = (2)$, $\mathbf{B} = (1, -1)$, $\mathbf{C} = (4, -2)^\top$, $\mathbf{D} = \begin{pmatrix} 5 & -1 \\ 8 & 2 \end{pmatrix}$, so

$$\mathbf{C}a_{11}^{-1}\mathbf{B} = \frac{1}{2} \begin{pmatrix} 4 \\ -2 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{A}/\mathbf{A}_1 = \begin{pmatrix} 3 & 1 \\ 9 & 1 \end{pmatrix},$$

which agrees with (a), as Theorem 24.2.1 says it must.

- (c) The second multiplier is $9/3 = 3$, and the second step turns $(0, 9, 1)$ into $(0, 0, 1 - 3) = (0, 0, -2)$. So the pivots are $2, 3, -2$. Now $\det \mathbf{A}_1 = 2$, $\det \mathbf{A}_2 = 10 - 4 = 6$, and $\det \mathbf{A} = 2(10 + 8) - 1(8 - 2) + (-1)(32 + 10) = 36 - 6 - 42 = -12$. The ratios are $2/1 = 2$, $6/2 = 3$, and $-12/6 = -2$: the pivots, as claimed.

Exercise. B2: Growth, with and without a swap

Let $\mathbf{A} = \begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix}$, and work in the six-digit decimal model.

- (a) Factor $\mathbf{A}$ without an exchange, compute g_2 , and compute $\widehat{\mathbf{L}}\widehat{\mathbf{U}}$ exactly.
- (b) Do the same after exchanging the rows.
- (c) Compare the two backward errors with the bound of Eq. 24.2.1.

Solution.

- (a) The multiplier is $\text{fl}(1/10^{-4}) = 10^4$, and $\widehat{u}_{22} = \text{fl}(1 - 10^4) = -9999.00$, which is exactly representable. So

$$\widehat{\mathbf{L}}\widehat{\mathbf{U}} = \begin{pmatrix} 1 & 0 \\ 10^4 & 1 \end{pmatrix} \begin{pmatrix} 10^{-4} & 1 \\ 0 & -9999 \end{pmatrix} = \begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} = \mathbf{A},$$

so here $\Delta\mathbf{A} = \mathbf{O}$: with six digits, $1 - 10^4$ is *not* rounded, and nothing is lost. The growth factor is $g_2 = 9999/1 = 9999$.

- (b) After the exchange the multiplier is 10^{-4} and $\widehat{u}_{22} = \text{fl}(1 - 10^{-4}) = 0.999900$, again exact, so $\widehat{\mathbf{L}}\widehat{\mathbf{U}} = \mathbf{P}\mathbf{A}$ exactly and $g_2 = 1$.
- (c) Both backward errors are zero, so both satisfy the bound. For $n = 2$ the entry-at-a-time algorithm and Chapter 3's row operations perform the very same three operations, so Eq. 24.2.1 applies to these runs as it stands, with g_2 read off the computed factors. The point is the *bound*, not the outcome: in (a) it is $2\gamma_2 g_2 \approx 4 \times 9999 u \approx 0.2$, which permits

a catastrophe, while in (b) it is about $4u$, which does not. Large growth does not force a large backward error; it removes the guarantee that there is none, and [Example \(The tiny pivot, in the model\)](#) shows what happens when the permission is taken up.

Exercise. B3: The Cholesky entry bound

Let $\mathbf{A} = \begin{pmatrix} 4 & -2 & 6 \\ -2 & 5 & -1 \\ 6 & -1 & 26 \end{pmatrix}$, whose Cholesky factor was computed in Example (A Cholesky Factorization, and One That Fails) to be $\mathbf{L}$ with rows $(2, 0, 0)$, $(-1, 2, 0)$, $(3, 1, 4)$.

- Check [Theorem 24.2.9](#) (a) entry by entry.
- Compute $|\mathbf{L}||\mathbf{L}^\top|$ and check that every entry is at most $\max_{p,q} |a_{pq}|$.
- What would the bound of [Eq. 24.2.1](#) require if $\mathbf{A}$ were factored by general elimination with partial pivoting instead?

Solution.

- The bound is $|\ell_{ij}| \leq \sqrt{a_{ii}}$. Row 1: $2 \leq \sqrt{4} = 2$, with equality. Row 2: $1, 2 \leq \sqrt{5} \approx 2.236$. Row 3: $3, 1, 4 \leq \sqrt{26} \approx 5.099$. All hold, and the row sums of squares are 4, 5, 26, which are a_{11}, a_{22}, a_{33} exactly.
- $|\mathbf{L}|$ has rows $(2, 0, 0)$, $(1, 2, 0)$, $(3, 1, 4)$, so

$$|\mathbf{L}||\mathbf{L}^\top| = \begin{pmatrix} 4 & 2 & 6 \\ 2 & 5 & 5 \\ 6 & 5 & 26 \end{pmatrix}.$$

The largest entry of $\mathbf{A}$ in modulus is 26, and every entry above is at most 26 — indeed the (i, k) entry is at most $\sqrt{a_{ii}a_{kk}}$: for instance $5 \leq \sqrt{5 \cdot 26} \approx 11.4$.

- It would require knowing g_3 for the computed factors, and the only bound available is [Theorem 24.2.8](#)'s exact-arithmetic $2^2 = 4$; granting the transfer between the two (the remark after [Theorem 24.2.6](#)), [Eq. 24.2.1](#) permits $\max |\Delta a_{ij}| \leq 3\gamma_3 \cdot 4 \cdot 26 = 12\gamma_3 \cdot 26$. [Theorem 24.2.9](#) (c) permits $(\gamma_4/(1 - \gamma_4)) \cdot 26$ instead: the factor $3g_3 = 12$ disappears entirely, and the price is the step from γ_3 to $\gamma_4/(1 - \gamma_4)$, which is about $\frac{4}{3}\gamma_3$. No pivoting is needed to get it.

24.2.9.3 C. Going deeper

Exercise. C1: Column diagonal dominance needs no pivoting

Call $\mathbf{A} \in M_n(\mathbb{R})$ **strictly column diagonally dominant** if $|a_{jj}| > \sum_{i \neq j} |a_{ij}|$ for every j .

- Show that partial pivoting makes no exchange at the first step.
- Show that $\mathbf{A}/\mathbf{A}_1$ is again strictly column diagonally dominant, and that each of its column sums $\sum_{i \geq 2} |a'_{ij}|$ is at most the corresponding column sum $\sum_{i \geq 1} |a_{ij}|$ of $\mathbf{A}$.

- (c) Deduce that elimination without any exchanges succeeds and that $g_n < 2$.

Hint: for (c), bound every entry of every $\mathbf{A}^{(k)}$ by a column sum of $\mathbf{A}$.

Solution.

- (a) Column 1 satisfies $|a_{11}| > \sum_{i \geq 2} |a_{i1}| \geq |a_{i1}|$ for each $i \geq 2$, so the entry of largest modulus in column 1 is already a_{11} , and it is nonzero.
- (b) Write $a'_{ij} = a_{ij} - (a_{i1}/a_{11})a_{1j}$ for $i, j \geq 2$. For the column sums, the triangle inequality gives

$$\sum_{i \geq 2} |a'_{ij}| \leq \sum_{i \geq 2} |a_{ij}| + \frac{|a_{1j}|}{|a_{11}|} \sum_{i \geq 2} |a_{i1}| \leq \sum_{i \geq 2} |a_{ij}| + |a_{1j}| = \sum_{i \geq 1} |a_{ij}|,$$

using $\sum_{i \geq 2} |a_{i1}| < |a_{11}|$. For dominance, fix $j \geq 2$. Then

$$\sum_{\substack{i \geq 2 \\ i \neq j}} |a'_{ij}| \leq \sum_{\substack{i \geq 2 \\ i \neq j}} |a_{ij}| + \frac{|a_{1j}|}{|a_{11}|} \sum_{\substack{i \geq 2 \\ i \neq j}} |a_{i1}| < (|a_{jj}| - |a_{1j}|) + \frac{|a_{1j}|}{|a_{11}|} (|a_{11}| - |a_{j1}|),$$

where the two strict bounds are column- j and column-1 dominance. The right side equals $|a_{jj}| - |a_{1j}||a_{j1}|/|a_{11}|$, which is at most $|a_{jj} - a_{j1}a_{1j}/a_{11}| = |a'_{jj}|$ by the reverse triangle inequality. So $\sum_{i \geq 2, i \neq j} |a'_{ij}| < |a'_{jj}|$.

- (c) By (a) and (b) and induction, every Schur complement $\mathbf{A}^{(k)}$ restricted to its active block is strictly column diagonally dominant, so no exchange is ever made and no pivot is zero. By (b) and induction, the column sums of the active blocks never increase, so every entry of every $\mathbf{A}^{(k)}$ is bounded by $\max_j \sum_i |a_{ij}|$ — and the rows that have been frozen into $\mathbf{U}$ were bounded the same way when they were created. Finally, dominance gives $\sum_i |a_{ij}| < 2|a_{jj}| \leq 2 \max_{p,q} |a_{pq}|$, so $g_n < 2$. Hence such matrices need no pivoting at all.

Exercise. C2: The growth bound is attained for every n

Let $\mathbf{Y}_n$ be the matrix of [Example \(A matrix that really grows\)](#): $y_{ii} = 1$, $y_{ij} = -1$ for $i > j$, $y_{in} = 1$ for all i , and $y_{ij} = 0$ for $j > i$ with $j < n$.

- (a) Prove that partial pivoting makes no exchange at any step, and that every multiplier equals -1 .
- (b) Prove by induction on k that after k steps the active trailing block is again of the form $\mathbf{Y}_{n-k}$, except that its last column is 2^k times the all-ones vector.
- (c) Deduce $g_n = 2^{n-1}$, and explain why this does not contradict the practical success of partial pivoting.

Solution.

- (a) and (b) together, by induction. At step $k = 1$, column 1 is $(1, -1, \dots, -1)$, every entry of modulus 1, so the pivot $y_{11} = 1$ is already of largest modulus and no exchange is made; the multipliers are $-1/1 = -1$. Subtracting -1 times row 1 from row i adds row 1 to row i . Row 1 is $(1, 0, \dots, 0, 1)$, so for $i \geq 2$ the entries in columns $2, \dots, n-1$ are unchanged, the entry in column 1 becomes 0, and the entry in column n becomes

$1 + 1 = 2$. The trailing $(n-1) \times (n-1)$ block therefore has 1 on its diagonal, -1 below, 0 above except in the last column, and 2 throughout its last column: it is $\mathbf{Y}_{n-1}$ with the last column doubled. Since doubling one column changes neither which entry of a column has largest modulus nor the ratios within the other columns, the same applies at every subsequent step, and after k steps the active block is $\mathbf{Y}_{n-k}$ with last column 2^k times the all-ones vector.

- (b) Every entry of $\mathbf{Y}_n$ has modulus 1, so $M_0 = 1$. By (b) the largest entry ever created is the (n, n) entry after step $n-1$, namely 2^{n-1} . So $g_n = 2^{n-1}$, meeting [Theorem 24.2.8](#) exactly.

This does not contradict anything practical: $\mathbf{Y}_n$ is a constructed matrix, and the theorem only says the *worst case* is 2^{n-1} . What the example does destroy is the hope of a theorem: no bound better than 2^{n-1} can be proved for partial pivoting, so any claim that partial pivoting is backward stable is a claim about the matrices one meets, not about all matrices. That is why [Theorem 24.2.9](#) matters — it is a genuine theorem, for a class of matrices identified in advance.