

24.3 Least Squares in Practice

Chapter 11 solved the least-squares problem twice. Section 4 of that chapter produced the normal equations $\mathbf{A}^* \mathbf{A} \mathbf{x} = \mathbf{A}^* \mathbf{b}$ (Theorem 11.4.2), and Section 8 rewrote them, once a $\mathbf{QR}$ factorization was available, as the triangular system $\mathbf{R} \mathbf{x} = \mathbf{Q}^* \mathbf{b}$. It then recommended the second route and postponed the reason: “forming $\mathbf{A}^* \mathbf{A}$ squares the sensitivity of the problem to rounding, while $(\dagger)$ never forms that product at all. Chapter 24 makes ‘sensitivity’ precise and proves the comparison; here we record only the practice.”

Chapter 16 supplied half of the reason. Its Theorem 16.8.4 proved $\kappa_2(\mathbf{A}^* \mathbf{A}) = \kappa_2(\mathbf{A})^2$ and then listed what was still missing: the conditioning of the fitting problem itself, which involves the residual; the arithmetic, which needed a model; and the comparison of the two algorithms in digits. Sections 1 and 2 of this chapter built the model and the vocabulary. This section pays the three debts, in that order, and ends with the reason the reflections of Chapter 11 are the right way to produce $\mathbf{Q}$ and $\mathbf{R}$.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, $\mathbf{A} \in M_{m \times n}(F)$ with $m \geq n$ and $\text{rank } \mathbf{A} = n$, and $\mathbf{b} \in F^m$ is nonzero. By Corollary 11.4.4 there is then exactly one least-squares solution $\mathbf{x}$, and $\mathbf{A}^* \mathbf{A}$ is invertible. All norms are $\|\cdot\|_2$ unless a subscript says otherwise.

24.3.1 How sensitive is the problem?

Before any algorithm is judged, the problem has to be measured. Write

$$\mathbf{r} = \mathbf{b} - \mathbf{A} \mathbf{x}, \quad \mathbf{A} \mathbf{x} = P_{\text{col}(\mathbf{A})} \mathbf{b},$$

the residual and the projection of Theorem 11.4.2 (b). Since $\mathbf{r} \perp \text{col}(\mathbf{A})$ and $\mathbf{A} \mathbf{x} \in \text{col}(\mathbf{A})$, Pythagoras (Theorem 11.1.3) gives $\|\mathbf{b}\|^2 = \|\mathbf{A} \mathbf{x}\|^2 + \|\mathbf{r}\|^2$. So there is a unique $\theta \in [0, \pi/2]$ with

$$\cos \theta = \frac{\|\mathbf{A} \mathbf{x}\|}{\|\mathbf{b}\|}, \quad \sin \theta = \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}, \quad (24.3.1)$$

the angle between $\mathbf{b}$ and the column space of $\mathbf{A}$. It is 0 exactly when the system is consistent and $\pi/2$ exactly when $\mathbf{A}^* \mathbf{b} = \mathbf{0}$, in which case $\mathbf{x} = \mathbf{0}$ and no relative error can be spoken of. **We assume $\theta < \pi/2$ throughout, so $\mathbf{x} \neq \mathbf{0}$.**

One number is now missing from Chapter 16’s account, and [Eq. 24.3.1](#) supplies it. Here is the sensitivity of $\mathbf{x}$ to a perturbation of the data.

Theorem 24.3.1. Conditioning of the Least-Squares Problem

Let $\mathbf{A} \in M_{m \times n}(F)$ have rank $n \leq m$, let $\mathbf{b} \neq \mathbf{0}$, let $\mathbf{x}$ be the least-squares solution of $\mathbf{A} \mathbf{z} = \mathbf{b}$, and let $\theta < \pi/2$ be as in [Eq. 24.3.1](#). Let $\mathbf{E} \in M_{m \times n}(F)$ and $\mathbf{f} \in F^m$, and for real t near 0 let $\mathbf{x}(t)$ be the least-squares solution of $(\mathbf{A} + t\mathbf{E})\mathbf{z} = \mathbf{b} + t\mathbf{f}$. Then $\mathbf{x}(t)$ is defined and differentiable near $t = 0$, with $\mathbf{x}(0) = \mathbf{x}$, and

$$\frac{\|\mathbf{x}'(0)\|}{\|\mathbf{x}\|} \leq \kappa_2(\mathbf{A}) \left(\varepsilon_{\mathbf{A}} + \frac{\varepsilon_{\mathbf{b}}}{\cos \theta} \right) + \kappa_2(\mathbf{A})^2 \varepsilon_{\mathbf{A}} \tan \theta,$$

where $\varepsilon_{\mathbf{A}} = \|\mathbf{E}\|/\|\mathbf{A}\|$ and $\varepsilon_{\mathbf{b}} = \|\mathbf{f}\|/\|\mathbf{b}\|$.

Idea. The least-squares solution is $\mathbf{x}(t) = \mathbf{N}(t)^{-1}\mathbf{A}(t)^*\mathbf{b}(t)$ with $\mathbf{N}(t) = \mathbf{A}(t)^*\mathbf{A}(t)$, which is a differentiable function of t because $\mathbf{N}(0)$ is invertible. Rather than differentiate that formula, differentiate the equation $\mathbf{N}(t)\mathbf{x}(t) = \mathbf{A}(t)^*\mathbf{b}(t)$ it comes from: the product rule gives three terms, two of them combine into $\mathbf{E}^*(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{E}^*\mathbf{r}$, and the residual appears by itself. Then read every norm off a singular value decomposition: $\|(\mathbf{A}^*\mathbf{A})^{-1}\| = \sigma_n^{-2}$ is where the square comes from, and the only inequality used to finish is $\sigma_1\|\mathbf{x}\| \geq \|\mathbf{A}\mathbf{x}\| = \|\mathbf{b}\|\cos\theta$.

Proof. Write $\mathbf{A}(t) = \mathbf{A} + t\mathbf{E}$, $\mathbf{b}(t) = \mathbf{b} + t\mathbf{f}$, and $\mathbf{N}(t) = \mathbf{A}(t)^*\mathbf{A}(t)$. Every entry of $\mathbf{N}(t)$ is a polynomial in t , so $\mathbf{N}$ is differentiable in the sense of Definition 16.6.1, and $\mathbf{N}(0) = \mathbf{A}^*\mathbf{A}$ is invertible by Corollary 11.4.4 (c). By Theorem 16.6.5, $\mathbf{N}(t)$ is invertible on an open interval around 0 and $t \mapsto \mathbf{N}(t)^{-1}$ is differentiable there. On that interval $\text{rank } \mathbf{A}(t) = n$, again by Corollary 11.4.4, so Theorem 11.4.2 has the unique solution

$$\mathbf{x}(t) = \mathbf{N}(t)^{-1}\mathbf{A}(t)^*\mathbf{b}(t),$$

differentiable by Lemma 16.6.4. (Conjugation is applied entrywise and commutes with differentiation in the real variable t , so $(\mathbf{A}(t)^*)' = \mathbf{E}^*$.)

Differentiate the identity $\mathbf{N}(t)\mathbf{x}(t) = \mathbf{A}(t)^*\mathbf{b}(t)$ at $t = 0$ with Lemma 16.6.4. Since $\mathbf{N}'(0) = \mathbf{E}^*\mathbf{A} + \mathbf{A}^*\mathbf{E}$,

$$(\mathbf{E}^*\mathbf{A} + \mathbf{A}^*\mathbf{E})\mathbf{x} + \mathbf{A}^*\mathbf{A}\mathbf{x}'(0) = \mathbf{E}^*\mathbf{b} + \mathbf{A}^*\mathbf{f}.$$

Move the first bracket across and collect the two terms carrying $\mathbf{E}^*$:

$$\mathbf{A}^*\mathbf{A}\mathbf{x}'(0) = \mathbf{E}^*(\mathbf{b} - \mathbf{A}\mathbf{x}) - \mathbf{A}^*(\mathbf{E}\mathbf{x}) + \mathbf{A}^*\mathbf{f} = \mathbf{E}^*\mathbf{r} - \mathbf{A}^*(\mathbf{E}\mathbf{x}) + \mathbf{A}^*\mathbf{f}.$$

Hence

$$\mathbf{x}'(0) = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{E}^*\mathbf{r} - (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*(\mathbf{E}\mathbf{x}) + (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{f}. \quad (*)$$

Now the two norms. Write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ as in Theorem 13.9.1, with $\sigma_1 \geq \dots \geq \sigma_n > 0$. As in the proof of Theorem 16.8.4, $\mathbf{A}^*\mathbf{A} = \mathbf{V}\text{diag}(\sigma_1^2, \dots, \sigma_n^2)\mathbf{V}^*$, so

$$(\mathbf{A}^*\mathbf{A})^{-1} = \mathbf{V}\text{diag}(\sigma_1^{-2}, \dots, \sigma_n^{-2})\mathbf{V}^*, \quad (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*,$$

where $\mathbf{\Sigma}^+$ is the $n \times m$ matrix with $\sigma_1^{-1}, \dots, \sigma_n^{-1}$ on its diagonal — the second identity because $\text{diag}(\sigma_i^{-2})\mathbf{\Sigma}^* = \mathbf{\Sigma}^+$. Unitary factors do not change singular values, rectangular ones included: for unitary $\mathbf{U}_1, \mathbf{V}_1$ of the right sizes, $(\mathbf{U}_1\mathbf{M}\mathbf{V}_1)^*(\mathbf{U}_1\mathbf{M}\mathbf{V}_1) = \mathbf{V}_1^*(\mathbf{M}^*\mathbf{M})\mathbf{V}_1$ is unitarily similar to $\mathbf{M}^*\mathbf{M}$, so the two have the same eigenvalues and Definition 13.9.1 reads the same singular values off them. Hence by Theorem 16.3.3 (c),

$$\|(\mathbf{A}^*\mathbf{A})^{-1}\| = \sigma_n^{-2}, \quad \|(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\| = \sigma_n^{-1}.$$

Applying these to (*) with Theorem 16.3.2 (a),

$$\|\mathbf{x}'(0)\| \leq \frac{\|\mathbf{E}\| \|\mathbf{r}\|}{\sigma_n^2} + \frac{\|\mathbf{E}\| \|\mathbf{x}\|}{\sigma_n} + \frac{\|\mathbf{f}\|}{\sigma_n}.$$

Divide by $\|\mathbf{x}\|$ and use $\|\mathbf{A}\| = \sigma_1$ (Theorem 16.3.3 (c)) and $\kappa_2(\mathbf{A}) = \sigma_1/\sigma_n$ (Definition 16.8.1). The middle term becomes $\|\mathbf{E}\|/\sigma_n = \kappa_2(\mathbf{A})\varepsilon_{\mathbf{A}}$ exactly. For the other two, note that $\sigma_1\|\mathbf{x}\| \geq \|\mathbf{Ax}\| = \|\mathbf{b}\| \cos \theta$, so

$$\frac{\|\mathbf{r}\|}{\sigma_1\|\mathbf{x}\|} \leq \frac{\|\mathbf{b}\| \sin \theta}{\|\mathbf{b}\| \cos \theta} = \tan \theta, \quad \frac{\|\mathbf{b}\|}{\sigma_1\|\mathbf{x}\|} \leq \frac{1}{\cos \theta}.$$

Therefore

$$\begin{aligned} \frac{\|\mathbf{E}\|\|\mathbf{r}\|}{\sigma_n^2\|\mathbf{x}\|} &= \frac{\sigma_1^2}{\sigma_n^2} \cdot \frac{\|\mathbf{E}\|}{\sigma_1} \cdot \frac{\|\mathbf{r}\|}{\sigma_1\|\mathbf{x}\|} \leq \kappa_2(\mathbf{A})^2 \varepsilon_{\mathbf{A}} \tan \theta, \\ \frac{\|\mathbf{f}\|}{\sigma_n\|\mathbf{x}\|} &= \frac{\sigma_1}{\sigma_n} \cdot \frac{\|\mathbf{f}\|}{\|\mathbf{b}\|} \cdot \frac{\|\mathbf{b}\|}{\sigma_1\|\mathbf{x}\|} \leq \frac{\kappa_2(\mathbf{A})\varepsilon_{\mathbf{b}}}{\cos \theta}. \end{aligned}$$

Adding the three estimates gives the stated bound. This proves the theorem. ■

What is linearized, and what is not. The theorem itself is exact: it bounds a derivative, and no term has been discarded. What is an approximation is the use one makes of it. Reading it as

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \lesssim \left[\kappa_2(\mathbf{A}) \left(1 + \frac{1}{\cos \theta}\right) + \kappa_2(\mathbf{A})^2 \tan \theta \right] \varepsilon \quad (24.3.2)$$

for data perturbed by $\|\Delta \mathbf{A}\| \leq \varepsilon \|\mathbf{A}\|$ and $\|\Delta \mathbf{b}\| \leq \varepsilon \|\mathbf{b}\|$ drops a remainder of size $O(\varepsilon^2)$, exactly as Theorem 16.8.3 (b) drops the factor $1/(1-r)$. For the $\varepsilon \approx 10^{-16}$ of the floating-point model that remainder is invisible, and every statement below that carries a $\lesssim$ is a statement of this kind.

Two readings of the bound matter.

- **When the residual is zero**, $\theta = 0$ and the bound is $2\kappa_2(\mathbf{A})\varepsilon$: the least-squares problem is exactly as sensitive as a square linear system with the same condition number. A square invertible $\mathbf{A}$ is always this case, and Theorem 16.8.3 is recovered.
- **When the residual is large**, the term $\kappa_2(\mathbf{A})^2 \tan \theta$ dominates the bound, and what it measures belongs to the *problem*: no algorithm can return an answer more accurate than data of that quality determine. This is the sentence Chapter 16 §08 promised, where the least-squares problem was said to have “a sensitivity that depends on $\kappa_2(\mathbf{A})$, on the residual $\|\mathbf{Ax}^+ - \mathbf{b}\|$ and on the angle between $\mathbf{b}$ and $\text{col}(\mathbf{A})$ ”. [Exercise \(C1: The squared term is attained\)](#) exhibits data on which the $\kappa_2(\mathbf{A})^2 \varepsilon_{\mathbf{A}} \tan \theta$ term is attained exactly, so it is not an artifact of the estimates.

Quick check. Take $m = n$ with $\mathbf{A}$ invertible. What does [Eq. 24.3.1](#) give for θ , and what does [Eq. 24.3.2](#) reduce to? Compare with Theorem 16.8.3.

Solution. For an invertible square $\mathbf{A}$ the system $\mathbf{Az} = \mathbf{b}$ is consistent, so its least-squares solution has residual $\mathbf{r} = \mathbf{0}$ and [Eq. 24.3.1](#) gives $\sin \theta = 0$, that is $\theta = 0$. Then $\cos \theta = 1$ and $\tan \theta = 0$, and [Eq. 24.3.2](#) reads $\|\Delta \mathbf{x}\|/\|\mathbf{x}\| \lesssim 2\kappa(\mathbf{A})\varepsilon$. That is Theorem 16.8.3 (a) and (b) added together, with the factor $1/(1-r)$ replaced by the $O(\varepsilon^2)$ that the linearization discards. The square case carries no κ^2 at all.

24.3.2 Two routes, and the digits they lose

The problem is measured; now the two algorithms. Both are run in the arithmetic of Definition 24.1.1, with u the unit roundoff and $\gamma_k = ku/(1 - ku)$ the constant of Lemma 24.1.1.

The **normal-equation route** forms $\mathbf{C} = \text{fl}(\mathbf{A}^* \mathbf{A})$ and $\mathbf{d} = \text{fl}(\mathbf{A}^* \mathbf{b})$ entry by entry as inner products, then solves $\mathbf{Cz} = \mathbf{d}$. The matrix it hands to the solver need **not** be positive definite. The exact $\mathbf{A}^* \mathbf{A}$ is: $\mathbf{x}^* \mathbf{A}^* \mathbf{A} \mathbf{x} = \|\mathbf{A} \mathbf{x}\|^2 > 0$ for $\mathbf{x} \neq \mathbf{0}$, the columns of $\mathbf{A}$ being independent. The computed $\mathbf{C}$ is Hermitian, only one triangle of it being formed, but its eigenvalues are those of $\mathbf{A}^* \mathbf{A}$ displaced by the rounding, and the smallest of them, $\sigma_n(\mathbf{A})^2$, can be displaced past 0 — [Example \(A matrix where one route survives and the other does not\)](#) below exhibits a $\mathbf{C}$ that comes out singular, on which the Cholesky factorization of Theorem 13.2.3 fails outright. **When $\mathbf{C}$ is positive definite** — which it is while the rounding stays below $\sigma_n(\mathbf{A})^2$ — Cholesky, and the analysis Section 2 gives it, apply.

The cost is mn^2 flops to form the product — there are $\frac{1}{2}n(n+1)$ entries to compute, each an inner product of length m costing $2m$ flops — plus $\frac{1}{3}n^3$ for the factorization (Proposition 24.2.3 (c)) and $2n^2$ for the two triangular solves. The **QR route** never forms that product: it computes a factorization $\mathbf{A} = \mathbf{QR}$ by reflections (Theorem 11.8.4), applies $\mathbf{Q}^*$ to $\mathbf{b}$, and solves the triangular system $\mathbf{Rx} = \mathbf{Q}^* \mathbf{b}$ by back substitution, which Theorem 24.2.5 showed to be backward stable. Its cost is $2mn^2 - \frac{2}{3}n^3$ flops by a sum of the same kind as Proposition 24.2.3, which we do not write out: about twice the first when $m \gg n$.

The rule of thumb of Theorem 24.1.3 says: forward relative error $\leq$ condition number $\times$ backward relative error. The two routes solve *different problems*, and that is the whole story. The first solves a square linear system whose matrix is $\mathbf{A}^* \mathbf{A}$, of condition number $\kappa_2(\mathbf{A})^2$ by Theorem 16.8.4. The second solves the least-squares problem itself, of sensitivity given by [Theorem 24.3.1](#). Here is the comparison. Part (a) is Theorem 24.1.3 applied to the system $\mathbf{A}^* \mathbf{A} \mathbf{x} = \mathbf{A}^* \mathbf{b}$, except that here the right-hand side is perturbed as well, so we give the short direct argument instead of appealing to it.

Theorem 24.3.2. The Normal Equations Lose Twice the Digits

Let $\mathbf{A} \in M_{m \times n}(F)$ have rank $n \leq m$, let $\mathbf{b} \neq \mathbf{0}$, let $\mathbf{x}$ be the least-squares solution, and let $\theta < \pi/2$ be as in [Eq. 24.3.1](#). Work in the model of Definition 24.1.1, and suppose $mu \leq \frac{1}{2}$.

- (a) **(Normal equations.)** Suppose $\mathbf{C} = \text{fl}(\mathbf{A}^* \mathbf{A})$ and $\mathbf{d} = \text{fl}(\mathbf{A}^* \mathbf{b})$ are formed by inner products, and that $\mathbf{Cz} = \mathbf{d}$ is then solved by a method that is backward stable in the sense of Definition 24.1.3, so that the computed $\hat{\mathbf{x}}$ satisfies $(\mathbf{C} + \Delta \mathbf{C})\hat{\mathbf{x}} = \mathbf{d} + \Delta \mathbf{d}$ with $\|\Delta \mathbf{C}\| \leq c_s u \|\mathbf{C}\|$ and $\|\Delta \mathbf{d}\| \leq c_s u \|\mathbf{d}\|$. Then there is a constant c , depending only on m , n , and c_s and of size $O(mn)$, such that if $c u \kappa_2(\mathbf{A})^2 < 1$ then

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{c u \kappa_2(\mathbf{A})^2 (1 + \frac{1}{\cos \theta})}{1 - c u \kappa_2(\mathbf{A})^2}.$$

The bound carries $\kappa_2(\mathbf{A})^2$ **even when the residual is zero**.

- (b) **(A backward stable least-squares solver.)** Suppose instead that $\hat{\mathbf{x}}$ is the exact least-squares solution of $(\mathbf{A} + \Delta \mathbf{A})\mathbf{z} = \mathbf{b} + \Delta \mathbf{b}$ for some $\Delta \mathbf{A}, \Delta \mathbf{b}$ with $\|\Delta \mathbf{A}\| \leq c' u \|\mathbf{A}\|$ and $\|\Delta \mathbf{b}\| \leq c' u \|\mathbf{b}\|$. Then, to first order in u as in [Eq. 24.3.2](#),

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \lesssim c' u \left[\kappa_2(\mathbf{A}) \left(1 + \frac{1}{\cos \theta}\right) + \kappa_2(\mathbf{A})^2 \tan \theta \right].$$

Idea. For (a), the two sources of error — forming the product, and solving the system — are both perturbations of the *exact* normal equations, and both are bounded by a constant times $u\|\mathbf{A}\|^2$ or $u\|\mathbf{A}\|\|\mathbf{b}\|$. Subtract the exact equations from the perturbed ones and multiply by $(\mathbf{A}^*\mathbf{A})^{-1}$, whose norm is σ_n^{-2} : the two factors of σ_1/σ_n appear there and nowhere else. The one nuisance is that the estimate produces $\|\hat{\mathbf{x}}\|$ rather than $\|\mathbf{x}\|$; subadditivity of the norm (Corollary 11.1.6) trades one for the other at the cost of the denominator, exactly as in Theorem 16.8.3 (b). For (b) there is nothing to do: the hypothesis says the computed answer is the exact answer to perturbed data, which is what Theorem 24.3.1 measures.

Proof. (a). We argue over $\mathbb{R}$, the field in which Theorem 24.1.2 is stated; over $\mathbb{C}$ each complex multiplication and addition is a bounded number of real ones, and the same argument gives the same bound with a larger constant c . Write $\mathbf{a}_1, \dots, \mathbf{a}_n$ for the columns of $\mathbf{A}$, so that $(\mathbf{A}^*\mathbf{A})_{ij} = \mathbf{a}_i^*\mathbf{a}_j$ and $(\mathbf{A}^*\mathbf{b})_i = \mathbf{a}_i^*\mathbf{b}$, each an inner product of length m . By Theorem 24.1.2, $\text{fl}(\mathbf{a}_i^*\mathbf{a}_j) = (\mathbf{a}_i + \Delta\mathbf{a})^*\mathbf{a}_j$ with $|\Delta\mathbf{a}| \leq \gamma_m|\mathbf{a}_i|$ entrywise, so

$$|c_{ij} - \mathbf{a}_i^*\mathbf{a}_j| \leq \gamma_m|\mathbf{a}_i|^\top|\mathbf{a}_j| \leq \gamma_m\|\mathbf{a}_i\|\|\mathbf{a}_j\|,$$

the last step by Cauchy–Schwarz (Theorem 11.1.5). Summing squares and using $\|\mathbf{M}\| \leq \|\mathbf{M}\|_F$ — true because $\|\mathbf{M}\|_F^2 = \text{tr}(\mathbf{M}^*\mathbf{M})$ is the sum of the eigenvalues of $\mathbf{M}^*\mathbf{M}$, hence of the $\sigma_i(\mathbf{M})^2$, while $\|\mathbf{M}\| = \sigma_1(\mathbf{M})$ by Theorem 16.3.3 (c) —

$$\|\mathbf{C} - \mathbf{A}^*\mathbf{A}\| \leq \|\mathbf{C} - \mathbf{A}^*\mathbf{A}\|_F \leq \gamma_m \left(\sum_i \|\mathbf{a}_i\|^2 \right) = \gamma_m \|\mathbf{A}\|_F^2 \leq n\gamma_m \|\mathbf{A}\|^2.$$

The same argument on $\mathbf{d}$ gives $\|\mathbf{d} - \mathbf{A}^*\mathbf{b}\| \leq \gamma_m\|\mathbf{b}\|\|\mathbf{A}\|_F \leq \sqrt{n}\gamma_m\|\mathbf{A}\|\|\mathbf{b}\|$.

Now add the solver's backward error. Put $\Delta = (\mathbf{C} + \Delta\mathbf{C}) - \mathbf{A}^*\mathbf{A}$ and $\mathbf{h} = (\mathbf{d} + \Delta\mathbf{d}) - \mathbf{A}^*\mathbf{b}$. Since $\|\mathbf{C}\| \leq (1 + n\gamma_m)\|\mathbf{A}\|^2$ and $\|\mathbf{d}\| \leq (1 + \sqrt{n}\gamma_m)\|\mathbf{A}\|\|\mathbf{b}\|$,

$$\|\Delta\| \leq cu\|\mathbf{A}\|^2, \quad \|\mathbf{h}\| \leq cu\|\mathbf{A}\|\|\mathbf{b}\|,$$

with $c = 2nm + c_s(1 + 2nm)$ serving, say, since the hypothesis $mu \leq \frac{1}{2}$ gives $\gamma_m \leq 2mu$, and $\sqrt{n} \leq n$.

The exact solution satisfies $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ and the computed one satisfies $(\mathbf{A}^*\mathbf{A} + \Delta)\hat{\mathbf{x}} = \mathbf{A}^*\mathbf{b} + \mathbf{h}$. Subtracting,

$$\mathbf{A}^*\mathbf{A}(\hat{\mathbf{x}} - \mathbf{x}) = \mathbf{h} - \Delta\hat{\mathbf{x}}, \quad \|\hat{\mathbf{x}} - \mathbf{x}\| \leq \sigma_n^{-2}(cu\|\mathbf{A}\|\|\mathbf{b}\| + cu\|\mathbf{A}\|^2\|\hat{\mathbf{x}}\|),$$

using $\|(\mathbf{A}^*\mathbf{A})^{-1}\| = \sigma_n^{-2}$ from the proof of Theorem 24.3.1. Divide by $\|\mathbf{x}\|$, write $\rho = \|\hat{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\|$, and use $\sigma_1\|\mathbf{x}\| \geq \|\mathbf{b}\|\cos\theta$ together with $\|\hat{\mathbf{x}}\| \leq \|\mathbf{x}\|(1 + \rho)$:

$$\rho \leq cu\kappa_2(\mathbf{A})^2 \left(\frac{1}{\cos\theta} + 1 + \rho \right).$$

Since $cu\kappa_2(\mathbf{A})^2 < 1$ we may collect the ρ terms and divide, which gives the stated bound. Nothing in the argument used $\mathbf{r}$, so it holds for $\theta = 0$ as it stands.

(b). By hypothesis $\hat{\mathbf{x}} = \mathbf{x}(1)$ for the family of Theorem 24.3.1 with $\mathbf{E} = \Delta\mathbf{A}$ and $\mathbf{f} = \Delta\mathbf{b}$, and $\varepsilon_{\mathbf{A}}, \varepsilon_{\mathbf{b}} \leq c'u$. The stated bound is Eq. 24.3.2 for that family. This proves the theorem. ■

In digits. A relative error of η means about $\log_{10}(1/\eta)$ correct decimal digits, and an arithmetic with unit roundoff $u \approx 10^{-16}$ offers about 16 of them. Part (a) says the normal-equation route returns an answer with relative error about $u\kappa_2(\mathbf{A})^2$: it **loses about $2\log_{10}\kappa_2(\mathbf{A})$ digits**, whatever the residual. Part (b) says a backward stable least-squares solver returns an answer with relative error about $u\kappa_2(\mathbf{A})$ when the residual is small, and so **loses about $\log_{10}\kappa_2(\mathbf{A})$ digits**.

The comparison Chapter 16 §08 left open has two halves, and only one of them is settled here. The half that is proved is the *implication*: a backward stable least-squares solver loses about $\log_{10}\kappa_2(\mathbf{A})$ digits, while forming $\mathbf{A}^*\mathbf{A}$ loses about twice that — and part (a), which carries no stability hypothesis at all, is unconditional.

The half that is **not** proved is that the reflection route of Chapter 11 §08 supplies such a solver. That is the accumulated statement of the remark after [Theorem 24.3.3](#) below, which this book records with credit and does not prove; what [Theorem 24.3.3](#) itself proves is the backward stability of **one** reflection. A further qualification is the content of the next warning.

Warning. **The two-versus-one comparison is a small-residual statement, and the reason is not symmetric.** For a large residual the bound in (b) also carries a $\kappa_2(\mathbf{A})^2$, through the term $\kappa_2(\mathbf{A})^2 \tan \theta$. The difference is *where that square comes from*. In (b) it is the sensitivity of the problem, certified by [Theorem 24.3.1](#), and no method whatever can avoid it. In (a) it is manufactured by the algorithm: it is there when $\theta = 0$, where the problem's own sensitivity is only $\kappa_2(\mathbf{A})$. So the honest form of the comparison is: **a backward stable method loses what the problem costs; forming $\mathbf{A}^*\mathbf{A}$ loses $\kappa_2(\mathbf{A})$ times more than that whenever $\tan \theta \lesssim 1/\kappa_2(\mathbf{A})$, and never less.**

Example. A matrix where one route survives and the other does not

Let $\varepsilon > 0$ and

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2\varepsilon \\ -\varepsilon \end{pmatrix}.$$

Find $\kappa_2(\mathbf{A})$, the exact least-squares solution, and the residual. Then run both routes in the model with $u = 2^{-53}$, for $\varepsilon = 10^{-6}$ and for $\varepsilon = 10^{-9}$, and compare.

Solution. Exact data. $\mathbf{A}^\top \mathbf{A} = \begin{pmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{pmatrix}$, which sends $(1, 1)$ to $(2 + \varepsilon^2)(1, 1)$ and $(1, -1)$ to $\varepsilon^2(1, -1)$. So its eigenvalues are $2 + \varepsilon^2$ and ε^2 , the singular values of $\mathbf{A}$ are $\sqrt{2 + \varepsilon^2}$ and ε , and

$$\kappa_2(\mathbf{A}) = \frac{\sqrt{2 + \varepsilon^2}}{\varepsilon}.$$

Also $\mathbf{A}(2, -1) = (2 - 1, 2\varepsilon, -\varepsilon) = \mathbf{b}$, so the system is **consistent**: the least-squares solution is $\mathbf{x} = (2, -1)$ exactly, the residual is $\mathbf{r} = \mathbf{0}$, and $\theta = 0$. By [Theorem 24.3.1](#) the problem's sensitivity is exactly $2\kappa_2(\mathbf{A})$ — there is no κ_2^2 in the problem at all.

$\varepsilon = 10^{-6}$. Here $\kappa_2(\mathbf{A}) = 1.4142136 \cdot 10^6$, so $u\kappa_2(\mathbf{A}) = 1.6 \cdot 10^{-10}$ and $u\kappa_2(\mathbf{A})^2 = 2.2 \cdot 10^{-4}$. Forming the product rounds $1 + \varepsilon^2 = 1 + 10^{-12}$ to 1.0000000000010000889; the last four digits are the rounding, and they are a relative perturbation of about 10^{-4} in the *difference*

ε^2 that carries the small eigenvalue. Cholesky and the two triangular solves then return

$$\hat{\mathbf{x}} = (1.99983348134991, -0.99983348134991),$$

a relative error of $\|\hat{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\| = 1.05 \cdot 10^{-4}$ — about four correct digits out of sixteen, and within a factor of 2 of the bound $u\kappa_2(\mathbf{A})^2 = 2.2 \cdot 10^{-4}$. (Entry by entry the relative errors are $8.3 \cdot 10^{-5}$ and $1.7 \cdot 10^{-4}$; it is the 2-norm figure that the theorem bounds.) The reflection route returns $(2, -1)$, every digit correct.

$\varepsilon = 10^{-9}$. Now $\kappa_2(\mathbf{A}) = 1.4142136 \cdot 10^9$ and $u\kappa_2(\mathbf{A})^2 = 222$, so part (a) promises nothing at all — and nothing is what it delivers. Since $\varepsilon^2 = 10^{-18}$ is below u , the computed product is

$$\mathbf{C} = \text{fl}(\mathbf{A}^\top \mathbf{A}) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

a **singular** matrix: its Cholesky factorization (Theorem 13.2.3) fails at the second pivot, and the route cannot even be run. The reflection route again returns $(2, -1)$.

The bound is an upper bound. For $\varepsilon = 10^{-9}$ the reflection route is entitled to lose $\log_{10} \kappa_2(\mathbf{A}) \approx 9$ digits and in fact loses none: the answer is exactly representable and the arithmetic happens to reach it. Part (b) is a worst case over perturbations, as Theorem 16.8.3 was. Part (a) is an upper bound too, and it is not attained exactly either; what this example shows is that it is not a pessimistic one — the normal-equation route does lose its $2 \log_{10} \kappa_2(\mathbf{A})$ digits here, to within a factor of 2 in the error.

Warning. Forming $\mathbf{A}^* \mathbf{A}$ is not forbidden, it is limited. If $\kappa_2(\mathbf{A}) = 10^3$ then $u\kappa_2(\mathbf{A})^2 \approx 10^{-10}$, and the normal equations return ten correct digits for half the arithmetic. The rule is quantitative, not moral: the route is safe exactly while $u\kappa_2(\mathbf{A})^2$ is small enough for the purpose, and it collapses — as [Example \(A matrix where one route survives and the other does not\)](#) shows — once $\kappa_2(\mathbf{A})$ approaches $u^{-1/2} \approx 10^8$. What is never acceptable is forming the product without knowing which side of that threshold one is on.

24.3.3 Why the reflection route is the stable one

Part (b) above assumed a backward stable least-squares solver. That hypothesis has to be earned, and earning it is what Chapter 11 §08's reflections are for. The claim to be supported is: applying a Householder reflection by the formula

$$\mathbf{H}_{\mathbf{w}} \mathbf{v} = \mathbf{v} - \frac{2\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{w}\|^2} \mathbf{w} \quad (24.3.3)$$

of Chapter 11 §08 — **applying it, never forming the matrix** — commits a backward error of size u , with no dependence on the conditioning of anything. Here is that statement, proved.

Theorem 24.3.3. Backward Stability of One Reflection

Work in the model of Definition 24.1.1 over $\mathbb{R}$, and suppose $(10m + 16)u \leq \frac{1}{4}$. Let $\mathbf{w} \in \mathbb{R}^m$ be nonzero and let $\mathbf{H} = \mathbf{H}_{\mathbf{w}}$ be the reflection of Definition 11.8.1, and suppose $\mathbf{w}$ and the input are stored exactly.

- (a) For $\mathbf{v} \in \mathbb{R}^m$, let $\hat{\mathbf{y}}$ be the vector computed from $\mathbf{v}$ and $\mathbf{w}$ by Eq. 24.3.3. Then

$$\hat{\mathbf{y}} = \mathbf{H}(\mathbf{v} + \Delta\mathbf{v}), \quad \|\Delta\mathbf{v}\|_2 \leq (10m + 16)u \|\mathbf{v}\|_2.$$

- (b) For $\mathbf{A} \in M_{m \times n}(\mathbb{R})$, let $\hat{\mathbf{B}}$ be the matrix whose columns are computed from those of $\mathbf{A}$ in this way. Then

$$\hat{\mathbf{B}} = \mathbf{H}(\mathbf{A} + \Delta\mathbf{A}), \quad \|\Delta\mathbf{A}\|_F \leq (10m + 16)u \|\mathbf{A}\|_F.$$

Idea. Three quantities are computed before anything is subtracted: $\|\mathbf{w}\|^2$, the inner product $\langle \mathbf{v}, \mathbf{w} \rangle$, and their quotient β . Each is an inner product or a division, so Theorem 24.1.2 and Lemma 24.1.1 bound them at once, and the crude estimate $|\beta| \leq 2\|\mathbf{v}\|/\|\mathbf{w}\|$ — which is Cauchy–Schwarz and nothing more — keeps every error proportional to $\|\mathbf{v}\|$, not to $\|\mathbf{v}\|/\|\mathbf{w}\|$. Then the forward error $\|\hat{\mathbf{y}} - \mathbf{H}\mathbf{v}\|$ is converted into a backward error for free, because $\mathbf{H}$ is an isometry and its own inverse: $\Delta\mathbf{v} = \mathbf{H}(\hat{\mathbf{y}} - \mathbf{H}\mathbf{v})$ has exactly the same length. Part (b) is part (a) column by column, since the Frobenius norm adds the columns in squares.

Proof. (a). Write $\nu = \|\mathbf{w}\|_2$, $\beta = 2\langle \mathbf{v}, \mathbf{w} \rangle/\nu^2$, so that $\mathbf{H}\mathbf{v} = \mathbf{v} - \beta\mathbf{w}$, and abbreviate $\gamma_k = ku/(1 - ku)$. The hypothesis $(10m + 16)u \leq \frac{1}{4}$ gives in particular $(m + 2)u \leq \frac{1}{4}$, so $\gamma_k \leq \frac{4}{3}ku \leq \frac{1}{3}$ for every $k \leq m + 2$.

Step 1: the two inner products. The computed $s = \text{fl}(\mathbf{w}^*\mathbf{w})$ is a sum of m nonnegative terms w_i^2 , each carrying at most m rounding factors, so by Lemma 24.1.1 $s = \sum_i w_i^2(1 + \theta_i)$ with $|\theta_i| \leq \gamma_m$; as every $w_i^2 \geq 0$ this gives $s = \nu^2(1 + \theta)$ with $|\theta| \leq \gamma_m$. By Theorem 24.1.2 the computed $p = \text{fl}(\mathbf{v}^*\mathbf{w})$ satisfies $p = \langle \mathbf{v}, \mathbf{w} \rangle + \eta$ with

$$|\eta| \leq \gamma_m |\mathbf{v}|^\top |\mathbf{w}| \leq \gamma_m \|\mathbf{v}\| \nu,$$

by Cauchy–Schwarz (Theorem 11.1.5).

Step 2: the coefficient. The computed $\hat{\beta}$ is $(2p/s)$ times two rounding factors, so $\hat{\beta} = (2p/s)(1 + \theta_2)$ with $|\theta_2| \leq \gamma_2$, by Lemma 24.1.1. Writing $2p/s = (\beta + \zeta)/(1 + \theta)$ with $\zeta = 2\eta/\nu^2$ and $|\zeta| \leq 2\gamma_m \|\mathbf{v}\|/\nu$,

$$\left| \frac{1 + \theta_2}{1 + \theta} - 1 \right| \leq \frac{\gamma_m + \gamma_2}{1 - \gamma_m} \leq \frac{3}{2}(\gamma_m + \gamma_2) \leq 2(m + 2)u,$$

using $1 - \gamma_m \geq \frac{2}{3}$ and $\gamma_k \leq \frac{4}{3}ku$; in particular the factor itself has modulus at most 2. Since $|\beta| \leq 2\|\mathbf{v}\|/\nu$ by Cauchy–Schwarz,

$$|\hat{\beta} - \beta| \leq |\beta| \cdot 2(m + 2)u + 2|\zeta| \leq \frac{\|\mathbf{v}\|}{\nu} (4(m + 2)u + 6mu) \leq Ku \frac{\|\mathbf{v}\|}{\nu},$$

with $K = 10m + 8$. In particular $|\hat{\beta}| \leq 3\|\mathbf{v}\|/\nu$, since $Ku \leq (10m + 16)u \leq \frac{1}{4}$.

Step 3: the subtraction. Entry i is computed as $\hat{y}_i = (v_i - \hat{\beta}w_i(1 + \delta_i))(1 + \delta'_i)$ with $|\delta_i|, |\delta'_i| \leq u$. Subtracting $(\mathbf{H}\mathbf{v})_i = v_i - \beta w_i$ and regrouping,

$$\hat{y}_i - (\mathbf{H}\mathbf{v})_i = (v_i - \hat{\beta}w_i)\delta'_i + (\beta - \hat{\beta})w_i - \hat{\beta}w_i\delta_i(1 + \delta'_i).$$

Take 2-norms over i , one term at a time. The first is at most $u(\|\mathbf{v}\| + |\hat{\beta}|\nu) \leq 4u\|\mathbf{v}\|$; the second is $|\beta - \hat{\beta}|\nu \leq Ku\|\mathbf{v}\|$; the third is at most $|\hat{\beta}|\nu u(1 + u) \leq 4u\|\mathbf{v}\|$. Hence

$$\|\hat{\mathbf{y}} - \mathbf{H}\mathbf{v}\|_2 \leq (K + 8)u\|\mathbf{v}\|_2 = (10m + 16)u\|\mathbf{v}\|_2.$$

Step 4: forward to backward. Put $\Delta\mathbf{v} = \mathbf{H}(\hat{\mathbf{y}} - \mathbf{H}\mathbf{v})$. Then $\mathbf{H}(\mathbf{v} + \Delta\mathbf{v}) = \mathbf{H}\mathbf{v} + \mathbf{H}^2(\hat{\mathbf{y}} - \mathbf{H}\mathbf{v}) = \hat{\mathbf{y}}$ by Proposition 11.8.2 (b), and $\|\Delta\mathbf{v}\|_2 = \|\hat{\mathbf{y}} - \mathbf{H}\mathbf{v}\|_2$ because $\mathbf{H}$ is unitary (Proposition 11.8.2 (c)). This is (a).

(b). Apply (a) to each column $\mathbf{a}_j$, obtaining $\Delta\mathbf{a}_j$ with $\|\Delta\mathbf{a}_j\|_2 \leq (10m + 16)u\|\mathbf{a}_j\|_2$, and let $\Delta\mathbf{A}$ have these columns. Then $\hat{\mathbf{B}} = \mathbf{H}(\mathbf{A} + \Delta\mathbf{A})$ column by column, and squaring and summing over j gives $\|\Delta\mathbf{A}\|_F \leq (10m + 16)u\|\mathbf{A}\|_F$. This proves the theorem. ■

The constants are crude on purpose; what matters is their shape. **The bound is a multiple of u and of m , and of nothing else.** No condition number appears, and none can: a reflection is an isometry, so it neither stretches nor shrinks, and there is no small quantity anywhere for a rounding error to be measured against.

Remark. **What the accumulation says, and that we do not prove it.** Running $\min(n, m - 1)$ reflections as in Theorem 11.8.4, one obtains computed $\hat{\mathbf{R}}$ and stored reflection vectors. The classical theorem — due to Wilkinson, and proved by inducting the argument above over the reflections while tracking how the perturbations compose — states that there is an **exactly** unitary $\tilde{\mathbf{Q}} \in M_m(\mathbb{R})$, not in general the product of the computed reflections, with

$$\tilde{\mathbf{Q}}\hat{\mathbf{R}} = \mathbf{A} + \Delta\mathbf{A}, \quad \|\Delta\mathbf{A}\|_F \leq c(m, n)u\|\mathbf{A}\|_F,$$

and that the resulting least-squares solution is the exact solution of a problem with data perturbed by $O(u)$ — the hypothesis of Theorem 24.3.2 (b). **This accumulated statement is not proved here**, and nothing in this book depends on it: every theorem above either proves its own case or carries backward stability as a stated hypothesis. The induction is not deep, but it is long, and the honest thing is to say so.

24.3.4 What goes wrong with Gram–Schmidt

Chapter 11 §02 warned that “the classical process is numerically fragile” and promised that this chapter would say why the reflection route is the stable one. Half of the answer is Theorem 24.3.3. The other half is what the Gram–Schmidt recursion of Theorem 11.2.2 does instead.

The recursion subtracts from $\mathbf{a}_k$ its components along the vectors already produced. When the columns are nearly dependent, what is left is a small difference of large quantities, and the leading digits cancel, exactly as in the subtraction of nearly equal numbers examined in Section 1. The computed $\mathbf{q}_k$ is then a unit vector in a noticeably wrong direction, and, worse, the error is not random: it is inherited by every later step. The measured consequence is a loss of *orthogonality*: the computed $\hat{\mathbf{Q}}$ satisfies $\hat{\mathbf{Q}}^\top \hat{\mathbf{Q}} \neq \mathbf{I}$ by an amount that grows with $\kappa_2(\mathbf{A})$. The classical analyses, which we state with credit and **do not prove here**, say how fast:

- for the **classical** process, $\|\hat{\mathbf{Q}}^\top \hat{\mathbf{Q}} - \mathbf{I}\|$ is of order $u\kappa_2(\mathbf{A})^2$ (Kielbasinski; Giraud, Langou, and Rozloznik);
- for the **modified** process, in which each new vector is projected against the $\mathbf{q}_i$ one at a

time as it is formed rather than all at once from the original $\mathbf{a}_k$, it is of order $u \kappa_2(\mathbf{A})$ (Bjorck);

- for the reflection route it is of order u , independently of $\kappa_2(\mathbf{A})$, which is the content of [Theorem 24.3.3](#).

Example. Three columns and three answers

Let $\varepsilon = 10^{-6}$ and

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \end{pmatrix}.$$

Show that $\kappa_2(\mathbf{A}) = 3/\varepsilon + O(\varepsilon)$, then run the classical and the modified recursion in the model with $u = 2^{-53}$ and compare the inner products of the computed vectors.

Solution. The condition number. $\mathbf{A}^\top \mathbf{A} = \begin{pmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$. Its characteristic polynomial factors as

$$\det(t\mathbf{I} - \mathbf{A}^\top \mathbf{A}) = (t - \varepsilon^2)(t^2 - (3 + \varepsilon^2)t + \varepsilon^2),$$

as one checks by expanding both sides. The quadratic has roots $t_\pm$ with $t_+ t_- = \varepsilon^2$ and $t_+ + t_- = 3 + \varepsilon^2$, so $t_+ = 3 + O(\varepsilon^2)$ and $t_- = \varepsilon^2/3 + O(\varepsilon^4)$. Hence $\sigma_1 = \sqrt{3} + O(\varepsilon^2)$ and $\sigma_3 = \varepsilon/\sqrt{3} + O(\varepsilon^3)$, and

$$\kappa_2(\mathbf{A}) = \frac{\sqrt{3}}{\varepsilon/\sqrt{3}} + O(\varepsilon) = \frac{3}{\varepsilon} + O(\varepsilon) = 3.000000 \cdot 10^6.$$

So $u\kappa_2(\mathbf{A}) = 3.3 \cdot 10^{-10}$ and $u\kappa_2(\mathbf{A})^2 = 1.0 \cdot 10^{-3}$.

The computed vectors. Both recursions produce the same $\mathbf{q}_1$ and $\mathbf{q}_2$, with $\langle \mathbf{q}_1, \mathbf{q}_2 \rangle = 6.3 \cdot 10^{-11}$: that is the unavoidable $u\kappa_2(\mathbf{A})$, and it is already five to six digits worse than u , being larger by a factor of $5.7 \cdot 10^5$. The two differ at the third step, which is where the cancellation bites:

$$\text{classical: } \langle \mathbf{q}_2, \mathbf{q}_3 \rangle = -8.89 \cdot 10^{-5},$$

$$\text{modified: } \langle \mathbf{q}_2, \mathbf{q}_3 \rangle = 0.$$

The classical figure is within a factor of 12 of $u\kappa_2(\mathbf{A})^2 = 1.0 \cdot 10^{-3}$; the modified figure is at the level of $u\kappa_2(\mathbf{A})$ or below. Two vectors at inner product $9 \cdot 10^{-5}$ are at about 89.995 degrees to each other, and a basis that is orthonormal only to five digits is worth five digits in everything computed from it.

24.3.5 The sign in the reflection, quantified

Chapter 11 §08 chose $\mathbf{w} = \mathbf{x} + \alpha\|\mathbf{x}\|\mathbf{e}_1$ rather than $\mathbf{w}' = \mathbf{x} - \alpha\|\mathbf{x}\|\mathbf{e}_1$, where α is the phase of x_1 , and said: “Chapter 24 quantifies the loss; the summary here is that both signs are correct mathematics and only one is correct arithmetic.” Here is the quantity.

Take $\mathbf{x} \in \mathbb{R}^m$ nonzero with $x_1 > 0$, so $\alpha = 1$, and write $\tilde{\mathbf{x}} = (x_2, \dots, x_m)$. Expanding as in Theorem 11.8.3,

$$\|\mathbf{w}\|^2 = 2\|\mathbf{x}\|(\|\mathbf{x}\| + x_1), \quad \|\mathbf{w}'\|^2 = 2\|\mathbf{x}\|(\|\mathbf{x}\| - x_1),$$

and $(\|\mathbf{x}\| + x_1)(\|\mathbf{x}\| - x_1) = \|\tilde{\mathbf{x}}\|^2$. Both vectors are built by the same two operations: compute $\|\mathbf{x}\|$, then add it to or subtract it from x_1 . The computed norm carries an error of absolute size about $u\|\mathbf{x}\|$, and every other entry is copied from $\mathbf{x}$ exactly. So in both cases the computed vector is $\mathbf{w} + \Delta\mathbf{w}$ with $\|\Delta\mathbf{w}\| \approx u\|\mathbf{x}\|$ — *the same absolute error*. What differs is what that error is measured against, because $\mathbf{H}_{\mathbf{w}}$ depends on $\mathbf{w}$ only through the line it spans. Writing $\mathbf{p} = \mathbf{w}/\|\mathbf{w}\|$ and $\mathbf{q} = (\mathbf{w} + \Delta\mathbf{w})/\|\mathbf{w} + \Delta\mathbf{w}\|$, Definition 11.8.1 gives $\mathbf{H}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}+\Delta\mathbf{w}} = -2(\mathbf{p}\mathbf{p}^* - \mathbf{q}\mathbf{q}^*)$, and

$$\|\mathbf{p}\mathbf{p}^* - \mathbf{q}\mathbf{q}^*\| \leq \|\mathbf{p}(\mathbf{p} - \mathbf{q})^*\| + \|(\mathbf{p} - \mathbf{q})\mathbf{q}^*\| = 2\|\mathbf{p} - \mathbf{q}\| \leq \frac{8\|\Delta\mathbf{w}\|}{\|\mathbf{w}\|}.$$

For the last step, abbreviate $\mathbf{s} = \Delta\mathbf{w}$ and split

$$\mathbf{p} - \mathbf{q} = \mathbf{w} \left(\frac{1}{\|\mathbf{w}\|} - \frac{1}{\|\mathbf{w} + \mathbf{s}\|} \right) - \frac{\mathbf{s}}{\|\mathbf{w} + \mathbf{s}\|}.$$

The second piece has norm $\|\mathbf{s}\|/\|\mathbf{w} + \mathbf{s}\|$, and the first has norm $|\|\mathbf{w} + \mathbf{s}\| - \|\mathbf{w}\||/\|\mathbf{w} + \mathbf{s}\|$, which is at most the same quantity by the reverse triangle inequality (Lemma 16.1.1); and $\|\mathbf{w} + \mathbf{s}\| \geq \|\mathbf{w}\|/2$ as soon as $\|\mathbf{s}\| \leq \|\mathbf{w}\|/2$, which is certainly the case here, $\|\mathbf{s}\|$ being of size $u\|\mathbf{x}\|$. So the computed reflection differs from the intended one by a small multiple of

$$u \frac{\|\mathbf{x}\|}{\|\mathbf{w}\|}, \quad \text{where} \quad \frac{\|\mathbf{x}\|}{\|\mathbf{w}\|} = \sqrt{\frac{\|\mathbf{x}\|}{2(\|\mathbf{x}\| + x_1)}} \leq \frac{1}{\sqrt{2}}.$$

The safe sign therefore never amplifies: its reflection is accurate to a small multiple of u , no matter what $\mathbf{x}$ is. The other sign amplifies by the factor

$$\frac{\|\mathbf{w}\|}{\|\mathbf{w}'\|} = \sqrt{\frac{\|\mathbf{x}\| + x_1}{\|\mathbf{x}\| - x_1}} = \frac{\|\mathbf{x}\| + x_1}{\|\tilde{\mathbf{x}}\|},$$

which is **unbounded**: it blows up precisely as $\mathbf{x}$ approaches a positive multiple of $\mathbf{e}_1$, where $\|\tilde{\mathbf{x}}\| \rightarrow 0$ and $\mathbf{w}' \rightarrow \mathbf{0}$.

Example. The same vector, the two signs

Let $\varepsilon = 10^{-8}$ and $\mathbf{x} = (1, \varepsilon, \varepsilon) \in \mathbb{R}^3$. Predict the amplification factor for each sign, then apply each computed reflection to $\mathbf{x}$ in the model with $u = 2^{-53}$ and measure how flat the answer is.

Solution. Prediction. $\|\tilde{\mathbf{x}}\| = \sqrt{2}\varepsilon = 1.41 \cdot 10^{-8}$ and $\|\mathbf{x}\| = \sqrt{1 + 2\varepsilon^2}$, so $\|\mathbf{x}\|/\|\mathbf{w}\| = 0.5$ for the safe sign, while

$$\frac{\|\mathbf{x}\|}{\|\mathbf{w}'\|} = \sqrt{\frac{\|\mathbf{x}\|}{2(\|\mathbf{x}\| - 1)}} = 7.07 \cdot 10^7, \quad u \frac{\|\mathbf{x}\|}{\|\mathbf{w}'\|} = 7.9 \cdot 10^{-9}.$$

The arithmetic. The computed $\|\mathbf{x}\|$ is exactly 1: the true value exceeds 1 by $\varepsilon^2 = 10^{-16}$, which is below the unit roundoff, so the addition returns 1. Hence

$$\hat{\mathbf{w}} = (2, \varepsilon, \varepsilon), \quad \hat{\mathbf{w}}' = (0, \varepsilon, \varepsilon).$$

The first is right to full relative accuracy in every entry. The second has lost its first entry

altogether: it should be -10^{-16} and it is 0. Applying [Eq. 24.3.3](#),

$$\mathbf{H}_{\hat{\mathbf{w}}} \mathbf{x} = (-1, 0, 0), \quad \mathbf{H}_{\hat{\mathbf{w}}'} \mathbf{x} = (1, -10^{-8}, -10^{-8}).$$

The safe sign flattens $\mathbf{x}$ exactly: the two entries that were supposed to be annihilated are true zeros. The other sign leaves behind entries of size 10^{-8} beside a vector of norm 1 — a failure of $1.4 \cdot 10^{-8}$ where 10^{-16} was available, and within a factor of 2 of the estimate $u\|\mathbf{x}\|/\|\mathbf{w}'\| = 7.9 \cdot 10^{-9}$. Eight digits, thrown away by a sign.

For $\varepsilon = 10^{-6}$ and $\varepsilon = 10^{-4}$ the same computation leaves residues of $1.3 \cdot 10^{-10}$ and $1.5 \cdot 10^{-13}$, against estimates $7.9 \cdot 10^{-11}$ and $7.9 \cdot 10^{-13}$. The estimates fall by exactly 100 each time, as $1/\|\tilde{\mathbf{x}}\|$ does; the residues fall by 112 and then by 825, because a residue depends on the actual rounding error committed in $\|\mathbf{x}\|$, which varies with the binary expansion of $\|\mathbf{x}\|$ and not only with its size. Over four orders of magnitude the estimate stays within a factor of 6 of the measurement, which is all an order-of-magnitude estimate promises.

Nonexample by minimal change. Change one character of [Example \(The same vector, the two signs\)](#): take $\mathbf{x} = (-1, \varepsilon, \varepsilon)$ with $\varepsilon = 10^{-8}$. The arithmetic is unaltered — the computed $\|\mathbf{x}\|$ is again exactly 1 — and the rule of Theorem 11.8.3 still works: the phase of x_1 is now $\alpha = -1$, so $\mathbf{w} = \mathbf{x} + \alpha\|\mathbf{x}\|\mathbf{e}_1 = (-2, \varepsilon, \varepsilon)$, again with $\|\mathbf{x}\|/\|\mathbf{w}\| = 0.5$ and again flattening $\mathbf{x}$ to $(1, 0, 0)$ exactly. What fails is the rule with the phase dropped, $\mathbf{w} = \mathbf{x} + \|\mathbf{x}\|\mathbf{e}_1$: it returns $(0, \varepsilon, \varepsilon)$, and that reflection leaves $\mathbf{x}$ with the same residue $1.4 \cdot 10^{-8}$ as before. So the clause carrying the weight is not the plus sign but the **phase** α : what the rule requires is that $\alpha\|\mathbf{x}\|$ be added to x_1 so that the first entry grows in modulus, never so that it cancels. Reading the plus sign literally is correct mathematics for every $\mathbf{x}$ and correct arithmetic only for $x_1 > 0$.

Warning. **The bad sign is not merely inaccurate; on one input it does not exist.** If $\mathbf{x} = c\mathbf{e}_1$ with $c > 0$, then $\mathbf{w}' = \mathbf{0}$ and Definition 11.8.1 does not apply at all. That is not an accident of the arithmetic but the limit of the cancellation: the quantity whose leading digits are being destroyed is $\|\mathbf{x}\| - x_1$, and on that input it is zero. An algorithm that is undefined at a point is inaccurate near it, and the sign rule of Theorem 11.8.3 is exactly the rule that keeps the point out of reach.

24.3.6 Exercises

24.3.6.1 A. Check your understanding

Exercise. A1

- Define θ for a least-squares problem $\mathbf{Az} \approx \mathbf{b}$ with $\text{rank } \mathbf{A} = n$, and say what $\theta = 0$ means about the system.
- State [Theorem 24.3.1](#), and say which of its two terms survives when the residual is zero.
- Explain in one sentence why $\kappa_2(\mathbf{A})^2$ appears in the normal-equation route even for a consistent system, and why it need not appear in the reflection route.
- True or false: “a matrix with $\kappa_2(\mathbf{A}) = 10^3$ should never be handled by the normal equations”. Justify your answer.
- What quantity does [Theorem 24.3.3](#) bound, and which quantity does it **not** depend

on?

Solution.

- (a) By Eq. 24.3.1, $\theta \in [0, \pi/2]$ is determined by $\sin \theta = \|\mathbf{r}\|/\|\mathbf{b}\|$ with $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ the residual, equivalently $\cos \theta = \|\mathbf{A}\mathbf{x}\|/\|\mathbf{b}\|$; it is the angle between $\mathbf{b}$ and $\text{col}(\mathbf{A})$. Then $\theta = 0$ says $\mathbf{r} = \mathbf{0}$, that is, the system $\mathbf{A}\mathbf{z} = \mathbf{b}$ is consistent.
- (b) See Theorem 24.3.1: the relative sensitivity is at most $\kappa_2(\mathbf{A})(\varepsilon_{\mathbf{A}} + \varepsilon_{\mathbf{b}}/\cos \theta) + \kappa_2(\mathbf{A})^2 \varepsilon_{\mathbf{A}} \tan \theta$. With $\mathbf{r} = \mathbf{0}$ we get $\tan \theta = 0$, and only the $\kappa_2(\mathbf{A})$ term survives, with $\cos \theta = 1$.
- (c) The normal-equation route solves a *different* linear system, the one with matrix $\mathbf{A}^*\mathbf{A}$, whose condition number is $\kappa_2(\mathbf{A})^2$ by Theorem 16.8.4, and the rounding committed in forming that matrix is magnified by its condition number, not by $\mathbf{A}$'s; the reflection route never forms the product, and $\kappa_2(\mathbf{R}) = \kappa_2(\mathbf{A})$.
- (d) False. By Theorem 24.3.2 (a) the loss is about $2 \log_{10} \kappa_2(\mathbf{A}) = 6$ digits out of sixteen, which is usually ample; the route costs about half the arithmetic. The warning after Example (A matrix where one route survives and the other does not) states the rule quantitatively: what matters is whether $u\kappa_2(\mathbf{A})^2$ is small enough for the purpose.
- (e) It bounds the **backward** error of applying one reflection by the formula Eq. 24.3.3: the computed output is the exact output for an input perturbed by at most $(10m + 16)u$ relatively. It does not depend on any condition number, only on m and u .

24.3.6.2 B. Practice

Exercise. B1: Reading the bound

Let $\mathbf{A} \in M_{50 \times 3}(\mathbb{R})$ have $\kappa_2(\mathbf{A}) = 10^5$, and suppose the data are accurate to $\varepsilon = 10^{-10}$ relatively. Using Eq. 24.3.2 with $u \approx 10^{-16}$ where needed:

- (a) Bound the relative error of the exact least-squares solution caused by the data error, when the residual is zero.
- (b) Do the same when $\tan \theta = 1$.
- (c) In case (b), how many digits does the normal-equation route lose on top of the data error, and how many does a backward stable method lose?

Solution.

- (a) With $\theta = 0$, Eq. 24.3.2 gives $2\kappa_2(\mathbf{A})\varepsilon = 2 \cdot 10^5 \cdot 10^{-10} = 2 \cdot 10^{-5}$: about five correct digits, and no algorithm can restore them.
- (b) With $\tan \theta = 1$ we have $\cos \theta = 1/\sqrt{2}$, so the bracket is $\kappa_2(1 + \sqrt{2}) + \kappa_2^2 = 2.4 \cdot 10^5 + 10^{10} \approx 10^{10}$, and the bound is $10^{10} \cdot 10^{-10} = 1$. The data alone determine nothing: the $\kappa_2^2 \tan \theta$ term has swallowed the problem.
- (c) The arithmetic contributes $u\kappa_2(\mathbf{A})^2 = 10^{-16} \cdot 10^{10} = 10^{-6}$ for the normal equations and $u(\kappa_2 + \kappa_2^2 \tan \theta) \approx 10^{-6}$ for a backward stable method as well, by Theorem 24.3.2. In this case the two lose the same six digits — because the κ_2^2 is the problem's, not the algorithm's. Both are far below the 10^0 already lost to the data, so the choice of method is irrelevant here; it is for $\tan \theta$ small that it decides everything.

Exercise. B2: Where the product goes singular

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$ as in [Example \(A matrix where one route survives and the other does not\)](#), in the model with $u = 2^{-53}$.

- For which ε does $\text{fl}(1 + \varepsilon^2) = 1$ hold, and what is $\text{fl}(\mathbf{A}^\top \mathbf{A})$ then?
- Express the threshold in (a) in terms of $\kappa_2(\mathbf{A})$ and u .
- Explain why the reflection route is untroubled at that threshold.

Solution.

- By Definition 24.1.1, $\text{fl}(1 + \delta) = 1$ exactly when δ is at most the largest quantity that rounds away beside 1, which is $u = 2^{-53} \approx 1.11 \cdot 10^{-16}$. So $\text{fl}(1 + \varepsilon^2) = 1$ for $\varepsilon^2 \leq u$, that is $\varepsilon \lesssim 1.05 \cdot 10^{-8}$. Then every entry of $\text{fl}(\mathbf{A}^\top \mathbf{A})$ is 1 and the computed matrix is $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, which is singular: the information distinguishing the two columns has been rounded away.
- From [Example \(A matrix where one route survives and the other does not\)](#), $\kappa_2(\mathbf{A}) = \sqrt{2 + \varepsilon^2}/\varepsilon \approx \sqrt{2}/\varepsilon$, so $\varepsilon^2 \leq u$ says $\kappa_2(\mathbf{A})^2 \geq 2/u$, that is $u\kappa_2(\mathbf{A})^2 \geq 2$. The normal equations collapse exactly when the bound of [Theorem 24.3.2](#) (a) exceeds 1, which is the same as saying that the route promises no correct digit.
- The reflection route never computes $1 + \varepsilon^2$. It works with $\mathbf{A}$ itself, whose entries record ε and not ε^2 , and by [Theorem 24.3.3](#) each reflection perturbs the data relatively by $O(u)$. A relative perturbation of 10^{-16} in $\varepsilon = 10^{-9}$ leaves ε intact; a relative perturbation of 10^{-16} in 1 destroys $\varepsilon^2 = 10^{-18}$. Squaring is what moves the small quantity below the rounding level.

Exercise. B3: One reflection, by hand

Let $\mathbf{x} = (3, 0, 4) \in \mathbb{R}^3$.

- Compute $\mathbf{w}$ and $\mathbf{w}'$ as in Theorem 11.8.3, and the two amplification factors $\|\mathbf{x}\|/\|\mathbf{w}\|$ and $\|\mathbf{x}\|/\|\mathbf{w}'\|$.
- Verify the identity $\|\mathbf{w}\|^2\|\mathbf{w}'\|^2 = 4\|\mathbf{x}\|^2\|\tilde{\mathbf{x}}\|^2$ on this example.
- Is the sign choice important for this $\mathbf{x}$? Justify your answer.

Solution.

- $\|\mathbf{x}\| = 5$ and $x_1 = 3 > 0$, so $\alpha = 1$, $\mathbf{w} = (8, 0, 4)$, and $\mathbf{w}' = (-2, 0, 4)$. Then $\|\mathbf{w}\|^2 = 80$ and $\|\mathbf{w}'\|^2 = 20$, so

$$\frac{\|\mathbf{x}\|}{\|\mathbf{w}\|} = \frac{5}{\sqrt{80}} = \frac{\sqrt{5}}{4} \approx 0.559, \quad \frac{\|\mathbf{x}\|}{\|\mathbf{w}'\|} = \frac{5}{\sqrt{20}} = \frac{\sqrt{5}}{2} \approx 1.118.$$

Both are of order 1. (The general formulas agree: $\|\mathbf{w}\|^2 = 2 \cdot 5 \cdot (5 + 3) = 80$ and $\|\mathbf{w}'\|^2 = 2 \cdot 5 \cdot (5 - 3) = 20$.)

- $\tilde{\mathbf{x}} = (0, 4)$ has $\|\tilde{\mathbf{x}}\|^2 = 16$, and $4\|\mathbf{x}\|^2\|\tilde{\mathbf{x}}\|^2 = 4 \cdot 25 \cdot 16 = 1600 = 80 \cdot 20$. The identity is $\|\mathbf{w}\|^2\|\mathbf{w}'\|^2 = 4\|\mathbf{x}\|^2(\|\mathbf{x}\| + x_1)(\|\mathbf{x}\| - x_1) = 4\|\mathbf{x}\|^2\|\tilde{\mathbf{x}}\|^2$.

- (c) No. The amplification factor of the bad sign is $(\|\mathbf{x}\| + x_1)/\|\tilde{\mathbf{x}}\| = 8/4 = 2$, so the two choices differ by a factor of 2 in accuracy, which is nothing. The sign matters only when $\mathbf{x}$ is close to a multiple of $\mathbf{e}_1$, that is when $\|\tilde{\mathbf{x}}\| \ll \|\mathbf{x}\|$; here $\|\tilde{\mathbf{x}}\| = 4$ is comparable with $\|\mathbf{x}\| = 5$.

24.3.6.3 C. Going deeper

Exercise. C1: The squared term is attained

Let $\varepsilon, \delta > 0$ and take

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \\ 0 & 0 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ \delta \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{f} = \mathbf{0}.$$

- Compute $\kappa_2(\mathbf{A})$, the least-squares solution $\mathbf{x}$, the residual $\mathbf{r}$, and $\tan \theta$.
- Using (*) in the proof of [Theorem 24.3.1](#), compute $\mathbf{x}'(0)$ exactly.
- Hence show that the squared term of [Theorem 24.3.1](#) is attained, in the sense that $\|\mathbf{x}'(0)\|/\|\mathbf{x}\|$ equals $\kappa_2(\mathbf{A})^2 \varepsilon_{\mathbf{A}} \tan \theta$ exactly, and explain what this says about any algorithm for this problem.

Solution.

- $\mathbf{A}^\top \mathbf{A} = \text{diag}(1, \varepsilon^2)$, so the singular values are 1 and ε and $\kappa_2(\mathbf{A}) = 1/\varepsilon$ by the rectangular clause of Definition 16.8.1. The normal equations $\text{diag}(1, \varepsilon^2)\mathbf{x} = \mathbf{A}^\top \mathbf{b} = (1, 0)$ give $\mathbf{x} = (1, 0)$. Then $\mathbf{Ax} = (1, 0, 0)$, so $\mathbf{r} = \mathbf{b} - \mathbf{Ax} = (0, 0, \delta)$ and, by [Eq. 24.3.1](#),

$$\tan \theta = \frac{\|\mathbf{r}\|}{\|\mathbf{Ax}\|} = \delta.$$

Also $\|\mathbf{E}\| = 1 = \|\mathbf{A}\|$, so $\varepsilon_{\mathbf{A}} = 1$, and $\varepsilon_{\mathbf{b}} = 0$.

- The three terms of (*). First, $\mathbf{E}^\top \mathbf{r} = \mathbf{e}_2(\mathbf{e}_3^\top \mathbf{r}) = \delta \mathbf{e}_2$, and $(\mathbf{A}^\top \mathbf{A})^{-1} = \text{diag}(1, \varepsilon^{-2})$, so the first term is $(\delta/\varepsilon^2)\mathbf{e}_2$. Second, $\mathbf{Ex} = \mathbf{e}_3(\mathbf{e}_2^\top \mathbf{x}) = \mathbf{0}$ because $x_2 = 0$, so the second term vanishes. Third, $\mathbf{f} = \mathbf{0}$. Hence

$$\mathbf{x}'(0) = \left(0, \frac{\delta}{\varepsilon^2}\right).$$

- $\|\mathbf{x}\| = 1$ and $\|\mathbf{x}'(0)\| = \delta/\varepsilon^2$, while

$$\kappa_2(\mathbf{A})^2 \varepsilon_{\mathbf{A}} \tan \theta = \frac{1}{\varepsilon^2} \cdot 1 \cdot \delta = \frac{\delta}{\varepsilon^2}.$$

The two agree exactly: the squared term is **attained**. The bound as a whole is not, and the reason is that the other two terms are nonnegative — here the first of them contributes a further $\kappa_2(\mathbf{A})\varepsilon_{\mathbf{A}} = 1/\varepsilon$, so the bound exceeds $\|\mathbf{x}'(0)\|/\|\mathbf{x}\|$ by exactly that much. It is the squared term that matters, and what it says is that the sensitivity is a property of the data $(\mathbf{A}, \mathbf{b})$ alone: a perturbation of $\mathbf{A}$ of relative size $\varepsilon_{\mathbf{A}}$ really does move the exact answer by $\kappa_2(\mathbf{A})^2 \tan \theta \varepsilon_{\mathbf{A}}$ relatively, so **no** algorithm, however stable, can return an answer better than that from data known only to within $\varepsilon_{\mathbf{A}}$. Compare [Theorem 24.3.2](#) (b): a backward stable method achieves this and no more.

Exercise. C2: Why the residual has to appear

Let $\mathbf{A} \in M_{m \times n}(F)$ have rank n , and let $\mathbf{x}^+ = \mathbf{A}^+ \mathbf{b}$ be the least-squares solution, where $\mathbf{A}^+$ is the pseudoinverse of Definition 13.12.1.

- Prove that $\mathbf{A}^+ = (\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^*$ when $\text{rank } \mathbf{A} = n$, and deduce $\|\mathbf{A}^+\| = 1/\sigma_n(\mathbf{A})$.
- Deduce from (*) in the proof of [Theorem 24.3.1](#) that if $\mathbf{E} = \mathbf{O}$ — only $\mathbf{b}$ is perturbed — then no $\kappa_2(\mathbf{A})^2$ term can appear, whatever the residual.
- Explain, in one or two sentences, why perturbing $\mathbf{A}$ is different: which term of (*) carries the residual, and why it cannot be present when $\mathbf{E} = \mathbf{O}$.

Solution.

- Write $\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*$ (Theorem 13.9.1). Since $\text{rank } \mathbf{A} = n$, $\sigma_1, \dots, \sigma_n > 0$, so $\mathbf{\Sigma}^+$ of Definition 13.12.1 has σ_i^{-1} on its diagonal and $\mathbf{A}^+ = \mathbf{V} \mathbf{\Sigma}^+ \mathbf{U}^*$. On the other hand $\mathbf{A}^* \mathbf{A} = \mathbf{V} (\mathbf{\Sigma}^* \mathbf{\Sigma}) \mathbf{V}^* = \mathbf{V} \text{diag}(\sigma_i^2) \mathbf{V}^*$, so

$$(\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* = \mathbf{V} \text{diag}(\sigma_i^{-2}) \mathbf{V}^* \mathbf{V} \mathbf{\Sigma}^+ \mathbf{U}^* = \mathbf{V} (\text{diag}(\sigma_i^{-2}) \mathbf{\Sigma}^+) \mathbf{U}^* = \mathbf{V} \mathbf{\Sigma}^+ \mathbf{U}^* = \mathbf{A}^+,$$

because $\text{diag}(\sigma_i^{-2}) \mathbf{\Sigma}^+$ is the $n \times m$ matrix with entries $\sigma_i^{-2} \sigma_i = \sigma_i^{-1}$ on the diagonal and zeros elsewhere. Unitary factors leave singular values alone — $(\mathbf{V} \mathbf{\Sigma}^+ \mathbf{U}^*)^* (\mathbf{V} \mathbf{\Sigma}^+ \mathbf{U}^*) = \mathbf{U} (\mathbf{\Sigma}^{++} \mathbf{\Sigma}^+) \mathbf{U}^*$ is unitarily similar to $\mathbf{\Sigma}^{++} \mathbf{\Sigma}^+$, so the two have the same eigenvalues and Definition 13.9.1 reads the same singular values off them — so $\|\mathbf{A}^+\| = \|\mathbf{\Sigma}^+\| = \max_i \sigma_i^{-1} = 1/\sigma_n$ by Theorem 16.3.3 (c).

- With $\mathbf{E} = \mathbf{O}$, (*) reduces to $\mathbf{x}'(0) = \mathbf{A}^+ \mathbf{f}$, so by (a)

$$\frac{\|\mathbf{x}'(0)\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{f}\|}{\sigma_n \|\mathbf{x}\|} = \kappa_2(\mathbf{A}) \varepsilon_{\mathbf{b}} \frac{\|\mathbf{b}\|}{\sigma_1 \|\mathbf{x}\|} \leq \frac{\kappa_2(\mathbf{A}) \varepsilon_{\mathbf{b}}}{\cos \theta},$$

a single factor of $\kappa_2(\mathbf{A})$ however large $\|\mathbf{r}\|$ is. (The residual enters only through $1/\cos \theta$, which is bounded as long as θ is bounded away from $\pi/2$.)

- The residual sits in the term $(\mathbf{A}^* \mathbf{A})^{-1} \mathbf{E}^* \mathbf{r}$, and that is the only term with $(\mathbf{A}^* \mathbf{A})^{-1}$ not paired with an $\mathbf{A}^*$ — which is exactly why it carries σ_n^{-2} rather than σ_n^{-1} . It is proportional to $\mathbf{E}$, so it vanishes when $\mathbf{E} = \mathbf{O}$. Perturbing $\mathbf{b}$ tilts the target inside a fixed column space; perturbing $\mathbf{A}$ tilts the column space itself, and the residual is the lever arm by which that tilt moves the answer.

Exercise. C3: A reflection applied to a matrix

Let $\mathbf{H} = \mathbf{H}_{\mathbf{w}}$ be a Householder reflection in $\mathbb{R}^m$ and let $\mathbf{A} \in M_{m \times n}(\mathbb{R})$.

- Using [Theorem 24.3.3](#) (b) and $\|\mathbf{M}\| \leq \|\mathbf{M}\|_F \leq \sqrt{n} \|\mathbf{M}\|$ for $\mathbf{M} \in M_{m \times n}$, show that the computed $\hat{\mathbf{B}}$ satisfies $\|\hat{\mathbf{B}} - \mathbf{H} \mathbf{A}\| \leq (10m + 16) \sqrt{n} u \|\mathbf{A}\|$.
- Deduce that $\sigma_i(\hat{\mathbf{B}})$ differs from $\sigma_i(\mathbf{A})$ by at most $(10m + 16) \sqrt{n} u \|\mathbf{A}\|$ for every i . *Hint: Chapter 20 §10 or Chapter 17 §09 bounds the perturbation of singular values by the norm of the perturbation.*
- Explain why the corresponding statement is false for the classical Gram–Schmidt process, citing [Example \(Three columns and three answers\)](#).

Solution.

- (a) By Theorem 24.3.3 (b), $\hat{\mathbf{B}} - \mathbf{H}\mathbf{A} = \mathbf{H}\Delta\mathbf{A}$ with $\|\Delta\mathbf{A}\|_F \leq (10m + 16)u\|\mathbf{A}\|_F$. Since $\mathbf{H}$ is unitary (Proposition 11.8.2 (c)) it leaves the 2-norm unchanged, so

$$\|\hat{\mathbf{B}} - \mathbf{H}\mathbf{A}\| = \|\Delta\mathbf{A}\| \leq \|\Delta\mathbf{A}\|_F \leq (10m + 16)u\|\mathbf{A}\|_F \leq (10m + 16)\sqrt{n}u\|\mathbf{A}\|.$$

- (b) $\mathbf{H}\mathbf{A}$ has the same singular values as $\mathbf{A}$: $(\mathbf{H}\mathbf{A})^*(\mathbf{H}\mathbf{A}) = \mathbf{A}^*\mathbf{H}^*\mathbf{H}\mathbf{A} = \mathbf{A}^*\mathbf{A}$, so Definition 13.9.1 reads the same list off the same matrix. Weyl's inequality for singular values (Theorem 17.9.4 with $j = 1$, or Corollary 17.9.5 directly) gives $|\sigma_i(\hat{\mathbf{B}}) - \sigma_i(\mathbf{H}\mathbf{A})| \leq \|\hat{\mathbf{B}} - \mathbf{H}\mathbf{A}\|$ for every i , and the right-hand side is bounded in (a).
- (c) The Gram–Schmidt recursion of Theorem 11.2.2 does not apply a unitary matrix to $\mathbf{A}$; it *builds* a matrix $\hat{\mathbf{Q}}$ out of $\mathbf{A}$ and hopes it is unitary. Example (Three columns and three answers) measures the damage: for $\kappa_2(\mathbf{A}) = 3 \cdot 10^6$ the computed vectors have inner products of size $9 \cdot 10^{-5}$, so $\hat{\mathbf{Q}}$ is unitary only to five digits, and $\|\hat{\mathbf{Q}}^\top \hat{\mathbf{Q}} - \mathbf{I}\|$ exceeds u by eleven orders of magnitude on that one example. (That it does so at the rate $u\kappa_2(\mathbf{A})^2$ in general is the classical estimate quoted above, which this book does not prove.) Any conclusion drawn from “unitary factors leave singular values alone” then inherits that error instead of none. The reflection route wins because the unitary matrix is never in question: it is exactly unitary by construction, and only its *input* is perturbed.