

16.5 Inverting a Perturbation of the Identity

Chapter 10 §10 summed the series $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \dots$ and identified its sum as $(\mathbf{I} - \mathbf{A})^{-1}$, under the exact hypothesis that every eigenvalue of $\mathbf{A}$ has modulus less than 1. What it could not say is how large that inverse is. This section reruns the argument with a norm in hand. The hypothesis becomes cruder — $\|\mathbf{A}\| < 1$ instead of $\rho(\mathbf{A}) < 1$ — and the conclusion becomes quantitative, which is the trade this chapter keeps making. The payoff is immediate: the invertible matrices turn out to form an open set, inversion turns out to be continuous, and a first estimate appears for how far the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ moves when $\mathbf{A}$ and $\mathbf{b}$ are disturbed.

Throughout, $\|\cdot\|$ is the operator norm on $M_n(\mathbb{C})$ induced by a norm on $\mathbb{C}^n$ (Definition 16.3.2), so that $\|\mathbf{I}\| = 1$ and $\|\mathbf{XY}\| \leq \|\mathbf{X}\|\|\mathbf{Y}\|$ by Theorem 16.3.2 (c) and (d). Both properties are used in every proof below, and the first of them is not free: the closing warning shows what happens without it.

16.5.1 The geometric series, with a number attached

The scalar identity behind everything here is $1 + r + r^2 + \dots = 1/(1 - r)$ for $0 \leq r < 1$. Chapter 10 §10 gave the matrix version of the *identity*. What a norm adds is the matrix version of the *inequality* $|1 + r + \dots| \leq 1/(1 - r)$: a ceiling on the inverse, expressed in the data.

Theorem 16.5.1. Neumann Series with a Bound

Let $\mathbf{A} \in M_n(\mathbb{C})$ with $\|\mathbf{A}\| < 1$, and write $a = \|\mathbf{A}\|$. Then:

- (a) $\mathbf{I} - \mathbf{A}$ is invertible, and $\sum_{k \geq 0} \mathbf{A}^k = (\mathbf{I} - \mathbf{A})^{-1}$;
- (b) $\|(\mathbf{I} - \mathbf{A})^{-1}\| \leq \frac{1}{1 - a}$;
- (c) for every $K \geq 0$, the partial sum $\mathbf{S}_K = \sum_{k=0}^K \mathbf{A}^k$ satisfies

$$\|(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{S}_K\| \leq \frac{a^{K+1}}{1 - a};$$

- (d) $\|(\mathbf{I} - \mathbf{A})^{-1}\| \geq \frac{1}{1 + a}$.

Idea. Part (a) is not new work: $\|\mathbf{A}\| < 1$ forces $\rho(\mathbf{A}) < 1$ by Theorem 16.4.1, which is exactly the hypothesis of Chapter 10's Theorem 10.10.6. Everything else is the tail-bounding move. Submultiplicativity turns $\|\mathbf{A}^k\|$ into a^k , so each partial sum is bounded by a scalar geometric sum; the bound then has to be carried across the limit, and what carries it is the fact that a norm is continuous. For (c), the tail of the series factors as $\mathbf{A}^{K+1}$ times the whole series again, so (b) bounds it. For (d), read $\mathbf{I} = (\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A})^{-1}$ through submultiplicativity.

Proof.

- (a) By Theorem 16.4.1 (a), $\rho(\mathbf{A}) \leq \|\mathbf{A}\| = a < 1$, so every eigenvalue of $\mathbf{A}$ has modulus less than 1. That is the hypothesis of Theorem 10.10.6, which gives both the invertibility of $\mathbf{I} - \mathbf{A}$ and the value of the sum.
- (b) Submultiplicativity (Theorem 16.3.2 (d)) gives $\|\mathbf{A}^k\| \leq a^k$ for every $k \geq 1$, by induction on k ; and $\|\mathbf{A}^0\| = \|\mathbf{I}\| = 1 = a^0$ by part (c) of the same theorem. Hence, by subadditivity applied K times,

$$\|\mathbf{S}_K\| \leq \sum_{k=0}^K \|\mathbf{A}^k\| \leq \sum_{k=0}^K a^k = \frac{1 - a^{K+1}}{1 - a} \leq \frac{1}{1 - a},$$

the last step because $a^{K+1} \geq 0$ and $1 - a > 0$. By (a) the sequence $(\mathbf{S}_K)$ converges entrywise to $(\mathbf{I} - \mathbf{A})^{-1}$, so $\|\mathbf{S}_K - (\mathbf{I} - \mathbf{A})^{-1}\| \rightarrow 0$ by Corollary 16.2.4. Then Lemma 16.1.1 gives

$$|\|\mathbf{S}_K\| - \|(\mathbf{I} - \mathbf{A})^{-1}\|| \leq \|\mathbf{S}_K - (\mathbf{I} - \mathbf{A})^{-1}\| \rightarrow 0,$$

so $\|\mathbf{S}_K\| \rightarrow \|(\mathbf{I} - \mathbf{A})^{-1}\|$. A nonstrict inequality survives a limit, so a limit of numbers each at most $1/(1 - a)$ is at most $1/(1 - a)$, which proves (b).

- (c) Fix K . For every $M \geq 1$, multiplying out and canceling,

$$\mathbf{S}_{K+M} - \mathbf{S}_K = \sum_{k=K+1}^{K+M} \mathbf{A}^k = \mathbf{A}^{K+1} \mathbf{S}_{M-1}.$$

Let $M \rightarrow \infty$. The left-hand side tends to $(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{S}_K$ by (a). The right-hand side tends to $\mathbf{A}^{K+1}(\mathbf{I} - \mathbf{A})^{-1}$, because

$$\|\mathbf{A}^{K+1} \mathbf{S}_{M-1} - \mathbf{A}^{K+1}(\mathbf{I} - \mathbf{A})^{-1}\| \leq \|\mathbf{A}^{K+1}\| \|\mathbf{S}_{M-1} - (\mathbf{I} - \mathbf{A})^{-1}\|,$$

which tends to 0 by (a) and Corollary 16.2.4 again. A sequence has at most one limit, so

$$(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{S}_K = \mathbf{A}^{K+1}(\mathbf{I} - \mathbf{A})^{-1}.$$

Taking norms and using submultiplicativity together with (b),

$$\|(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{S}_K\| \leq \|\mathbf{A}^{K+1}\| \|(\mathbf{I} - \mathbf{A})^{-1}\| \leq \frac{a^{K+1}}{1 - a}.$$

- (d) $1 = \|\mathbf{I}\| = \|(\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A})^{-1}\| \leq \|\mathbf{I} - \mathbf{A}\| \|(\mathbf{I} - \mathbf{A})^{-1}\|$, and $\|\mathbf{I} - \mathbf{A}\| \leq \|\mathbf{I}\| + \|\mathbf{A}\| = 1 + a$ by subadditivity. Since $1 + a > 0$ we may divide, which proves (d). This proves the theorem. ■

Parts (b) and (d) together pin the inverse between two numbers computed from a alone:

$$\frac{1}{1 + a} \leq \|(\mathbf{I} - \mathbf{A})^{-1}\| \leq \frac{1}{1 - a}.$$

As a approaches 1 the upper bound escapes to infinity while the lower one stays near $\frac{1}{2}$, and the gap is real rather than an artifact of the proof: $\mathbf{A} = a\mathbf{I}$ attains the upper bound, and $\mathbf{A} = -a\mathbf{I}$ attains the lower one.

Example. The bound and the truth

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1/2 \\ 1/3 & 0 \end{pmatrix}.$$

Using the ∞ -norm, check the hypothesis, compute $(\mathbf{I} - \mathbf{A})^{-1}$ exactly, and compare it with the bound of [Theorem 16.5.1](#) (b). Then compare $\mathbf{S}_3$ with the true inverse and with the tail bound of (c).

Solution. The hypothesis. The absolute row sums are $\frac{1}{2}$ and $\frac{1}{3}$, so $\|\mathbf{A}\|_\infty = \frac{1}{2} < 1$ by [Theorem 16.3.3](#) (b). So [Theorem 16.5.1](#) applies with $a = \frac{1}{2}$.

The inverse. $\mathbf{I} - \mathbf{A} = \begin{pmatrix} 1 & -1/2 \\ -1/3 & 1 \end{pmatrix}$ has determinant $1 - \frac{1}{6} = \frac{5}{6}$, so by [Theorem 0.9.8](#)

$$(\mathbf{I} - \mathbf{A})^{-1} = \frac{6}{5} \begin{pmatrix} 1 & 1/2 \\ 1/3 & 1 \end{pmatrix} = \begin{pmatrix} 6/5 & 3/5 \\ 2/5 & 6/5 \end{pmatrix},$$

whose largest absolute row sum is $\frac{6}{5} + \frac{3}{5} = \frac{9}{5} = 1.8$. The bound is $1/(1 - \frac{1}{2}) = 2$. So the estimate is correct and loses only ten percent.

The partial sum. Here $\mathbf{A}^2 = \frac{1}{6}\mathbf{I}$, so $\mathbf{A}^3 = \frac{1}{6}\mathbf{A}$ and

$$\mathbf{S}_3 = (1 + \frac{1}{6})(\mathbf{I} + \mathbf{A}) = \frac{7}{6} \begin{pmatrix} 1 & 1/2 \\ 1/3 & 1 \end{pmatrix}.$$

Hence $(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{S}_3 = (\frac{6}{5} - \frac{7}{6})(\mathbf{I} + \mathbf{A}) = \frac{1}{30}(\mathbf{I} + \mathbf{A})$, with ∞ -norm $\frac{1}{30} \cdot \frac{3}{2} = \frac{1}{20} = 0.05$. The bound of (c) with $K = 3$ is $(\frac{1}{2})^4/(1 - \frac{1}{2}) = \frac{1}{8} = 0.125$. Again correct, and this time loose by a factor of 2.5 — the price of replacing four matrices by four copies of one number.

16.5.2 Which theorem is stronger

The two Neumann theorems must be compared honestly, because neither contains the other.

Chapter 10's [Theorem 10.10.6](#) assumes $\rho(\mathbf{A}) < 1$ and concludes that the series converges, with sum $(\mathbf{I} - \mathbf{A})^{-1}$; and its hypothesis is **sharp**, since the statement is an “if and only if”. [Theorem 16.5.1](#) assumes $\|\mathbf{A}\| < 1$, which by [Theorem 16.4.1](#) is a **strictly stronger** demand: $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$ always, and the inequality is often strict. For

$$\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$$

we have $\rho(\mathbf{A}) = 0$, so Chapter 10's theorem applies and returns $\sum_k \mathbf{A}^k = \mathbf{I} + \mathbf{A}$; but $\|\mathbf{A}\| = 2$ in each of the 1-, 2-, and ∞ -norms, so this section's theorem says nothing at all about it.

In the other direction, when this section's theorem does apply it says strictly more: a ceiling on the inverse, a rate at which the partial sums approach it, and a floor as well. Chapter 10 could not have stated any of the three, having no way to measure a matrix.

There is a third thing to say, and it dissolves most of the apparent conflict. If $\rho(\mathbf{A}) < 1$, then by [Theorem 16.4.3](#) there is a norm $\|\cdot\|_t$ with $\|\mathbf{A}\|_t \leq \rho(\mathbf{A}) + \varepsilon < 1$, and [Theorem 16.5.1](#) applies **in that norm**. For the matrix above, taking $\mathbf{D} = \text{diag}(1, 4)$ and $\|\mathbf{x}\|_t = \|\mathbf{D}\mathbf{x}\|_\infty$ gives $\|\mathbf{A}\|_t = \|\mathbf{D}\mathbf{A}\mathbf{D}^{-1}\|_\infty = \frac{1}{2}$, so the bound reads $\|(\mathbf{I} - \mathbf{A})^{-1}\|_t \leq 2$; and indeed $\mathbf{D}(\mathbf{I} + \mathbf{A})\mathbf{D}^{-1} = \begin{pmatrix} 1 & 1/2 \\ 0 & 1 \end{pmatrix}$

has ∞ -norm $\frac{3}{2} \leq 2$. So the quantitative theorem can always be brought to bear on a matrix Chapter 10 accepts — at the cost of changing the norm, and the bound it then produces is a bound in that new norm, which may be a strange unit of measurement.

Quick check. Both theorems need a hypothesis. Which of the two hypotheses, $\rho(\mathbf{A}) < 1$ and $\|\mathbf{A}\| < 1$, can be checked by looking at the entries of $\mathbf{A}$ and doing no linear algebra?

Solution. The second, for a suitable norm: $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_\infty$ are the largest absolute column and row sums (Theorem 16.3.3), read straight off the entries. Deciding $\rho(\mathbf{A}) < 1$ requires the eigenvalues, that is, the roots of a degree- n polynomial. This is the practical reason the cruder hypothesis is worth having: it is the one a reader can verify at a glance, and Corollary 16.4.2 is the same observation used in the other direction.

Warning. **The hypothesis $\|\mathbf{A}\| < 1$ is sufficient and never necessary.** A matrix can be as far from $\mathbf{I}$ as one likes and still have $\mathbf{I} - \mathbf{A}$ invertible: $\mathbf{A} = 5\mathbf{I}$ gives $(\mathbf{I} - \mathbf{A})^{-1} = -\frac{1}{4}\mathbf{I}$. What fails when $\|\mathbf{A}\| \geq 1$ is not the invertibility but the *argument*, and with it the bound. Read $\|\mathbf{A}\| < 1$ as “ $\mathbf{A}$ is a small perturbation”, not as a criterion.

16.5.3 The invertible matrices form an open set

The theorem above concerns perturbations of $\mathbf{I}$. Every invertible matrix can be moved to $\mathbf{I}$ by multiplication, so the general statement follows at once, and it is the statement several earlier arguments wanted: a matrix near an invertible matrix is invertible, and its inverse is near.

Corollary 16.5.2. Invertibility Is an Open Condition, and Inversion Is Continuous

Let $\mathbf{B} \in M_n(\mathbb{C})$ be invertible and let $\mathbf{E} \in M_n(\mathbb{C})$ satisfy

$$\|\mathbf{E}\| < \frac{1}{\|\mathbf{B}^{-1}\|}.$$

Write $r = \|\mathbf{B}^{-1}\|\|\mathbf{E}\| < 1$. Then:

- (a) $\mathbf{B} + \mathbf{E}$ is invertible, and $\|(\mathbf{B} + \mathbf{E})^{-1}\| \leq \frac{\|\mathbf{B}^{-1}\|}{1 - r}$;
- (b) $\|(\mathbf{B} + \mathbf{E})^{-1} - \mathbf{B}^{-1}\| \leq \frac{\|\mathbf{B}^{-1}\|^2\|\mathbf{E}\|}{1 - r}$.

Consequently $\text{GL}_n(\mathbb{C})$ is an open subset of $M_n(\mathbb{C})$: around every invertible $\mathbf{B}$ the ball of radius $1/\|\mathbf{B}^{-1}\|$ consists of invertible matrices. And inversion is continuous on it: if $\mathbf{B}_m \rightarrow \mathbf{B}$ with $\mathbf{B}$ invertible, then $\mathbf{B}_m$ is invertible for all large m and $\mathbf{B}_m^{-1} \rightarrow \mathbf{B}^{-1}$.

Idea. Factor the perturbed matrix as $\mathbf{B} + \mathbf{E} = \mathbf{B}(\mathbf{I} + \mathbf{B}^{-1}\mathbf{E})$ — legitimate because $\mathbf{B}$ is invertible — and notice that $\|\mathbf{B}^{-1}\mathbf{E}\| \leq r < 1$, so the second factor is a perturbation of the identity of exactly the kind [Theorem 16.5.1](#) handles, with $\mathbf{A} = -\mathbf{B}^{-1}\mathbf{E}$. For (b), the difference of two inverses always factors as $\mathbf{X}^{-1} - \mathbf{Y}^{-1} = \mathbf{X}^{-1}(\mathbf{Y} - \mathbf{X})\mathbf{Y}^{-1}$, and both inverses

are now bounded.

Proof. Put $\mathbf{A} = -\mathbf{B}^{-1}\mathbf{E}$. By submultiplicativity, $\|\mathbf{A}\| \leq \|\mathbf{B}^{-1}\|\|\mathbf{E}\| = r < 1$.

- (a) Since $\mathbf{B}$ is invertible, $\mathbf{B} + \mathbf{E} = \mathbf{B}(\mathbf{I} + \mathbf{B}^{-1}\mathbf{E}) = \mathbf{B}(\mathbf{I} - \mathbf{A})$. By [Theorem 16.5.1](#) (a) the factor $\mathbf{I} - \mathbf{A}$ is invertible, so $\mathbf{B} + \mathbf{E}$ is a product of invertible matrices and is invertible, with

$$(\mathbf{B} + \mathbf{E})^{-1} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}^{-1}.$$

By submultiplicativity and [Theorem 16.5.1](#) (b),

$$\|(\mathbf{B} + \mathbf{E})^{-1}\| \leq \|(\mathbf{I} - \mathbf{A})^{-1}\| \|\mathbf{B}^{-1}\| \leq \frac{\|\mathbf{B}^{-1}\|}{1 - \|\mathbf{A}\|} \leq \frac{\|\mathbf{B}^{-1}\|}{1 - r},$$

the last step because $\|\mathbf{A}\| \leq r$ and $t \mapsto 1/(1 - t)$ increases on $[0, 1)$.

- (b) Multiplying the identity $\mathbf{B} - (\mathbf{B} + \mathbf{E}) = -\mathbf{E}$ on the left by $(\mathbf{B} + \mathbf{E})^{-1}$ and on the right by $\mathbf{B}^{-1}$ gives

$$(\mathbf{B} + \mathbf{E})^{-1} - \mathbf{B}^{-1} = -(\mathbf{B} + \mathbf{E})^{-1}\mathbf{E}\mathbf{B}^{-1}.$$

Taking norms and applying submultiplicativity twice and then (a),

$$\|(\mathbf{B} + \mathbf{E})^{-1} - \mathbf{B}^{-1}\| \leq \frac{\|\mathbf{B}^{-1}\|}{1 - r} \|\mathbf{E}\| \|\mathbf{B}^{-1}\| = \frac{\|\mathbf{B}^{-1}\|^2 \|\mathbf{E}\|}{1 - r}.$$

For the two consequences: openness is the statement of (a) with $\mathbf{E}$ ranging over the ball of radius $1/\|\mathbf{B}^{-1}\|$. For continuity, let $\mathbf{B}_m \rightarrow \mathbf{B}$ with $\mathbf{B}$ invertible, and put $\mathbf{E}_m = \mathbf{B}_m - \mathbf{B}$. By [Corollary 16.2.4](#), $\|\mathbf{E}_m\| \rightarrow 0$, so $\|\mathbf{E}_m\| < 1/(2\|\mathbf{B}^{-1}\|)$ for all large m ; for those m , part (a) applies with $r \leq \frac{1}{2}$ and $\mathbf{B}_m$ is invertible, and part (b) gives

$$\|\mathbf{B}_m^{-1} - \mathbf{B}^{-1}\| \leq 2\|\mathbf{B}^{-1}\|^2 \|\mathbf{E}_m\| \rightarrow 0.$$

This proves the corollary. ■

Example. How far an invertible matrix can be pushed

Take $\mathbf{B} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ and work with $\|\cdot\|_\infty$. Then $\mathbf{B}^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{1}{2} \end{pmatrix}$, whose largest absolute row sum is $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$, so $\|\mathbf{B}^{-1}\|_\infty = \frac{3}{4}$. [Corollary 16.5.2](#) therefore keeps $\mathbf{B} + \mathbf{E}$ invertible for every $\mathbf{E}$ with

$$\|\mathbf{E}\|_\infty < \frac{1}{\|\mathbf{B}^{-1}\|_\infty} = \frac{4}{3}.$$

For a perturbation of size $\|\mathbf{E}\|_\infty \leq \frac{1}{2}$ we get $r = \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8}$, and part (a) bounds

$$\|(\mathbf{B} + \mathbf{E})^{-1}\|_\infty \leq \frac{3/4}{1 - 3/8} = \frac{6}{5}.$$

The bound is honest rather than tight: pushing every entry of $\mathbf{E}$ to $\pm\frac{1}{4}$ gives at worst $\|(\mathbf{B} + \mathbf{E})^{-1}\|_\infty = \frac{12}{11} \approx 1.09$, against the promised 1.2. What the corollary buys is not the sharpest constant but a radius inside which no invertibility check is needed at all.

Part (a) has a reading worth recording, since the relative-error theorem of §08 is the same sentence with the constant named: **the distance from an invertible $\mathbf{B}$ to the nearest singular matrix is at least $1/\|\mathbf{B}^{-1}\|$** . A large $\|\mathbf{B}^{-1}\|$ is therefore not an abstract complaint; it says that $\mathbf{B}$ is close, in the literal sense of the norm, to a matrix that cannot be inverted at all.

Remark. Chapter 12 §01 proved that every matrix has diagonalizable matrices arbitrarily close to it (Corollary 12.1.6) and used the word “close” before this chapter existed. [Corollary 16.5.2](#) is the first of the two facts that argument needs; the second is the continuity of the eigenvalues, and the two are put together in §07.

16.5.4 Perturbing a solve

The question the rest of the chapter circles is this. A linear system $\mathbf{Ax} = \mathbf{b}$ is solved; the data $\mathbf{A}$ and $\mathbf{b}$ were only known approximately; how wrong can $\mathbf{x}$ be? With [Corollary 16.5.2](#) available the answer is three lines, and it already contains the quantity §08 will name.

Theorem 16.5.3. Relative Error in a Perturbed Solve

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible, let $\mathbf{b} \neq \mathbf{0}$, and let $\mathbf{x}$ be the solution of $\mathbf{Ax} = \mathbf{b}$. Let $\mathbf{E} \in M_n(\mathbb{C})$ and $\mathbf{f} \in \mathbb{C}^n$ be perturbations with

$$r = \|\mathbf{A}^{-1}\| \|\mathbf{E}\| < 1,$$

and let $\tilde{\mathbf{x}}$ be the solution of $(\mathbf{A} + \mathbf{E})\tilde{\mathbf{x}} = \mathbf{b} + \mathbf{f}$, which exists and is unique by [Corollary 16.5.2](#). Then

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\| \|\mathbf{A}^{-1}\|}{1 - r} \left(\frac{\|\mathbf{E}\|}{\|\mathbf{A}\|} + \frac{\|\mathbf{f}\|}{\|\mathbf{b}\|} \right).$$

Idea. Subtract the two equations to isolate $\mathbf{A}(\tilde{\mathbf{x}} - \mathbf{x})$; what is left on the other side is $\mathbf{f} - \mathbf{E}\tilde{\mathbf{x}}$. Apply $\mathbf{A}^{-1}$ and take norms. The one awkward term is $\|\tilde{\mathbf{x}}\|$, which is not data; write $\tilde{\mathbf{x}} = \mathbf{x} + (\tilde{\mathbf{x}} - \mathbf{x})$ and move the resulting copy of $\|\tilde{\mathbf{x}} - \mathbf{x}\|$ to the left, which is where the factor $1/(1 - r)$ comes from. Finally $\|\mathbf{b}\| \leq \|\mathbf{A}\| \|\mathbf{x}\|$ converts $1/\|\mathbf{x}\|$ into $\|\mathbf{A}\|/\|\mathbf{b}\|$, and the two perturbations appear only through their relative sizes.

Proof. Write $\mathbf{d} = \tilde{\mathbf{x}} - \mathbf{x}$. Subtracting $\mathbf{Ax} = \mathbf{b}$ from $(\mathbf{A} + \mathbf{E})\tilde{\mathbf{x}} = \mathbf{b} + \mathbf{f}$ gives $\mathbf{A}\tilde{\mathbf{x}} + \mathbf{E}\tilde{\mathbf{x}} - \mathbf{Ax} = \mathbf{f}$, that is, $\mathbf{Ad} = \mathbf{f} - \mathbf{E}\tilde{\mathbf{x}}$. Since $\mathbf{A}$ is invertible, $\mathbf{d} = \mathbf{A}^{-1}(\mathbf{f} - \mathbf{E}\tilde{\mathbf{x}})$, and Theorem 16.3.2 (a) with subadditivity gives

$$\|\mathbf{d}\| \leq \|\mathbf{A}^{-1}\| (\|\mathbf{f}\| + \|\mathbf{E}\| \|\tilde{\mathbf{x}}\|).$$

Now $\|\tilde{\mathbf{x}}\| \leq \|\mathbf{x}\| + \|\mathbf{d}\|$, so

$$\|\mathbf{d}\| \leq \|\mathbf{A}^{-1}\| \|\mathbf{f}\| + \|\mathbf{A}^{-1}\| \|\mathbf{E}\| \|\mathbf{x}\| + r \|\mathbf{d}\|,$$

using $\|\mathbf{A}^{-1}\| \|\mathbf{E}\| = r$ on the last term. Since $r < 1$ we may subtract $r \|\mathbf{d}\|$ and divide by $1 - r > 0$:

$$\|\mathbf{d}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1 - r} (\|\mathbf{f}\| + \|\mathbf{E}\| \|\mathbf{x}\|).$$

Divide by $\|\mathbf{x}\|$, which is nonzero because $\mathbf{b} \neq \mathbf{0}$:

$$\frac{\|\mathbf{d}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|}{1-r} \left(\frac{\|\mathbf{f}\|}{\|\mathbf{x}\|} + \|\mathbf{E}\| \right).$$

Finally $\|\mathbf{b}\| = \|\mathbf{Ax}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$, so $1/\|\mathbf{x}\| \leq \|\mathbf{A}\|/\|\mathbf{b}\|$. Substituting this in the first term and writing $\|\mathbf{E}\| = \|\mathbf{A}\| \cdot \|\mathbf{E}\|/\|\mathbf{A}\|$ in the second gives the stated inequality. This proves the theorem. ■

The shape of the conclusion is the point. On the right, the two perturbations appear only as **relative** sizes, $\|\mathbf{E}\|/\|\mathbf{A}\|$ and $\|\mathbf{f}\|/\|\mathbf{b}\|$; on the left is the relative size of the error. Between them stands one number built from $\mathbf{A}$ alone,

$$\|\mathbf{A}\| \|\mathbf{A}^{-1}\|,$$

inflated slightly by $1/(1-r)$, which is close to 1 whenever the perturbation of $\mathbf{A}$ is small. That number is the condition number of §08. Everything that section does is to give it a name, compute it for the 2-norm, and show that the bound is attained.

Example. A perturbation of the matrix

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} 0 & 0 \\ 0 & 1/10 \end{pmatrix}, \quad \mathbf{f} = \mathbf{0}.$$

Solve both systems in the ∞ -norm, compute the relative error, and compare it with [Theorem 16.5.3](#).

Solution. The two solutions. $\det \mathbf{A} = 1$, so $\mathbf{A}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$ by Theorem 0.9.8, and $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = (3-2, -3+4) = (1, 1)$. For the perturbed matrix, $\mathbf{A} + \mathbf{E} = \begin{pmatrix} 2 & 1 \\ 1 & 11/10 \end{pmatrix}$ has determinant $\frac{22}{10} - 1 = \frac{6}{5}$, so

$$\tilde{\mathbf{x}} = \frac{5}{6} \begin{pmatrix} 11/10 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \frac{5}{6} \begin{pmatrix} 13/10 \\ 1 \end{pmatrix} = \begin{pmatrix} 13/12 \\ 5/6 \end{pmatrix}.$$

The error. $\tilde{\mathbf{x}} - \mathbf{x} = (\frac{1}{12}, -\frac{1}{6})$, so $\|\tilde{\mathbf{x}} - \mathbf{x}\|_{\infty} = \frac{1}{6}$ and, since $\|\mathbf{x}\|_{\infty} = 1$, the relative error is $\frac{1}{6} \approx 0.167$.

The bound. The absolute row sums give $\|\mathbf{A}\|_{\infty} = 3$ and $\|\mathbf{A}^{-1}\|_{\infty} = 3$, so $\|\mathbf{A}\|_{\infty} \|\mathbf{A}^{-1}\|_{\infty} = 9$. Also $\|\mathbf{E}\|_{\infty} = \frac{1}{10}$ and $r = 3 \cdot \frac{1}{10} = \frac{3}{10} < 1$. With $\mathbf{f} = \mathbf{0}$ the bound is

$$\frac{9}{1 - \frac{3}{10}} \cdot \frac{1/10}{3} = \frac{9}{7/10} \cdot \frac{1}{30} = \frac{3}{7} \approx 0.429.$$

So $0.167 \leq 0.429$: correct, and loose by a factor of about 2.6. Note the accounting. The matrix was disturbed by 3.3% of its size and the answer moved by 16.7%, an amplification of about 5; the theorem promised at most $9/(1-0.3) \approx 12.9$.

Warning. All four bounds of this section need $\|\mathbf{I}\| = 1$, which is a property of **induced norms only**. The Frobenius norm is submultiplicative (Example (The Frobenius norm is submultiplicative)) but has $\|\mathbf{I}_n\|_F = \sqrt{n}$. Take $\mathbf{A} = \mathbf{O}$ in $M_2(\mathbb{C})$: then $\|\mathbf{A}\|_F = 0 < 1$, while

$$\|(\mathbf{I} - \mathbf{A})^{-1}\|_F = \|\mathbf{I}_2\|_F = \sqrt{2} > 1 = \frac{1}{1 - \|\mathbf{A}\|_F},$$

so Theorem 16.5.1 (b) is false for it. The repair is easy — Exercise (C3: Without $\|\mathbf{I}\| = 1$) — but it is not the same inequality, and using the Frobenius norm here is the most common way to arrive at a wrong constant.

16.5.5 Exercises

16.5.5.1 A. Check your understanding

Exercise. A1

- State Theorem 16.5.1 in full, including which kind of norm it assumes.
- Chapter 10 §10 already proved that $\sum_k \mathbf{A}^k$ converges to $(\mathbf{I} - \mathbf{A})^{-1}$ when $\rho(\mathbf{A}) < 1$. Say precisely what this section adds, and what it gives up.
- Determine whether the following statement is correct, and justify your answer: if $\mathbf{I} - \mathbf{A}$ is invertible then $\|\mathbf{A}\| < 1$.
- Where in the proof of Theorem 16.5.1 is $\|\mathbf{I}\| = 1$ used?
- State what Corollary 16.5.2 says about the distance from an invertible matrix to the singular matrices.

Solution.

- If $\|\cdot\|$ is the operator norm induced by a vector norm and $\|\mathbf{A}\| = a < 1$, then $\mathbf{I} - \mathbf{A}$ is invertible with $\sum_{k \geq 0} \mathbf{A}^k = (\mathbf{I} - \mathbf{A})^{-1}$; moreover $\|(\mathbf{I} - \mathbf{A})^{-1}\| \leq 1/(1 - a)$, the partial sums satisfy $\|(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{S}_K\| \leq a^{K+1}/(1 - a)$, and $\|(\mathbf{I} - \mathbf{A})^{-1}\| \geq 1/(1 + a)$.
- It adds three numbers: a ceiling on the inverse, a geometric rate for the partial sums, and a floor. It gives up sharpness of the hypothesis: $\rho(\mathbf{A}) < 1$ is necessary and sufficient for convergence, whereas $\|\mathbf{A}\| < 1$ is only sufficient, and strictly stronger by Theorem 16.4.1.
- Incorrect. $\mathbf{A} = 5\mathbf{I}$ has $\mathbf{I} - \mathbf{A} = -4\mathbf{I}$, invertible, while $\|\mathbf{A}\| = 5$ in every induced norm. Invertibility of $\mathbf{I} - \mathbf{A}$ says only $1 \notin \text{spec}(\mathbf{A})$.
- Twice, and both times to make a geometric sum start at 1: in (b), where $\|\mathbf{A}^0\| = \|\mathbf{I}\| = 1$ is the $k = 0$ term of $\sum_k a^k$; and in (d), where $1 = \|\mathbf{I}\|$ is the left-hand side of the estimate.
- It is at least $1/\|\mathbf{B}^{-1}\|$: every $\mathbf{E}$ with $\|\mathbf{E}\| < 1/\|\mathbf{B}^{-1}\|$ leaves $\mathbf{B} + \mathbf{E}$ invertible.

16.5.5.2 B. Practice

Exercise. B1: Check, sum, compare

Let $\mathbf{A} = \begin{pmatrix} 1/4 & 1/4 \\ 0 & 1/2 \end{pmatrix}$.

- (a) Verify the hypothesis of [Theorem 16.5.1](#) for $\|\cdot\|_\infty$.
- (b) Compute $(\mathbf{I} - \mathbf{A})^{-1}$ exactly and compare $\|(\mathbf{I} - \mathbf{A})^{-1}\|_\infty$ with the bound of (b) and with the lower bound of (d).

Solution.

- (a) The absolute row sums are $\frac{1}{2}$ and $\frac{1}{2}$, so $\|\mathbf{A}\|_\infty = \frac{1}{2} < 1$ by Theorem 16.3.3 (b).
- (b) $\mathbf{I} - \mathbf{A} = \begin{pmatrix} 3/4 & -1/4 \\ 0 & 1/2 \end{pmatrix}$, upper triangular with determinant $\frac{3}{8}$, so by Theorem 0.9.8

$$(\mathbf{I} - \mathbf{A})^{-1} = \frac{8}{3} \begin{pmatrix} 1/2 & 1/4 \\ 0 & 3/4 \end{pmatrix} = \begin{pmatrix} 4/3 & 2/3 \\ 0 & 2 \end{pmatrix}.$$

The absolute row sums are 2 and 2, so $\|(\mathbf{I} - \mathbf{A})^{-1}\|_\infty = 2$. The upper bound is $1/(1 - \frac{1}{2}) = 2$, attained exactly; the lower bound is $1/(1 + \frac{1}{2}) = \frac{2}{3}$, comfortably satisfied. The upper bound is attained because both rows of $\mathbf{A}$ have the same absolute sum and no cancellation occurs anywhere in the geometric series.

Exercise. B2: A perturbed right-hand side

Let $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$, $\mathbf{b} = (4, 3)$, and $\mathbf{f} = (1/10, 0)$, with $\mathbf{E} = \mathbf{O}$. Solve both systems, compute the relative error in the ∞ -norm, and compare with [Theorem 16.5.3](#).

Solution. $\det \mathbf{A} = 5$, so $\mathbf{A}^{-1} = \frac{1}{5} \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ by Theorem 0.9.8, and $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \frac{1}{5}(8 - 3, -4 + 9) = (1, 1)$.

With $\mathbf{b} + \mathbf{f} = (\frac{41}{10}, 3)$ we get $\tilde{\mathbf{x}} = \frac{1}{5}(\frac{82}{10} - 3, -\frac{41}{10} + 9) = (\frac{52}{50}, \frac{49}{50})$. Hence $\tilde{\mathbf{x}} - \mathbf{x} = (\frac{2}{50}, -\frac{1}{50})$ and $\|\tilde{\mathbf{x}} - \mathbf{x}\|_\infty = \frac{1}{25} = 0.04$, a relative error of 0.04 since $\|\mathbf{x}\|_\infty = 1$.

For the bound: $\|\mathbf{A}\|_\infty = 4$, $\|\mathbf{A}^{-1}\|_\infty = \frac{4}{5}$, so $\|\mathbf{A}\|_\infty \|\mathbf{A}^{-1}\|_\infty = \frac{16}{5}$. Here $\mathbf{E} = \mathbf{O}$, so $r = 0$, and $\|\mathbf{f}\|_\infty / \|\mathbf{b}\|_\infty = \frac{1/10}{4} = \frac{1}{40}$. The bound is $\frac{16}{5} \cdot \frac{1}{40} = \frac{2}{25} = 0.08$. So $0.04 \leq 0.08$, loose by a factor of exactly 2.

Exercise. B3: Staying invertible

Let $\mathbf{B} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$. Using $\|\cdot\|_\infty$, find a radius $\delta > 0$ such that every $\mathbf{E}$ with $\|\mathbf{E}\|_\infty < \delta$ leaves $\mathbf{B} + \mathbf{E}$ invertible, and bound $\|(\mathbf{B} + \mathbf{E})^{-1}\|_\infty$ for $\|\mathbf{E}\|_\infty \leq \frac{1}{4}$. Is your δ the largest possible?

Solution. $\mathbf{B}^{-1} = \text{diag}(\frac{1}{2}, 2)$, so $\|\mathbf{B}^{-1}\|_\infty = 2$ and [Corollary 16.5.2](#) gives $\delta = 1/\|\mathbf{B}^{-1}\|_\infty = \frac{1}{2}$.

For $\|\mathbf{E}\|_\infty \leq \frac{1}{4}$ we get $r = 2 \cdot \frac{1}{4} = \frac{1}{2}$, so part (a) of the corollary gives

$$\|(\mathbf{B} + \mathbf{E})^{-1}\|_\infty \leq \frac{2}{1 - \frac{1}{2}} = 4.$$

It is the largest possible here. Taking $\mathbf{E} = \text{diag}(0, -\frac{1}{2})$ gives $\|\mathbf{E}\|_\infty = \frac{1}{2}$ and $\mathbf{B} + \mathbf{E} = \text{diag}(2, 0)$, which is singular. So no radius larger than $\frac{1}{2}$ works, and the estimate $1/\|\mathbf{B}^{-1}\|$ is attained by this $\mathbf{B}$. (It is not attained by every $\mathbf{B}$ in every norm; [Exercise \(C2: Distance to the singular matrices\)](#) returns to this.)

16.5.5.3 C. Going deeper

Exercise. C1: The inverse to first order

Let $\|\mathbf{A}\| < 1$.

- (a) Prove that $\|(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{I} - \mathbf{A}\| \leq \frac{\|\mathbf{A}\|^2}{1 - \|\mathbf{A}\|}$.
 (b) Deduce that for a fixed $\mathbf{M} \in M_n(\mathbb{C})$ and small real t ,

$$(\mathbf{I} - t\mathbf{M})^{-1} = \mathbf{I} + t\mathbf{M} + \mathbf{R}(t), \quad \|\mathbf{R}(t)\|/|t| \rightarrow 0$$

as $t \rightarrow 0$.

Hint for (a): this is [Theorem 16.5.1](#) (c) with a particular K .

Solution.

- (a) Take $K = 1$ in [Theorem 16.5.1](#) (c): $\mathbf{S}_1 = \mathbf{I} + \mathbf{A}$, so

$$\|(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{I} - \mathbf{A}\| \leq \frac{\|\mathbf{A}\|^{1+1}}{1 - \|\mathbf{A}\|} = \frac{\|\mathbf{A}\|^2}{1 - \|\mathbf{A}\|}.$$

- (b) If $\mathbf{M} = \mathbf{O}$ the claim is trivial, so assume $\mathbf{M} \neq \mathbf{O}$ and let $|t| < 1/(2\|\mathbf{M}\|)$. Then $\|t\mathbf{M}\| = |t|\|\mathbf{M}\| < \frac{1}{2}$, so (a) applies with $\mathbf{A} = t\mathbf{M}$ and gives, with $\mathbf{R}(t) = (\mathbf{I} - t\mathbf{M})^{-1} - \mathbf{I} - t\mathbf{M}$,

$$\|\mathbf{R}(t)\| \leq \frac{t^2\|\mathbf{M}\|^2}{1 - |t|\|\mathbf{M}\|} \leq 2t^2\|\mathbf{M}\|^2,$$

using $1 - |t|\|\mathbf{M}\| > \frac{1}{2}$. Hence $\|\mathbf{R}(t)\|/|t| \leq 2|t|\|\mathbf{M}\|^2 \rightarrow 0$ as $t \rightarrow 0$. So $t \mapsto (\mathbf{I} - t\mathbf{M})^{-1}$ is differentiable at $t = 0$ with derivative $\mathbf{M}$, which is the case $\mathbf{A}(t) = \mathbf{I} - t\mathbf{M}$ of the general formula for differentiating an inverse, proved in §06.

Exercise. C2: Distance to the singular matrices

Let $\mathbf{B} \in M_n(\mathbb{C})$ be invertible and let $\|\cdot\|$ be an induced matrix norm.

- (a) Prove that every singular $\mathbf{S} \in M_n(\mathbb{C})$ satisfies $\|\mathbf{B} - \mathbf{S}\| \geq 1/\|\mathbf{B}^{-1}\|$.
 (b) For $\mathbf{B} = \text{diag}(1, \varepsilon)$ with $0 < \varepsilon < 1$ and the 2-norm, exhibit a singular $\mathbf{S}$ attaining the bound.

(c) Hence say what $\|\mathbf{B}^{-1}\|_2$ measures.

Solution.

- (a) Suppose $\|\mathbf{B} - \mathbf{S}\| < 1/\|\mathbf{B}^{-1}\|$. Apply [Corollary 16.5.2](#) with $\mathbf{E} = \mathbf{S} - \mathbf{B}$: since $\|\mathbf{E}\| < 1/\|\mathbf{B}^{-1}\|$, the matrix $\mathbf{B} + \mathbf{E} = \mathbf{S}$ is invertible, contradicting the assumption that $\mathbf{S}$ is singular. Hence $\|\mathbf{B} - \mathbf{S}\| \geq 1/\|\mathbf{B}^{-1}\|$ for every singular $\mathbf{S}$.
- (b) $\mathbf{B}^{-1} = \text{diag}(1, 1/\varepsilon)$, a normal matrix, so its singular values are the moduli of its eigenvalues and $\|\mathbf{B}^{-1}\|_2 = 1/\varepsilon$ by [Theorem 16.3.3 \(c\)](#). The bound is therefore ε . Take $\mathbf{S} = \text{diag}(1, 0)$, which is singular; then $\mathbf{B} - \mathbf{S} = \text{diag}(0, \varepsilon)$ and $\|\mathbf{B} - \mathbf{S}\|_2 = \varepsilon$. So the bound is attained.
- (c) The reciprocal $1/\|\mathbf{B}^{-1}\|_2$ is a lower bound for the distance from $\mathbf{B}$ to the singular matrices, attained in the example. So $\|\mathbf{B}^{-1}\|_2$ is large exactly when $\mathbf{B}$ is close to being singular, which is the geometric content behind the condition number of §08.

Exercise. C3: Without $\|\mathbf{I}\| = 1$

Let $\|\cdot\|$ be any matrix norm on $M_n(\mathbb{C})$ in the sense of [Definition 16.3.1](#), not necessarily induced, and let $\|\mathbf{A}\| < 1$.

- (a) Prove that $\mathbf{I} - \mathbf{A}$ is invertible and that

$$\|(\mathbf{I} - \mathbf{A})^{-1}\| \leq \|\mathbf{I}\| + \frac{\|\mathbf{A}\|}{1 - \|\mathbf{A}\|}.$$

- (b) Verify that this bound is correct for the Frobenius norm on $M_2(\mathbb{C})$ with $\mathbf{A} = \mathbf{O}$, where the bound of [Theorem 16.5.1 \(b\)](#) fails.
- (c) Explain why the argument of (a) cannot be improved to $1/(1 - \|\mathbf{A}\|)$ for a general matrix norm.

Solution.

- (a) By [Theorem 16.4.1 \(b\)](#), which holds for every matrix norm, $\rho(\mathbf{A}) \leq \|\mathbf{A}\| < 1$; so [Theorem 10.10.6](#) applies and gives both the invertibility of $\mathbf{I} - \mathbf{A}$ and $\mathbf{S}_K \rightarrow (\mathbf{I} - \mathbf{A})^{-1}$. Submultiplicativity gives $\|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k$ for $k \geq 1$, so separating the term $k = 0$,

$$\|\mathbf{S}_K\| \leq \|\mathbf{I}\| + \sum_{k=1}^K \|\mathbf{A}\|^k \leq \|\mathbf{I}\| + \frac{\|\mathbf{A}\|}{1 - \|\mathbf{A}\|}.$$

As in the proof of [Theorem 16.5.1 \(b\)](#), [Corollary 16.2.4](#) and [Lemma 16.1.1](#) let the bound pass to the limit.

- (b) With $\mathbf{A} = \mathbf{O}$, the right-hand side is $\|\mathbf{I}_2\|_F + 0 = \sqrt{2}$, and the left-hand side is $\|\mathbf{I}_2\|_F = \sqrt{2}$. The bound holds, with equality; the bound $1/(1 - 0) = 1$ of [Theorem 16.5.1 \(b\)](#) does not.
- (c) The case $\mathbf{A} = \mathbf{O}$ already decides it: the left-hand side is $\|\mathbf{I}\|$, so any valid bound must be at least $\|\mathbf{I}\|$, and $1/(1 - \|\mathbf{A}\|) = 1$ there. A norm with $\|\mathbf{I}\| > 1$ therefore falsifies the sharper form at once, and by [Theorem 16.3.2 \(c\)](#) an induced norm never has $\|\mathbf{I}\| > 1$. This is exactly the difference between a matrix norm and an induced one.