

Chapter 2

Sensitivity, Errors and Norms

Two difficulties arise when we solve systems of linear equations or perform other matrix computations:

- (i) Errors in matrix elements.

Matrix elements may be contaminated with errors from measurements or previous computations, or they may simply not be known exactly. Merely inputting numbers into a computer or calculator can cause errors (e.g. when $1/3$ is stored as $.33333333$). To account for all these situations, we say that the matrix elements are afflicted with *uncertainties* or are *perturbed*. In general, perturbations of the inputs cause difficulties when the outputs are “sensitive” to changes in the inputs.

- (ii) Errors in algorithms.

Algorithms may not compute an exact solution, because computing the exact solution may not be necessary, may take too long, may require too much storage, or may not be practical. Furthermore, arithmetic operations in finite precision may not be performed exactly.

In this book, we focus on perturbations of inputs, and how these perturbations affect the outputs.

2.1 Sensitivity and Conditioning

In real life, *sensitive* means¹: “acutely affected by external stimuli”, “easily offended or emotionally hurt”, or “responsive to slight changes”. A sensitive person can be easily upset by small events, such as having to wait in line for a few minutes. Hardware can be sensitive: A very slight turn of a faucet may change the water from freezing cold to scalding hot. The slightest turn of the steering wheel when driving on an icy surface can send the car careening into a spin. Organs can be

¹The Concise Oxford English Dictionary

sensitive: Healthy skin may not even feel the prick of a needle, while it may cause extreme pain on burnt skin.

It is no different in mathematics. Steep functions, for instance, can be sensitive to small perturbations in the input.

Example. Let $f(x) = 9^x$ and consider the effect of a small perturbation to the input of $f(50) = 9^{50}$, such as

$$f(50.5) = \sqrt{9} 9^{50} = 3f(50).$$

Here a 1 percent change in the input causes a 300 percent change of the output. \square

Systems of linear equations are sensitive when a small modification in the matrix or the right-hand side causes a large change in the solution.

Example 2.1 The linear system $Ax = b$ with

$$A = \begin{pmatrix} 1/3 & 1/3 \\ 1/3 & .3 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

has the solution

$$x = \begin{pmatrix} -27 \\ 30 \end{pmatrix}.$$

However, a small change of the (2, 2) element from .3 to 1/3 results in the total loss of the solution, because the system $\tilde{A}x = b$ with

$$\tilde{A} = \begin{pmatrix} 1/3 & 1/3 \\ 1/3 & 1/3 \end{pmatrix},$$

has no solution. \blacksquare

A linear system like the one above whose solution is sensitive to small perturbations in the matrix is called *ill-conditioned*. Here is another example of ill-conditioning.

Example. The linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 + \epsilon \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad 0 < \epsilon \ll 1,$$

has the solution

$$x = \frac{1}{\epsilon} \begin{pmatrix} -2 - \epsilon \\ 2 \end{pmatrix}.$$

But changing the (2, 2) element of A from $1 + \epsilon$ to 1 results in the loss of the solution, because the linear system $\tilde{A}x = b$ with

$$\tilde{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

has no solution. This happens regardless of how small ϵ is. \square

An ill-conditioned linear system can also be sensitive to small perturbations in the right-hand side, as the next example shows.

Example 2.2 The linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 + \epsilon \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad 0 < \epsilon \ll 1,$$

has the solution $x = (2 \ 0)^T$. Changing the leading element in the right hand side from 2 to $2 + \epsilon$ alters the solution radically. That is, the system $A\tilde{x} = \tilde{b}$ with

$$\tilde{b} = \begin{pmatrix} 2 \\ 2 + \epsilon \end{pmatrix}$$

has the solution $\tilde{x} = (1 \ 1)^T$, which is completely different from x . \blacksquare

Important. Ill-conditioning of a linear system has nothing to do with how we compute the solution. Ill-conditioning is a property of the linear system. Hence there is, in general, nothing you can do about ill-conditioning.

In an ill-conditioned linear system, errors in the matrix or in the right-hand side can be amplified so that the errors in the solution are much larger. Our aim is to determine which properties of a linear system are responsible for ill-conditioning, and how one can quantify ill-conditioning.

2.2 Absolute and Relative Errors

To quantify ill-conditioning, we need to assess the size of errors.

Example. Suppose you have $y = 10$ dollars in your bank account. But the bank makes a mistake, and subtracts 20 dollars from your account, so that your account now has a negative balance of $\tilde{y} = -10$ dollars. The account is overdrawn, and all kinds of bad consequences ensue.

Now imagine this happens to Bill Gatez. He has $g = 10^{11}$ dollars in his account, and if the bank subtracts by mistake 20 dollars from his balance, he still has $\tilde{g} = 10^{11} - 20$ dollars.

In both cases, the bank makes the same error,

$$y - \tilde{y} = g - \tilde{g} = 20.$$

But you are much worse off than Bill Gatez. You are in now debt, while Bill Gatez has so much money, he may not even notice the error. In your case, the error is larger than your credit; while in Bill Gatez's case, the error is only a tiny part of his fortune.

How can we express mathematically that the bank's error is much worse for you than for Bill Gatez? We can compare the error to the balance in your account: $\frac{y-\tilde{y}}{y} = 2$. This shows that the error is twice as large as your original balance. For Bill Gatez we obtain $\frac{y-\tilde{y}}{y} = 2 \cdot 10^{-10}$, so that the error is only a tiny fraction of his balance. Now it's clear that the bank's error is much more serious for you than it is for Bill Gatez. \square

A difference like $y - \tilde{y}$ measures an *absolute error*, while $\frac{y-\tilde{y}}{y}$ and $\frac{y-\tilde{y}}{\tilde{y}}$ measure *relative errors*. We use relative errors if we want to know how large the error is when compared to the original quantity. Often we are not interested in the signs of the errors, so we consider absolute values $\frac{|y-\tilde{y}|}{|y|}$ and $\frac{|y-\tilde{y}|}{|\tilde{y}|}$.

Definition 2.3. *If the scalar \tilde{x} is an approximation to the scalar x then we call $|x - \tilde{x}|$ an absolute error. If $x \neq 0$ then we call $\frac{|x-\tilde{x}|}{|x|}$ a relative error. If $\tilde{x} \neq 0$ then $\frac{|x-\tilde{x}|}{|\tilde{x}|}$ is also a relative error.*

A relative error close to or larger than 1 means that an approximation is totally inaccurate. To see this, suppose that $\frac{|x-\tilde{x}|}{|x|} \geq 1$. Then $|x - \tilde{x}| \geq |x|$, which means that the absolute error is larger than the quantity we are trying to compute. If we approximate $x = 0$ by $\tilde{x} \neq 0$, however small, then the relative error is always $\frac{|0-\tilde{x}|}{|\tilde{x}|} = 1$. Thus, the only approximation to 0 that has a small relative error is 0 itself.

In contrast to an absolute error, a relative error can give information about how many digits two numbers have in common. As a rule of thumb, if

$$\frac{|x - \tilde{x}|}{|x|} \leq 5 \cdot 10^{-d},$$

then we say that the numbers x and \tilde{x} agree to d decimal digits.

Example. If $x = 1$ and $\tilde{x} = 1.003$ then $\frac{|x-\tilde{x}|}{|x|} = 3 \cdot 10^{-3} \leq 5 \cdot 10^{-3}$, so that x and \tilde{x} agree to three decimal digits.

According to the above definition, the numbers $x = 1$ and $\hat{x} = .997$ also agree to three decimal digits because $\frac{|x-\hat{x}|}{|x|} = 3 \cdot 10^{-3} \leq 5 \cdot 10^{-3}$. \square

2.3 Floating Point Arithmetic.

Many computations in science and engineering are carried out in floating point arithmetic, where all real numbers are represented by a finite set of floating point numbers. All floating point numbers are stored in the same, fixed number of bits regardless of how small or how large they are. Many computers are based on IEEE double precision arithmetic where a floating point number is stored in 64 bits.

The floating point representation \hat{x} of a real number x differs from x by a

factor close to one, and satisfies²

$$\hat{x} = x(1 + \epsilon_x), \quad \text{where } |\epsilon_x| \leq u.$$

Here u is the “unit round off” that specifies the accuracy of floating point arithmetic. In IEEE double precision arithmetic $u = 2^{-53} \approx 10^{-16}$. If $x \neq 0$ then

$$\frac{|x - \hat{x}|}{|x|} \leq |\epsilon_x|.$$

This means conversion to floating point representation causes relative errors. We say that a floating point number \hat{x} is a *relative perturbation* of the exact number x .

Since floating point arithmetic causes relative perturbations in the inputs, it makes sense to determine relative – rather than absolute – errors in the output. As a consequence, we will pay more attention to relative errors than to absolute errors.

The question now is how elementary arithmetic operations are affected when they are performed on numbers contaminated with small relative perturbations, such as floating point numbers. We start with subtraction.

2.4 Conditioning of Subtraction

Subtraction is the only elementary operation that is sensitive to relative perturbations. The analogy below of the captain and the battleship can help us understand why.

Example. To find out how much he weighs, the captain first weighs the battleship with himself on it, and then he steps off the battle ship and weighs it without himself on it. At the end he subtracts the two weights. Intuitively we have a fuzzy feeling for why this should not give an accurate estimate for the captain’s weight. Below we explain why.

Let \tilde{x} represent the weight of the battleship plus captain, and \tilde{y} the weight of the battleship without the captain, where

$$\tilde{x} = 112233\underline{9}, \quad \tilde{y} = 112233\underline{7}.$$

Due to the limited precision of the scale, the underlined digits are uncertain and may be in error. The captain computes as his weight $\tilde{x} - \tilde{y} = \underline{2}$. This difference is totally inaccurate because it is derived from uncertainties, while all the accurate digits have cancelled out. This is an example of “catastrophic cancellation”. \square

Catastrophic cancellation occurs when we subtract two numbers that are uncertain, and when the difference between these two numbers is as small as the uncertainties. We will now show that catastrophic cancellation occurs when subtraction is ill-conditioned with regard to relative errors.

Let \tilde{x} be a perturbation of the scalar x , and \tilde{y} a perturbation of the scalar y . We bound the error in $\tilde{x} - \tilde{y}$ in terms of the errors in \tilde{x} and \tilde{y} .

²We assume that x lies in the range of normalized floating point numbers, so that no underflow or overflow occurs.

Absolute Error. From

$$|(\tilde{x} - \tilde{y}) - (x - y)| \leq |\tilde{x} - x| + |\tilde{y} - y|$$

we see that the absolute error in the difference is bounded by the absolute errors in the inputs. Therefore we say that subtraction is *well-conditioned in the absolute sense*. In the above example, the last digit of \tilde{x} and \tilde{y} is uncertain, so that $|\tilde{x} - x| \leq 9$ and $|\tilde{y} - y| \leq 9$, and the absolute error is bounded by $|(\tilde{x} - \tilde{y}) - (x - y)| \leq 18$.

Relative Error. However the relative error in the difference can be much larger than the relative error in the inputs. In the above example we can estimate the relative error from

$$\frac{|(\tilde{x} - \tilde{y}) - (x - y)|}{|\tilde{x} - \tilde{y}|} \leq 18/2 = 9,$$

which suggests that the computed difference $\tilde{x} - \tilde{y}$ is completely inaccurate.

In general, this severe loss of accuracy can occur when we subtract two nearly equal numbers that are in error. The bound in Fact 2.4 below shows that subtraction can be *ill-conditioned in the relative sense* if the difference is much smaller in magnitude than the inputs.

Fact 2.4 (Relative Conditioning of Subtraction) Let x, y, \tilde{x} , and \tilde{y} be scalars. If $x \neq 0, y \neq 0$ and $x \neq y$, then

$$\underbrace{\frac{|(\tilde{x} - \tilde{y}) - (x - y)|}{|x - y|}}_{\text{relative error in output}} \leq \kappa \underbrace{\max \left\{ \frac{|\tilde{x} - x|}{|x|}, \frac{|\tilde{y} - y|}{|y|} \right\}}_{\text{relative error in input}},$$

where

$$\kappa = \frac{|x| + |y|}{|x - y|}.$$

The positive number κ is a *relative condition number for subtraction* because it quantifies how relative errors in the input can be amplified, and how sensitive subtraction can be to relative errors in the input. When $\kappa \gg 1$, subtraction is ill-conditioned in the relative sense, and is called catastrophic cancellation.

If we don't know x, y or $x - y$, but want an estimate of the condition number we can use instead the bound

$$\frac{|(\tilde{x} - \tilde{y}) - (x - y)|}{|\tilde{x} - \tilde{y}|} \leq \tilde{\kappa} \max \left\{ \frac{|\tilde{x} - x|}{|\tilde{x}|}, \frac{|\tilde{y} - y|}{|\tilde{y}|} \right\}, \quad \tilde{\kappa} = \frac{|\tilde{x}| + |\tilde{y}|}{|\tilde{x} - \tilde{y}|},$$

provided $\tilde{x} \neq 0, \tilde{y} \neq 0$ and $\tilde{x} \neq \tilde{y}$,

Remark 2.5. *Catastrophic cancellation does not occur when we subtract two numbers that are exact.*

Catastrophic cancellation can only occur when we subtract two numbers that have relative errors. It is the amplification of these relative errors that leads to catastrophe.

Exercises

1. Relative Conditioning of Multiplication.

Let $x, y, \tilde{x}, \tilde{y}$ be nonzero scalars. Show:

$$\left| \frac{xy - \tilde{x}\tilde{y}}{xy} \right| \leq (2 + \epsilon)\epsilon, \quad \text{where } \epsilon = \max \left\{ \left| \frac{x - \tilde{x}}{x} \right|, \left| \frac{y - \tilde{y}}{y} \right| \right\},$$

and if $\epsilon \leq 1$ then

$$\left| \frac{xy - \tilde{x}\tilde{y}}{xy} \right| \leq 3\epsilon.$$

Therefore, if the relative error in the inputs is not too large then the condition number of multiplication is at most 3. We can conclude that multiplication is well-conditioned in the relative sense, provided the inputs have small relative perturbations.

2. Relative Conditioning of Division.

Let $x, y, \tilde{x}, \tilde{y}$ be nonzero scalars, and

$$\epsilon = \max \left\{ \left| \frac{x - \tilde{x}}{x} \right|, \left| \frac{y - \tilde{y}}{y} \right| \right\}.$$

Show: If $\epsilon < 1$ then

$$\left| \frac{x/y - \tilde{x}/\tilde{y}}{x/y} \right| \leq \frac{2\epsilon}{1 - \epsilon},$$

and if $\epsilon < 1/2$ then

$$\left| \frac{x/y - \tilde{x}/\tilde{y}}{x/y} \right| \leq 4\epsilon.$$

Therefore, if the relative error in the operands is not too large then the condition number of division is at most 4. We can conclude that division is well-conditioned in the relative sense, provided the inputs have small relative perturbations.

2.5 Vector Norms

In the context of linear system solution, the error in the solution constitutes a vector. If we do not want to pay attention to individual components of the error, perhaps because there are too many components, then we can combine all errors into a single number. This is akin to a grade point average which combines all grades into a single number. Mathematically this “combining” is accomplished by norms. We start with vector norms, which measure the length of a vector.

Definition 2.6. A vector norm $\|\cdot\|$ is a function from \mathbb{C}^n to \mathbb{R} with three properties:

N1: $\|x\| \geq 0$ for all $x \in \mathbb{C}^n$, and $\|x\| = 0$ if and only if $x = 0$.

N2: $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{C}^n$ (triangle inequality)

N3: $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{C}$, $x \in \mathbb{C}^n$.

The vector p-norms below are useful for computational purposes, as well as analysis.

Fact 2.7 (Vector p-Norms) Let $x \in \mathbb{C}^n$ with elements $x = (x_1 \ \dots \ x_n)^T$. The p -norm

$$\|x\|_p = \left(\sum_{j=1}^n |x_j|^p \right)^{1/p}, \quad p \geq 1,$$

is a vector norm.

Example.

- If e_j is a canonical vector then $\|e_j\|_p = 1$ for $p \geq 1$.
- If $e = (1 \ 1 \ \dots \ 1)^T \in \mathbb{R}^n$ then

$$\|e\|_1 = n, \quad \|e\|_\infty = 1, \quad \|e\|_p = n^{1/p}, \quad 1 < p < \infty.$$

□

The three p-norms below are the most popular, because they are easy to compute.

- One norm: $\|x\|_1 = \sum_{j=1}^n |x_j|$
- Two (or Euclidean) norm: $\|x\|_2 = \sqrt{\sum_{j=1}^n |x_j|^2} = \sqrt{x^*x}$
- Infinity (or maximum) norm: $\|x\|_\infty = \max_{1 \leq j \leq n} |x_j|$

Example. If $x = (1 \ 2 \ \dots \ n)^T \in \mathbb{R}^n$ then

$$\|x\|_1 = \frac{1}{2}n(n+1), \quad \|x\|_2 = \sqrt{\frac{1}{6}n(n+1)(2n+1)}, \quad \|x\|_\infty = n.$$

□

The inequalities below bound inner products in terms of norms.

Fact 2.8 Let $x, y \in \mathbb{C}^n$. Then

Hölder inequality: $|x^*y| \leq \|x\|_1 \|y\|_\infty$

Cauchy-Schwartz inequality: $|x^*y| \leq \|x\|_2 \|y\|_2$.

Moreover, $|x^*y| = \|x\|_2 \|y\|_2$ if and only if x and y are multiples of each other.

Example. Let $x \in \mathbb{C}^n$ with elements $x = (x_1 \ \cdots \ x_n)$. The Hölder inequality and Cauchy-Schwartz inequality imply, respectively,

$$\left| \sum_{i=1}^n x_i \right| \leq n \max_{1 \leq i \leq n} |x_i|, \quad \left| \sum_{i=1}^n x_i \right| \leq \sqrt{n} \|x\|_2.$$

□

Definition 2.9. A non-zero vector $x \in \mathbb{C}^n$ is called unit-norm vector in the $\|\cdot\|$ norm if $\|x\| = 1$. The vector $x/\|x\|$ has unit norm.

Example. Let e be the $n \times 1$ vector of all ones. Then

$$1 = \|e\|_\infty = \left\| \frac{1}{n} e \right\|_1 = \left\| \frac{1}{\sqrt{n}} e \right\|_2.$$

The canonical vectors e_i have unit norm in any p-norm. □

Normwise Errors. We determine how much information the norm of an error gives about individual, componentwise errors.

Definition 2.10. If \tilde{x} is an approximation to a vector $x \in \mathbb{C}^n$, then $\|x - \tilde{x}\|$ is a normwise absolute error. If $x \neq 0$ or $\tilde{x} \neq 0$ then $\frac{\|x - \tilde{x}\|}{\|x\|}$ and $\frac{\|x - \tilde{x}\|}{\|\tilde{x}\|}$ are normwise relative errors.

How much do we lose when we replace componentwise errors by normwise errors? For vectors $x, \tilde{x} \in \mathbb{C}^n$, the infinity norm is equal to the largest absolute error,

$$\|x - \tilde{x}\|_\infty = \max_{1 \leq j \leq n} |x_j - \tilde{x}_j|.$$

For the one and two norms we have

$$\max_{1 \leq j \leq n} |x_j - \tilde{x}_j| \leq \|x - \tilde{x}\|_1 \leq n \max_{1 \leq j \leq n} |x_j - \tilde{x}_j|$$

and

$$\max_{1 \leq j \leq n} |x_j - \tilde{x}_j| \leq \|x - \tilde{x}\|_2 \leq \sqrt{n} \max_{1 \leq j \leq n} |x_j - \tilde{x}_j|.$$

Hence absolute errors in the one and two norms can overestimate the worst componentwise error by a factor that depends on the vector length n .

Unfortunately, normwise relative errors give much less information about componentwise relative errors.

Example. Let \tilde{x} be an approximation to a vector x where

$$x = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}, \quad 0 < \epsilon \ll 1, \quad \tilde{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The normwise relative error $\frac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty} = \epsilon$ is small. However, the componentwise relative error in the second component, $\frac{|x_2 - \tilde{x}_2|}{|x_2|} = 1$, shows that \tilde{x}_2 is a totally inaccurate approximation to x_2 in the relative sense. \square

The preceding example illustrates that a normwise relative error can be small, even if individual vector elements have a large relative error. In the infinity norm, for example, the normwise relative error only bounds the relative error corresponding to a component of x with the largest magnitude. To see this, let $|x_k| = \|x\|_\infty$. Then

$$\frac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty} = \frac{\max_{1 \leq j \leq n} |x_j - \tilde{x}_j|}{|x_k|} \geq \frac{|x_k - \tilde{x}_k|}{|x_k|}.$$

For the normwise relative errors in the one and two norm we incur additional factors that depend on the vector length n ,

$$\frac{\|x - \tilde{x}\|_1}{\|x\|_1} \geq \frac{1}{n} \frac{|x_k - \tilde{x}_k|}{|x_k|}, \quad \frac{\|x - \tilde{x}\|_2}{\|x\|_2} \geq \frac{1}{\sqrt{n}} \frac{|x_k - \tilde{x}_k|}{|x_k|}.$$

Therefore normwise relative errors give no information about relative errors in components of smaller magnitude. If relative errors in individual vector components are important, then do not use normwise errors.

Remark 2.11. *When measuring the normwise relative error of an approximation \tilde{x} to x , the question is which error to measure, $\frac{\|x - \tilde{x}\|}{\|x\|}$ or $\frac{\|x - \tilde{x}\|}{\|\tilde{x}\|}$? If $\|\tilde{x}\| \approx \|x\|$ then the two errors are about the same. In general, the two errors are related as follows. Let $x \neq 0$, $\tilde{x} \neq 0$, and*

$$\epsilon = \frac{\|x - \tilde{x}\|}{\|x\|}, \quad \tilde{\epsilon} = \frac{\|x - \tilde{x}\|}{\|\tilde{x}\|}.$$

If $\epsilon < 1$ then

$$\frac{\epsilon}{1 + \epsilon} \leq \tilde{\epsilon} \leq \frac{\epsilon}{1 - \epsilon}.$$

This follows from $\tilde{\epsilon} = \epsilon \|x\| / \|\tilde{x}\|$ and $1 - \tilde{\epsilon} \leq \|x\| / \|\tilde{x}\| \leq 1 + \tilde{\epsilon}$.

Exercises

- (i) Let $x \in \mathbb{C}^n$. Prove: $\|x\|_2 \leq \sqrt{\|x\|_1 \|x\|_\infty}$.
- (ii) For each equality below, determine a class of vectors that satisfy the equality:

$$\|x\|_1 = \|x\|_\infty, \quad \|x\|_1 = n \|x\|_\infty, \quad \|x\|_2 = \|x\|_\infty, \quad \|x\|_2 = \sqrt{n} \|x\|_\infty.$$
- (iii) Give examples of vectors $x, y \in \mathbb{C}^n$ with $x^*y \neq 0$ for which $|x^*y| = \|x\|_1 \|y\|_\infty$. Also find examples for $|x^*y| = \|x\|_2 \|y\|_2$.
- (iv) The p-norm of a vector does not change when the vector is permuted. Prove: If P is a permutation matrix then $\|Px\|_p = \|x\|_p$.

- (v) The two norm vector does not change when the vector is multiplied by a unitary matrix.
 Prove: If the matrix $V \in \mathbb{C}^{n \times n}$ is unitary then $\|Vx\|_2 = \|x\|_2$ for any vector $x \in \mathbb{C}^n$.
- (vi) Prove: If $Q \in \mathbb{C}^{n \times n}$ is unitary and $x \in \mathbb{C}^n$ is a non-zero vector with $Qx = \lambda x$, where λ is a scalar, then $|\lambda| = 1$.

1. Verify that the vector p-norms do indeed satisfy the three properties of a vector norm in Definition 2.6.
2. Reverse Triangle Inequality:
 Let $x, y \in \mathbb{C}^n$ and $\|\cdot\|$ a vector norm. Prove: $\left| \|x\| - \|y\| \right| \leq \|x - y\|$.
3. Theorem of Pythagoras.
 Prove: If $x, y \in \mathbb{C}^n$ and $x^*y = 0$ then $\|x \pm y\|_2^2 = \|x\|_2^2 + \|y\|_2^2$.
4. Parallelogram Equality.
 Let $x, y \in \mathbb{C}^n$. Prove that $\|x + y\|_2^2 + \|x - y\|_2^2 = 2(\|x\|_2^2 + \|y\|_2^2)$
5. Polarization Identity.
 Let $x, y \in \mathbb{C}^n$. Prove that $\Re(x^*y) = \frac{1}{4}(\|x + y\|_2^2 - \|x - y\|_2^2)$, where $\Re(\alpha)$ is the real part of a complex number α .
6. Let $x \in \mathbb{C}^n$. Prove:

$$\begin{aligned} \|x\|_2 &\leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \\ \|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty \\ \|x\|_\infty &\leq \|x\|_1 \leq n\|x\|_\infty \end{aligned}$$

7. Let $A \in \mathbb{C}^{n \times n}$ be nonsingular. Show that $\|x\|_A = \|Ax\|_p$ is a vector norm.

2.6 Matrix Norms

We need to separate matrices from vectors inside the norms. To see this, let $Ax = b$ be a nonsingular linear system, and $A\tilde{x} = \tilde{b}$ a perturbed system. The normwise absolute error is $\|x - \tilde{x}\| = \|A^{-1}(b - \tilde{b})\|$. In order to isolate the perturbation and derive a bound of the form $\|A^{-1}\| \|b - \tilde{b}\|$, we have to define a norm for matrices.

Definition 2.12. A matrix norm $\|\cdot\|$ is a function from $\mathbb{C}^{m \times n}$ to \mathbb{R} with three properties:

- N1:** $\|A\| \geq 0$ for all $A \in \mathbb{C}^{n \times m}$. And $\|A\| = 0$ if and only if $A = 0$.
- N2:** $\|A + B\| \leq \|A\| + \|B\|$ for all $A, B \in \mathbb{C}^{n \times m}$ (triangle inequality)
- N3:** $\|\alpha A\| = |\alpha| \|A\|$ for all $\alpha \in \mathbb{C}$, $A \in \mathbb{C}^{n \times m}$.

Because of the triangle inequality, matrix norms are well-conditioned, in the absolute sense and in the relative sense.

Fact 2.13 If $A, E \in \mathbb{C}^{m \times n}$ then $|\|A + E\| - \|A\|| \leq \|E\|$.

Proof. The triangle inequality implies $\|A + E\| \leq \|A\| + \|E\|$, hence $\|A + E\| - \|A\| \leq \|E\|$. Similarly $\|A\| = \|(A + E) - E\| \leq \|A + E\| + \|E\|$, so that $-\|E\| \leq \|A + E\| - \|A\|$. The result follows from

$$-\|E\| \leq \|A + E\| - \|A\| \leq \|E\|.$$

□

The matrix p-norms below are based on the vector p-norms, and measure how much a matrix can stretch a unit-norm vector.

Fact 2.14 (Matrix p-Norms) Let $A \in \mathbb{C}^{n \times m}$. The *p*-norm

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p=1} \|Ax\|_p$$

is a matrix norm.

Remark 2.15. The matrix *p*-norms are extremely useful because they satisfy the following submultiplicative inequality.

Let $A \in \mathbb{C}^{m \times n}$ and $y \in \mathbb{C}^n$, then

$$\|Ay\|_p \leq \|A\|_p \|y\|_p.$$

This is clearly true for $y = 0$, and for $y \neq 0$ it follows from

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \geq \frac{\|Ay\|_p}{\|y\|_p}.$$

The matrix one-norm is equal to the maximal absolute column sum.

Fact 2.16 (One Norm) Let $A \in \mathbb{C}^{m \times n}$. Then

$$\|A\|_1 = \max_{1 \leq j \leq n} \|Ae_j\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|.$$

Proof.

- The definition of p-norms implies

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 \geq \|Ae_j\|_1, \quad 1 \leq j \leq n.$$

Hence $\|A\|_1 \geq \max_{1 \leq j \leq n} \|Ae_j\|_1$.

- Let $y = (y_1 \ \dots \ y_n)^T$ be a vector with $\|A\|_1 = \|Ay\|_1$ and $\|y\|_1 = 1$. Viewing the matrix vector product Ay as a linear combination of columns of A , see §1.5, and applying the triangle inequality for vector norms gives

$$\begin{aligned} \|A\|_1 = \|Ay\|_1 &= \|y_1 A e_1 + \dots + y_n A e_n\|_1 \leq |y_1| \|A e_1\|_1 + \dots + |y_n| \|A e_n\|_1 \\ &\leq (|y_1| + \dots + |y_n|) \max_{1 \leq j \leq n} \|A e_j\|_1. \end{aligned}$$

From $|y_1| + \dots + |y_n| = \|y\|_1 = 1$ follows $\|A\|_1 \leq \max_{1 \leq j \leq n} \|A e_j\|_1$.

□

The matrix infinity norm is equal to the maximal absolute row sum.

Fact 2.17 (Infinity Norm) Let $A \in \mathbb{C}^{m \times n}$. Then

$$\|A\|_\infty = \max_{1 \leq i \leq m} \|A^* e_i\|_1 = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|.$$

Proof. Denote the rows of A by $r_i^* = e_i^* A$, and let r_k have the largest one-norm, $\|r_k\|_1 = \max_{1 \leq i \leq m} \|r_i\|_1$.

- Let y be a vector with $\|A\|_\infty = \|Ay\|_\infty$ and $\|y\|_\infty = 1$. Then

$$\|A\|_\infty = \|Ay\|_\infty = \max_{1 \leq i \leq m} |r_i^* y| \leq \max_{1 \leq i \leq m} \|r_i\|_1 \|y\|_\infty = \|r_k\|_1,$$

where the inequality follows from Fact 2.8. Hence $\|A\|_\infty \leq \max_{1 \leq i \leq m} \|r_i\|_1$.

- For any vector y with $\|y\|_\infty = 1$ we have $\|A\|_\infty \geq \|Ay\|_\infty \geq |r_k^* y|$. Now we show how to choose the elements of y such that $|r_k^* y| = \|r_k\|_1$. Let $r_k^* = (\rho_1 \ \dots \ \rho_n)$ be the elements of r_k^* . Choose the elements of y such that $\rho_j y_j = |\rho_j|$. That is, if $\rho_j = 0$ then $y_j = 0$, and otherwise $y_j = |\rho_j|/\rho_j$. Then $\|y\|_\infty = 1$, and $|r_k^* y| = \sum_{j=1}^n \rho_j y_j = \sum_{j=1}^n |\rho_j| = \|r_k\|_1$. Hence

$$\|A\|_\infty \geq |r_k^* y| = \|r_k\|_1 = \max_{1 \leq i \leq m} \|r_i\|_1.$$

□

The p -norms satisfy the following *submultiplicative inequality*.

Fact 2.18 (Norm of a Product) If $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{n \times p}$ then

$$\|AB\|_p \leq \|A\|_p \|B\|_p.$$

Proof. Let $x \in \mathbb{C}^p$ such that $\|AB\|_p = \|ABx\|_p$ and $\|x\|_p = 1$. Applying Remark 2.15 twice gives

$$\|AB\|_p = \|ABx\|_p \leq \|A\|_p \|Bx\|_p \leq \|A\|_p \|B\|_p \|x\|_p = \|A\|_p \|B\|_p.$$

□

Since the computation of the two norm is more involved, we postpone it until later. However, even without knowing how to compute it, we can still derive several useful properties of the two norm. If x is a column vector then $\|x\|_2^2 = x^*x$. We show below that an analogous property holds for matrices. We also show that a matrix and its transpose have the same two norm.

Fact 2.19 (Two Norm) Let $A \in \mathbb{C}^{m \times n}$. Then

$$\|A^*\|_2 = \|A\|_2, \quad \|A^*A\|_2 = \|A\|_2^2.$$

Proof. The definition of the two norm implies that for some $x \in \mathbb{C}^n$ with $\|x\|_2 = 1$ we have $\|A\|_2 = \|Ax\|_2$. The definition of the *vector* two norm implies

$$\|A\|_2^2 = \|Ax\|_2^2 = x^*A^*Ax \leq \|x\|_2\|A^*Ax\|_2 \leq \|A^*A\|_2,$$

where the first inequality follows from the Cauchy-Schwartz inequality in Fact 2.8 and the second inequality from the two norm of A^*A . Hence $\|A\|_2^2 \leq \|A^*A\|_2$. Fact 2.18 implies $\|A^*A\|_2 \leq \|A^*\|_2\|A\|_2$. As a consequence

$$\|A\|_2^2 \leq \|A^*A\|_2 \leq \|A^*\|_2\|A\|_2, \quad \|A\|_2 \leq \|A^*\|_2.$$

The same reasoning applied to AA^* gives

$$\|A^*\|_2^2 \leq \|AA^*\|_2 \leq \|A\|_2\|A^*\|_2, \quad \|A^*\|_2 \leq \|A\|_2.$$

Therefore $\|A^*\|_2 = \|A\|_2$ and $\|A^*A\|_2 = \|A\|_2^2$. □

If we omit a piece of a matrix, the norm does not increase but it can decrease.

Fact 2.20 (Norm of Submatrix) Let $A \in \mathbb{C}^{m \times n}$. If B is a submatrix of A then $\|B\|_p \leq \|A\|_p$.

Exercises

- (i) Let $D \in \mathbb{C}^{n \times n}$ is a diagonal matrix with diagonal elements d_{jj} . Show that $\|D\|_p = \max_{1 \leq j \leq n} |d_{jj}|$.
- (ii) Let $A \in \mathbb{C}^{n \times n}$ be nonsingular. Show: $\|A\|_p\|A^{-1}\|_p \geq 1$.
- (iii) Show: If P is a permutation matrix then $\|P\|_p = 1$.
- (iv) Let $P \in \mathbb{R}^{m \times m}$, $Q \in \mathbb{R}^{n \times n}$ be permutation matrices and $A \in \mathbb{C}^{m \times n}$. Show: $\|PAQ\|_p = \|A\|_p$.
- (v) Let $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$ be unitary. Show: $\|U\|_2 = \|V\|_2 = 1$, and $\|UBV\|_2 = \|B\|_2$ for any $B \in \mathbb{C}^{m \times n}$.
- (vi) Let $x \in \mathbb{C}^n$. Show: $\|x^*\|_2 = \|x\|_2$ without using Fact 2.19.
- (vii) Let $x \in \mathbb{C}^n$. Is $\|x\|_1 = \|x^*\|_1$, and $\|x\|_\infty = \|x^*\|_\infty$? Why or why not?

(viii) Let $x \in \mathbb{C}^n$ be the vector of all ones. Determine

$$\|x\|_1, \quad \|x^*\|_1, \quad \|x\|_\infty, \quad \|x^*\|_\infty, \quad \|x\|_2, \quad \|x^*\|_2.$$

(ix) For each of the two equalities, determine a class of matrices A that satisfy the equality: $\|A\|_\infty = \|A\|_1$, and $\|A\|_\infty = \|A\|_1 = \|A\|_2$.

(x) Let $A \in \mathbb{C}^{m \times n}$. Then $\|A\|_\infty = \|A^*\|_1$.

1. Verify that the matrix p-norms do indeed satisfy the three properties of a matrix norm in Definition 2.12.
2. Let $A \in \mathbb{C}^{m \times n}$. Prove:

$$\begin{aligned} \max_{i,j} |a_{ij}| &\leq \|A\|_2 \leq \sqrt{mn} \max_{i,j} |a_{ij}| \\ \frac{1}{\sqrt{n}} \|A\|_\infty &\leq \|A\|_2 \leq \sqrt{m} \|A\|_\infty \\ \frac{1}{\sqrt{m}} \|A\|_1 &\leq \|A\|_2 \leq \sqrt{n} \|A\|_1 \end{aligned}$$

3. Norms of Outer Products.

Let $x \in \mathbb{C}^m$ and $y \in \mathbb{C}^n$. Show:

$$\|xy^*\|_2 = \|x\|_2 \|y\|_2, \quad \|xy^*\|_\infty = \|x\|_\infty \|y\|_1.$$

4. Given an approximate solution z , here is the matrix perturbation of smallest two norm that 'realizes' z , in the sense that the perturbed system has z as a solution.

Let $A \in \mathbb{C}^{n \times n}$, $Ax = b$ and $z \neq 0$. Show: Among all matrices E with $(A + E)z = b$ the matrix $E_0 = (b - Az)z^\dagger$ has smallest two norm, where $z^\dagger = (z^* z)^{-1} z^*$.

5. Norms of Idempotent Matrices.

Show: If $A \neq 0$ is idempotent then $\|A\|_p \geq 1$. If A is also Hermitian then $\|A\|_2 = 1$.

6. Let $A \in \mathbb{C}^{n \times n}$. Show that, among all Hermitian matrices, $\frac{1}{2}(A + A^*)$ is the matrix that is closest to A in the two norm.

2.7 Conditioning of Matrix Addition and Multiplication

We derive normwise relative bounds for matrix addition and subtraction, as well as for matrix multiplication.

Fact 2.21 (Matrix Addition) Let $U, V, \tilde{U}, \tilde{V} \in \mathbb{C}^{m \times n}$ such that $U, V, U + V \neq 0$. Then

$$\frac{\|\tilde{U} + \tilde{V} - (U + V)\|_p}{\|U + V\|_p} \leq \frac{\|U\|_p + \|V\|_p}{\|U + V\|_p} \max\{\epsilon_U, \epsilon_V\},$$

where

$$\epsilon_U = \frac{\|\tilde{U} - U\|_p}{\|U\|_p}, \quad \epsilon_V = \frac{\|\tilde{V} - V\|_p}{\|V\|_p}.$$

Proof. The triangle inequality implies

$$\begin{aligned} \|\tilde{U} + \tilde{V} - (U + V)\|_p &\leq \|\tilde{U} - U\|_p + \|\tilde{V} - V\|_p = \|U\|_p \epsilon_U + \|V\|_p \epsilon_V \\ &\leq (\|U\|_p + \|V\|_p) \max\{\epsilon_U, \epsilon_V\}. \end{aligned}$$

□

The condition number for adding, or subtracting, the matrices U and V is $(\|U\|_p + \|V\|_p)/\|U + V\|_p$. It is analogous to the condition number for scalar subtraction in Fact 2.4. If $\|U\|_p + \|V\|_p \approx \|U + V\|_p$ then matrix addition $U + V$ is well-conditioned in the normwise relative sense. But if $\|U\|_p + \|V\|_p \gg \|U + V\|_p$ then the matrix addition $U + V$ is ill-conditioned in the normwise relative sense.

Fact 2.22 (Matrix Multiplication) Let $U, \tilde{U} \in \mathbb{C}^{m \times n}$, $V, \tilde{V} \in \mathbb{C}^{n \times p}$ such that $U, V, UV \neq 0$. Then

$$\frac{\|\tilde{U}\tilde{V} - UV\|_p}{\|UV\|_p} \leq \frac{\|U\|_p \|V\|_p}{\|UV\|_p} (\epsilon_U + \epsilon_V + \epsilon_U \epsilon_V),$$

where

$$\epsilon_U = \frac{\|\tilde{U} - U\|_p}{\|U\|_p}, \quad \epsilon_V = \frac{\|\tilde{V} - V\|_p}{\|V\|_p}.$$

Proof. If $\tilde{U} = U + E$ and $\tilde{V} = V + F$ then

$$\tilde{U}\tilde{V} - UV = (U + E)(V + F) - UV = UF + EV + EF.$$

Now take norms, apply the triangle inequality and divide by $\|UV\|_p$. □

Fact 2.22 shows that the normwise relative condition number for multiplying matrices U and V is $\|U\|_p \|V\|_p / \|UV\|_p$. If $\|U\|_p \|V\|_p \approx \|UV\|_p$ then the matrix multiplication UV is well-conditioned in the normwise relative sense. However, if $\|U\|_p \|V\|_p \gg \|UV\|_p$ then the matrix multiplication UV is ill-conditioned in the normwise relative sense.

Exercises

- (i) What is the two norm condition number of a product where one of the matrices is unitary?
- (ii) Normwise absolute condition number for matrix multiplication when one of the matrices is perturbed.

Let $U, V \in \mathbb{C}^{n \times n}$, and U be nonsingular. Show

$$\frac{\|F\|_p}{\|U^{-1}\|_p} \leq \|U(V + F) - UV\|_p \leq \|U\|_p \|F\|_p.$$

(iii) Here is a bound on the normwise relative error for matrix multiplication with regard to the perturbed product.

Let $U \in \mathbb{C}^{m \times n}$, $V \in \mathbb{C}^{n \times m}$. Show: If $(U + E)(V + F) \neq 0$ then

$$\frac{\|(U + E)(V + F) - UV\|_p}{\|(U + E)(V + F)\|_p} \leq \frac{\|U + E\|_p \|V + F\|_p}{\|(U + E)(V + F)\|_p} (\epsilon_U + \epsilon_V + \epsilon_U \epsilon_V),$$

where

$$\epsilon_U = \frac{\|E\|_p}{\|U + E\|_p}, \quad \epsilon_V = \frac{\|F\|_p}{\|V + F\|_p}.$$

2.8 Conditioning of Matrix Inversion

We determine the sensitivity of the inverse to perturbations in the matrix.

We start by bounding the inverse of a perturbed identity matrix. If the norm of the perturbation is sufficiently small then the perturbed identity matrix is nonsingular.

Fact 2.23 (Inverse of Perturbed Identity) If $A \in \mathbb{C}^{n \times n}$ and $\|A\|_p < 1$ then $I + A$ is nonsingular and

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

If also $\|A\|_p \leq 1/2$ then $\|(I + A)^{-1}\|_p \leq 2$.

Proof. Suppose, to the contrary, that $\|A\|_p < 1$ and $I + A$ is singular. Then there is a vector $x \neq 0$ such that $(I + A)x = 0$. Hence $\|x\|_p = \|Ax\|_p \leq \|A\|_p \|x\|_p$ implies $\|A\|_p \geq 1$, a contradiction.

- Lower bound: $I = (I + A)(I + A)^{-1}$ implies

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

- Upper bound: From

$$I = (I + A)(I + A)^{-1} = (I + A)^{-1} + A(I + A)^{-1}$$

follows

$$1 = \|I\|_p \geq \|(I + A)^{-1}\|_p - \|A(I + A)^{-1}\|_p \geq (1 - \|A\|_p) \|(I + A)^{-1}\|_p.$$

If $\|A\|_p \leq 1/2$ then $1/(1 - \|A\|_p) \leq 2$.

□

Below is the corresponding result for inverses of general matrices.

Corollary 2.24 (Inverse of Perturbed Matrix). *Let $A \in \mathbb{C}^{n \times n}$ be nonsingular and $\|A^{-1}E\|_p < 1$. Then $A + E$ is nonsingular and*

$$\|(A + E)^{-1}\|_p \leq \frac{\|A^{-1}\|_p}{1 - \|A^{-1}E\|_p}.$$

If also $\|A^{-1}\|_p\|E\|_p \leq 1/2$ then $\|(A + E)^{-1}\|_p \leq 2\|A^{-1}\|_p$.

Proof. Since A is nonsingular, we can write $A + E = A(I + A^{-1}E)$. From $\|A^{-1}E\|_p < 1$ follows with Fact 2.23 that $I + A^{-1}E$ is nonsingular. Hence $A + E$ is nonsingular. Its inverse can be written as $(A + E)^{-1} = (I + A^{-1}E)^{-1}A^{-1}$. Now take norms and apply Fact 2.23.

The second assertion follows from $\|A^{-1}E\|_p \leq \|A^{-1}\|_p\|E\|_p \leq 1/2$. \square

Corollary 2.24 implies that if the perturbation E is sufficiently small then $\|(A + E)^{-1}\|_p$ exceeds $\|A^{-1}\|_p$ by a factor of at most two.

We use the above bounds to derive normwise condition numbers for the inverses of general nonsingular matrices. A perturbation of a nonsingular matrix remains nonsingular if the perturbation is small enough in the normwise relative sense.

Fact 2.25 If $A \in \mathbb{C}^{n \times n}$ is nonsingular and $\|A^{-1}E\|_p < 1$ then

$$\|(A + E)^{-1} - A^{-1}\|_p \leq \|A^{-1}\|_p \frac{\|A^{-1}E\|_p}{1 - \|A^{-1}E\|_p}.$$

If also $\|A^{-1}\|_p\|E\|_p \leq 1/2$ then

$$\frac{\|(A + E)^{-1} - A^{-1}\|_p}{\|A^{-1}\|_p} \leq 2\kappa_p(A) \frac{\|E\|_p}{\|A\|_p},$$

where $\kappa_p(A) = \|A\|_p\|A^{-1}\|_p \geq 1$.

Proof. Corollary 2.24 implies that $A + E$ is nonsingular. Abbreviating $F = A^{-1}E$, we obtain for the absolute difference

$$(A + E)^{-1} - A^{-1} = (I + F)^{-1}A^{-1} - A^{-1} = ((I + F)^{-1} - I)A^{-1} = -(I + F)^{-1}FA^{-1},$$

where the last equation follows from $(I + F)^{-1}(I + F) = I$. Taking norms and applying the first bound in Fact 2.23 yields

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

If $\|A^{-1}\|_p\|E\|_p \leq 1/2$ then the second bound in Fact 2.23 implies

$$\|(A + E)^{-1} - A^{-1}\|_p \leq 2\|F\|_p\|A^{-1}\|_p,$$

where

$$\|F\|_p \leq \|A^{-1}\|_p \|E\|_p = \|A\|_p \|A^{-1}\|_p \frac{\|E\|_p}{\|A\|_p} = \kappa(A) \frac{\|E\|_p}{\|A\|_p}.$$

The lower bound for $\kappa(A)$ follows from

$$1 = \|I\|_p = \|AA^{-1}\|_p \leq \|A\|_p \|A^{-1}\|_p = \kappa_p(A).$$

□

Remark 2.26. We can conclude the following from Fact 2.25:

- The inverse of A is well-conditioned in the absolute sense if its norm is “small”. In particular, the perturbed matrix is nonsingular if the perturbation has small enough norm.
- The inverse of A is well-conditioned in the relative sense if $\kappa_p(A)$ is “close to” 1. Note that $\kappa_p(A) \geq 1$.

Definition 2.27. Let $A \in \mathbb{C}^{n \times n}$ be nonsingular. The number $\kappa_p(A) = \|A\|_p \|A^{-1}\|_p$ is a norm-wise relative condition number of A with respect to inversion.

According to Fact 2.25, a perturbed matrix $A+E$ is nonsingular if $\|A^{-1}E\|_p < 1$. Is this bound pessimistic, or is it tight? Does it imply that if $\|A^{-1}E\|_p = 1$ then $A+E$ can be singular? The answer is “yes”. We illustrate this now for the two norm.

Example 2.28 Let $A \in \mathbb{C}^{n \times n}$ be nonsingular. We show how to construct an outer product E such that $\|A^{-1}E\|_2 = 1$ and $A+E$ is singular.

Set $E = -yx^*/\|x\|_2^2$, where $x \neq 0$ and $y \neq 0$ are vectors we still need to choose. Since E is an outer product, Exercise 3 in Section 2.6 implies

$$\|A^{-1}E\|_2 = \frac{\|(A^{-1}y)x^*\|_2}{\|x\|_2^2} = \frac{\|A^{-1}y\|_2 \|x\|_2}{\|x\|_2^2} = \frac{\|A^{-1}y\|_2}{\|x\|_2}.$$

Choosing $x = A^{-1}y$ gives $\|A^{-1}E\|_2 = 1$ and $(A+E)x = Ax + Ex = Ax - y = 0$. Since $(A+E)x = 0$ for $x \neq 0$, the matrix $A+E$ must be singular.

Therefore, if A is nonsingular, $y \neq 0$ is any vector, $x = A^{-1}y$, and $E = yx^*/\|x\|_2^2$ then $\|A^{-1}E\|_2 = 1$ and $A+E$ is singular. ■

Exercise 3 in Section 2.6 implies that the two norm of the perturbation in Example 2.28 is $\|E\|_2 = \|y\|_2/\|x\|_2 = \|y\|_2/\|A^{-1}y\|_2$. What is the smallest two norm a matrix E can have that makes $A+E$ singular? We show that the smallest norm such an E can have is equal to $1/\|A^{-1}\|_2$.

Fact 2.29 (Absolute Distance to Singularity) Let $A \in \mathbb{C}^{n \times n}$ be nonsingular. Then

$$\min \{ \|E\|_2 : A+E \text{ is singular} \} = \frac{1}{\|A^{-1}\|_2}.$$

Proof. Let $E \in \mathbb{C}^{n \times n}$ be any matrix such that $A + E$ is singular. Then there is a vector $x \neq 0$ so that $(A + E)x = 0$. Hence $\|x\|_2 = \|A^{-1}Ex\|_2 \leq \|A^{-1}\|_2 \|E\|_2 \|x\|_2$ implies $\|E\|_2 \geq 1/\|A^{-1}\|_2$. Since this is true for any E that makes $A + E$ singular, $1/\|A^{-1}\|_2$ is a lower bound for the absolute distance of A to singularity,

Now we show that there is a matrix E_0 that achieves equality. Construct E_0 as in Example 2.28, and choose the vector y such that $\|A^{-1}\|_2 = \|A^{-1}y\|_2$ and $\|y\|_2 = 1$. Then $\|E_0\|_2 = \|y\|_2 \|A^{-1}y\|_2 = 1/\|A^{-1}\|_2$. \square

Corollary 2.30 (Relative Distance to Singularity). *Let $A \in \mathbb{C}^{n \times n}$ be nonsingular. Then*

$$\min \left\{ \frac{\|E\|_2}{\|A\|_2} : A + E \text{ is singular} \right\} = \frac{1}{\kappa_2(A)},$$

where $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$.

Therefore, matrices that are ill-conditioned with respect to inversion are close to singular, and vice versa. In other words, matrices that are close to being singular have sensitive inverses.

The example below illustrates that absolute and relative distance to singularity are not the same.

Example. Just because a matrix is close to singularity in the absolute sense, does not imply that it is also close to singularity in the relative sense. To see this, let

$$A = \begin{pmatrix} \epsilon & \epsilon \\ 0 & \epsilon \end{pmatrix}, \quad 0 < \epsilon \ll 1, \quad A^{-1} = \begin{pmatrix} \frac{1}{\epsilon} & \frac{1}{\epsilon} \\ 0 & \frac{1}{\epsilon} \end{pmatrix}.$$

Exercise 2 in §2.6 implies for a $n \times n$ matrix B that $\|B\|_2 \leq n \max_{ij} |b_{ij}|$. Hence $\epsilon \leq \|A\|_2 \leq 2\epsilon$ and $\frac{1}{\epsilon} \leq \|A^{-1}\|_2 \leq \frac{2}{\epsilon}$. Therefore

$$\frac{\epsilon}{2} \leq \frac{1}{\|A^{-1}\|_2} \leq \epsilon, \quad \frac{1}{4} \leq \frac{1}{\kappa_2(A)} \leq 1,$$

so that A is close to singularity in the absolute sense, but far from singularity in the relative sense. \square

Exercises

- (i) Let $A \in \mathbb{C}^{n \times n}$ be unitary. Show: $\kappa_2(A) = 1$.
- (ii) Let $A, B \in \mathbb{C}^{n \times n}$ be nonsingular. Show: $\kappa_p(AB) \leq \kappa_p(A)\kappa_p(B)$.
- (iii) Residuals for Matrix Inversion.

Let $A, A + E \in \mathbb{C}^{n \times n}$ be nonsingular, and $Z = (A + E)^{-1}$. Show that

$$\|AZ - I_n\|_p \leq \|E\|_p \|Z\|_p, \quad \|ZA - I_n\|_p \leq \|E\|_p \|Z\|_p.$$

1. For small enough perturbations, the identity matrix is well-conditioned with respect to inversion, in the normwise absolute and relative sense.

Show: If $A \in \mathbb{C}^{n \times n}$ and $\|A\|_p < 1$ then

$$\|(I + A)^{-1} - I\|_p \leq \frac{\|A\|_p}{1 - \|A\|_p},$$

and if $\|A\|_p \leq 1/2$ then

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

2. If the norm of A is small enough then $(I + A)^{-1} \approx I - A$.

Let $A \in \mathbb{C}^{n \times n}$ and $\|A\|_p \leq 1/2$. Show:

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

3. One can also bound the relative error with regard to $(A + E)^{-1}$.

Let A and $A + E$ be nonsingular. Show

$$\frac{\|(A + E)^{-1} - A^{-1}\|_p}{\|(A + E)^{-1}\|_p} \leq \kappa_p(A) \frac{\|E\|_p}{\|A\|_p}.$$

4. A matrix $A \in \mathbb{C}^{n \times n}$ is called *strictly column diagonally dominant* if

$$\sum_{i=1, i \neq j}^n |a_{ij}| < |a_{jj}|, \quad 1 \leq j \leq n.$$

Show: A strictly column diagonally dominant matrix is nonsingular.

5. Let $A \in \mathbb{C}^{n \times n}$ be nonsingular. Show that $\kappa_p(A) \geq \|A\|_p / \|A - B\|_p$ for any singular matrix $B \in \mathbb{C}^{n \times n}$.