## Chapter 1. Error Estimation

This handout replaces, and extends, portions of sections 3.6, 5.6, 6.3, and 6.4 of Noble/Daniel. The objective is to connect the somewhat abstract error bound inequalities of Secton 6.4 to the accuracy of individual measurements.

The main result of this chapter is this: *The order of magnitude of the condition number of a matrix is the number of digits of accuracy that might be lost in the process of solving a linear system.*

Notation. Throughout, lower case latin and greek characters will denote numbers, upper case characters will denote matrices and bold lower case characters will denote vectors. The $i^{\text{th}}$ component of the vector $\mathbf{a}$ is $a_i$ and the entry in the $i^{\text{th}}$ row and $j^{\text{th}}$ column of the matrix $A$ is $a_{ij}$.

### 1.1. Relative Error.

The concept of relative error makes precise the somewhat inexact idea of a number being correct to $n$ digits. It is clear what we mean when we say that the numbers 123451 and 123452 agree to five digits—the difference occurs in the sixth digit. Occasionally, however, simply counting the number of digits that agree can be inadequate. One reason for this is carries. If we change 9999 by adding 1 then the change is in the fourth digit but the new number is 10000 and the two numbers have no digits in common. Since the two numbers have different numbers of digits it is not even clear in which digit the change can be said to have occured. Another problem is rounding. The numbers 1232 and 1238 differ in the fourth digit. If we round them off to three digits we get 1230 and 1240—so the rounded numbers agree to only two, not three, digits. To help clarify the situation we introduce some terminology.

Definition 1.1 (Order of Magnitude). *If $a \neq 0$, write $a = \pm f \times 10^n$ where $.1 \leq f < 1$. The exponent $n$ is the order of magnitude of $a$, the number $f$ is the mantissa of $a$ and $\pm$ is the sign of $a$. (We do not assign an order of magnitude, mantissa or sign to the number zero.)*

The number $a$ has order of magnitude $n$ if and only if $10^{n-1} \leq |a| < 10^n$. Thus $a = 1$ has order of magnitude 1 and $a = 10$ has order of magnitude 2. This definition, which is not quite standard, is convenient for error estimation computations.

The statement that numbers $a$ and $b$ agree to $n$ digits is roughly the same as the statement that $(a-b)/a$ has order of magnitude about $-n$. For the previous examples, if $a = 123451$ and $b = 123452$ then $(a-b)/a \approx -.81 \times 10^{-5}$ and $(b-a)/b \approx .81 \times 10^{-5}$. If $a = 10000$ and $b = 9999$ then $(a-b)/a = .1 \times 10^{-4}$

and $(b-a)/b \approx -.1 \times 10^{-4}$. For $a = 1232$ and $b = 1238$ the two ratios are approximately $\pm .5 \times 10^{-2}$.

The statement that $a$ and $b$ agree to a certain number of digits is intuitive and common in informal scientific discourse. In formal situations the algebraicaly more tractable ratio $(a-b)/a$ is used. The terminology we will use throughout is given in the next two definitions

Definition 1.2 (Absolute Error). *The absolute error of $b$ as an approximation to $a$ is $|a-b|$.*

Definition 1.3 (Relative Error). *The relative error of $b$ as an approximation to $a \neq 0$ is $|a-b|/|a|$.*

The absolute error of $b$ as an approximation to $a$ is the same as the absolute error of $a$ as an approximation to $b$, but relative error is not so symmetric. The relative error of $b$ as an approximation to $a$ is different from the relative error of $a$ as an approximation to $b$. But they usually are not very different.

Proposition 1.4. *Let $a, b$ be nonzero and $0 < \epsilon$. If*

$$\left|\frac{a-b}{b}\right| \leq \epsilon$$

*then*

$$1 - \epsilon \leq \left|\frac{a}{b}\right| \leq 1 + \epsilon.$$

*If further, $0 < \epsilon < 1$, then*

$$\left|\frac{a-b}{a}\right| \leq \frac{\epsilon}{1-\epsilon}.$$

Proof: Recall that for any two real numbers $a, b$ we have $|a-b| \geq ||a| - |b|| = ||b| - |a||$. Thus

$$\epsilon \geq \frac{||a| - |b||}{|b|} \quad \text{or} \quad -\epsilon \leq \frac{|a|}{|b|} - 1 \leq \epsilon.$$

whence the first conclusion. If $0 < \epsilon < 1$, taking reciprocals the first conclusion becomes

$$\frac{1}{1-\epsilon} \geq \frac{|b|}{|a|} \geq \frac{1}{1+\epsilon}$$

and so

$$\frac{|a-b|}{|a|} = \frac{|b|}{|a|}\frac{|a-b|}{|b|} \leq \frac{|a|}{|b|}\epsilon \leq \frac{\epsilon}{1-\epsilon}. \qquad \blacksquare$$

To see why $\epsilon/(1-\epsilon)$ is not much different than $\epsilon$ for $0 < \epsilon < 1$ expand the denominator in a geometric series. Then

$$\frac{\epsilon}{1-\epsilon} = \epsilon(1+\epsilon+\epsilon^2+\cdots) = \epsilon+\epsilon^2+\epsilon^3+\cdots .$$

If, for example, $\epsilon = .001 = 10^{-3}$ then

$$\epsilon+\epsilon^2+\epsilon^3+\cdots = .001001001\ldots .$$

Since the function $\epsilon/(1-\epsilon)$ has a vertical asymtote at $\epsilon = 1$ the numbers $\epsilon$ and $\epsilon/(1-\epsilon)$ are not always close. But for $0 < \epsilon < .5$ the two numbers are comparable—although carries ocassionaly produce orders of magnitude that formally differ by one (see exercises).

The difference between $|a-b|/|a|$ and $|a-b|/|b|$ is often ignored in practice but we will not do so.

Example 1.5. *A standard meter stick is used to measure a length of wire. The resulting estimate is 21.3 centimeters. Find a bound on the relative error in the measurement as an approximation to the true value.*

Solution: Note the problem: if the true length of the wire is $a$ and the measurement is $b$, then we are asked for $|a-b|/|a|$. But we don't know $a$. Proposition 1.4 is used to get around this problem.

A standard meter stick is graduated in millimeters. Assuming that we can identify the closest graduation to the exact measurement, the absolute error, $|a-b|$, would be at most .5 mm. Thus if $a$ is the exact length and $b = 21.3$ then $|a-b|/|b| \leq .0005/.213 \approx .002347 = \epsilon$ and, by Proposition 1.4, a bound on the relative error of $b$ as an approximation to $a$ is $\epsilon/(1-\epsilon) \approx .002352$. ■

Let's take a closer look at the relationship between relative error and the 'number of correct digits'. Suppose that $|a-b|/|a| = \epsilon$ and, for simplicity, suppose that $a > 0$. Then $b = a \pm a\epsilon$. Suppose $a$ has order of magnitude $n$ and $\epsilon$ has order of magnitude $-s$. Write $a = \alpha \times 10^n$ and $\epsilon = e \times 10^{-s}$ where $.1 \leq \alpha, e < 1$. Further assume $s \geq 1$. Then $a\epsilon = \alpha e \times 10^{n-s}$ and $.01 \leq \alpha e < 1$. So $b$ is obtained by modifying $a$ *after* the $s^{\text{th}}$ or even $(s+1)^{\text{th}}$ digit. If no carries occur, this means that $a$ and $b$ agree to $s$ digits.

*The statement '$b$ agrees with $a$ to $s$ digits' will be taken to mean, in this text, that the relative error of $b$ as an approximation to $a$ has order of magnitude $-s$.*

Many calculations are done with limited precision using decimal approximations to actual values. For such calculations the mere act of entering a number via a keyboard introduces some error. The size of this error, often referred to as *round-off error* depends upon the machine and processor used. A small four function calculator may not distinguish between $\frac{1}{3}$ and .333333 whereas advanced symbolic processors may allow the user to decide the number of digits to be carried in the computation. Whether the precision of a series of calculations is built into the hardware or specified by the user there is a number, called the *machine epsilon*, $\epsilon_{\text{mach}}$, which is the accuracy limit for the basic arithmetic operations. In particular, if the number $a \neq 0$ is replaced in the machine by the approximation $b$ then

$$\frac{|a-b|}{|a|} \leq \epsilon_{\text{mach}}.$$

(If $a = 0$ then $b = 0$ as well and there is no error.)

Many statistical and scientific software pacakages conform to the IEEE 32-bit arithmetic standard. For these routines the default machine epsilon is $\epsilon_{\text{mach}} = 2^{-24} + 2^{-47} \approx .6 \times 10^{-7}$. Thus these routines cannot be expected to give results more accurate than 7 digits.

The relative error is not defined when $a$, the number being approximated, is zero. In this case absolute error must be used. Of course we will not usally know if $a$ is zero or not. The most information we usually have is a bound on the relative error. From the first formula of the conclusion of Proposition 1.4 we see that $a$ cannot be zero unless $\epsilon \geq 1$. This leads to the following rule of thumb.

*If the bound on the relative error is near 1, estimates of absolute error should be used instead.*

Exercises

1. What are the orders of magnitude of: a) 100, b) 1, c) .1 d)123/456?
2. A bound on the relative error of $b \neq 0$ as an approximation to $a \neq 0$ is .1. Find a bound on the relative error of $a$ as an approximation to $b$.
3. The height of a tree is measured to be 500 feet to within half a foot. Find a bound on the relative error of the measurement as an approximation to the true height. To how many digits does the measurement agree with the true height?
4. The manual for an AC voltmeter states that the readings are accurate to ±2%.

   a) The meter registers 119.4 volts when inserted in a wall socket. Find a bound on absolute error and the relative error of the reading as an approximation to the true value.

   b) Show that your answer for the relative error in part a) does not depend upon the particular voltage measured.

c) Suppose that the accuracy is given in the manual as $k\%$ where $k$ is a number less than 100. What will the bound on the relative error of any measurement be?

5. Show that if $a$ and $b$ have the same order of magnitude, then

$$\frac{1}{10} < \left|\frac{a}{b}\right| < 10.$$

Hint: If the common order of magnitude is $n$, then $10^{n-1} \leq |a|,\ |b| < 10^n$.

6. Assume that $0 < \epsilon < .5$.

a) Show that $\epsilon$ and $\epsilon/(1-\epsilon)$ differ by at most one order of magnitude.
Hint: If $\epsilon$ has order of magnitude zero then $.1 \leq \epsilon < .5$. In this case show that $.1 < \epsilon/(1-\epsilon) < 1$. If $\epsilon$ has order of magnitude $-n$ with $n \geq 1$ then $10^{-n-1} \leq \epsilon < 10^{-n}$. In this case show that $10^{-n-1} < \epsilon/(1-\epsilon) < 10^{-n+1}$.

b) If $\epsilon$ has order of magnitude $-n$ and the orders of magnitude of $\epsilon$ and $\epsilon/(1-\epsilon)$ differ by 1, show that $f \times 10^{-n} \leq \epsilon < 10^{-n}$ where $f = 1/(1+10^{-n})$. What is $f$ for $n = 1, 2, 3$? (These are the intervals in which carries change the order of magnitude.)
Hint: Start with the inequality $10^{-n} \leq \epsilon/(1-\epsilon)$.

### 1.2. Vectors.

Recall the definition of a vector norm.

DEFINITION 1.6 (VECTOR NORM). *Let $V$ be a vector space over the real field $\mathbf{R}$. A vector norm on $V$ is a real valued function $\|\cdot\| : V \to \mathbf{R}$ such that*

(1) *For every $\mathbf{x} \in V$, $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.*
(2) *For every $\mathbf{x} \in V$ and every $\alpha \in \mathbf{R}$, $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$.*
(3) *(Triangle Inequality) For every $\mathbf{x}, \mathbf{y} \in \mathbf{R}$, $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.*

For purposes of error estimation we are interested in three vector norms on $\mathbf{R}^n$, the space of real $n$-tuples. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbf{R}^n$ and define

$$\begin{aligned}
\|\mathbf{x}\|_1 &= |x_1| + |x_2| + \cdots + |x_n| \\
\|\mathbf{x}\|_2 &= \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \\
\|\mathbf{x}\|_\infty &= \max(|x_1|, |x_2|, \ldots, |x_n|)
\end{aligned}$$

We refer to these as the 1-norm, the 2-norm and the Max-norm respectively.

If $\|\cdot\|$ is a norm on $\mathbf{R}^n$ and $c \geq 0$ then the *sphere of radius* $c$, centered at the origin, with respect to this norm is the set of vectors $\mathbf{x}$ such that $\|\mathbf{x}\| = c$. For $n = 2$, the plane, the sphere of radius $c$ for the 2-norm is the circle of radius $c$ centered at the origin. Sets of concentric spheres centered at the origin for the other two norms are shown in Figures  and . The geometry of these spheres will be important subsequently.

Recall that for vectors $\mathbf{a}$ and $\mathbf{b}$ in a normed vector space the *distance* between $\mathbf{a}$ and $\mathbf{b}$ is defined to be $\|\mathbf{a} - \mathbf{b}\|$.

Definition 1.7 (Relative Error for Vectors). *For vectors* $\mathbf{a} \neq \mathbf{0}$ *and* $\mathbf{b}$ *in a normed vector space the relative error of* $\mathbf{b}$ *as an approximation to* $\mathbf{a}$ *is* $\|\mathbf{a} - \mathbf{b}\|/\|\mathbf{a}\|$. *The absolute error of* $\mathbf{b}$ *as an approximation to* $\mathbf{a}$ *is* $\|\mathbf{a} - \mathbf{b}\|$.

It is natural to ask what the relative error of $\mathbf{b}$ as an approximation to $\mathbf{a}$ tells us about the relative error of $b_i$ as an approximation to $a_i$ and conversely. For an arbitrary norm not much can be said, but the next defintion isolates two properties that a norm may have which help in this regard.

Definition 1.8. *Let* $\|\cdot\|$ *be a vector norm.*

(1) *We say that* $\|\cdot\|$ *has Property I if when* $|a_i| \leq |b_i|$ *for all* $i$, *then* $\|\mathbf{a}\| \leq \|\mathbf{b}\|$.
(2) *We say that* $\|\cdot\|$ *has Property II if* $|a_i| \leq \|\mathbf{a}\|$ *for all* $i$.

With one exception, all the norms used in this chapter have Properties I and II. The matrix 2-norm, defined in the next section, has Property II but not Property I.

Property I allows us to go from bounds on the relative errors of the $b_i$ as approximations to a bound on the relative error of $\mathbf{b}$ as an approximation to $\mathbf{a}$.

Proposition 1.9. *Let* $\|\cdot\|$ *be a vector norm with Property I. If*

$$\frac{|a_i - b_i|}{|a_i|} \leq \epsilon_i, \text{ for all } i$$

*then*

$$\frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{a}\|} \leq \max_i \epsilon_i.$$

Proof:

$$\begin{aligned} \frac{|a_i - b_i|}{|a_i|} \leq \epsilon_i, \text{ for all } i &\implies |a_i - b_i| \leq \epsilon_i |a_i|, \text{ for all } i \\ &\implies |a_i - b_i| \leq (\max_i \epsilon_i)|a_i|, \text{ for all } i \end{aligned}$$

$$
\begin{aligned}
&\Longrightarrow\ |a_i - b_i| \le |(\max_i \epsilon_i) a_i|, \text{ for all } i\\
&\Longrightarrow\ \|\mathbf{a} - \mathbf{b}\| \le \|(\max_i \epsilon_i)\mathbf{a}\|\\
&\Longrightarrow\ \|\mathbf{a} - \mathbf{b}\| \le (\max_i \epsilon_i)\|\mathbf{a}\|\\
&\Longrightarrow\ \frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{a}\|} \le \max_i \epsilon_i. \qquad \blacksquare
\end{aligned}
$$

If the numbers $b_i$ are approximations to unknown numbers $a_i$ as in Example 1.5 and one of the $b_i$ is zero then we do not have an $\epsilon_i$ and the above proposition does not apply. If, however, $a_i$ is known to be zero as well, then the proposition may apply to the shortened vectors obtained by omitting the $i^{\text{th}}$ components. For our three standard norms, the original vectors and the shortened vectors have the same norm so the proposition will apply to the original vectors. This is the case where the numbers $b_i$ result from storing approximations to the numbers $a_i$ in a computer or calculator. Thus if $\mathbf{b}$ is the vector obtained by storing the components of $\mathbf{a}$ in a machine, then

$$\frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{a}\|} \le \epsilon_{\text{mach}}.$$

Property II allows us to go from a bound on the relative error of $\mathbf{b}$ as an approximation to $\mathbf{a}$ to a bound on the relative error of $b_i$ as an approximation to $a_i$. Before we can proceed, however, we need a vector norm version of Proposition 1.4. The inequality of the next propostion is geometrically clear in the plane for the 2-norm since any segment from the inner circle to the outer circle must be at least as long as the difference in the radii (Figure ).

PROPOSITION 1.10. *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be vectors in a normed vector space. Then*

$$\|\mathbf{a} - \mathbf{b}\| \ge |\|\mathbf{a}\| - \|\mathbf{b}\||.$$

PROOF: From the triangle inequality,

$$\|\mathbf{a}\| = \|\mathbf{a} - \mathbf{b} + \mathbf{b}\| \le \|\mathbf{a} - \mathbf{b}\| + \|\mathbf{b}\|$$

hence

$$\|\mathbf{a}\| - \|\mathbf{b}\| \le \|\mathbf{a} - \mathbf{b}\|.$$

Interchanging $\mathbf{a}$ and $\mathbf{b}$ above we have

$$\|\mathbf{b}\| - \|\mathbf{a}\| \le \|\mathbf{a} - \mathbf{b}\|.$$

(Since $\|\mathbf{a} - \mathbf{b}\| = \|\mathbf{b} - \mathbf{a}\|$ by the second property of norms with $\alpha = -1$.) But $|\|\mathbf{a}\| - \|\mathbf{b}\||$ is either $\|\mathbf{a}\| - \|\mathbf{b}\|$ or $\|\mathbf{b}\| - \|\mathbf{a}\|$. $\blacksquare$

Proposition 1.11. *Let* **a** *and* **b** *be nonzero vectors in a normed vector space and let* $0 < \epsilon$*. If*

$$\frac{\|\mathbf{a}-\mathbf{b}\|}{\|\mathbf{b}\|} \leq \epsilon$$

*then*

$$1-\epsilon \leq \frac{\|\mathbf{a}\|}{\|\mathbf{b}\|} \leq 1+\epsilon$$

*and, if* $0 < e < 1$,

$$\frac{\|\mathbf{a}-\mathbf{b}\|}{\|\mathbf{a}\|} \leq \frac{\epsilon}{1-\epsilon}.$$

Proof: By the previous proposition,

$$\epsilon \geq \frac{\|\mathbf{a}\|-\|\mathbf{b}\|}{\|\mathbf{b}\|} \geq \frac{|\|\mathbf{a}\|-\|\mathbf{b}\||}{\|\mathbf{b}\|}$$

so apply Proposition 1.4 ■

The next proposition gives a recipe for finding a bound on the relative error of $b_i$ as an approximation to $a_i$ given the relative error of **b** as an approximation to **a**.

Proposition 1.12. *Let* $\|\cdot\|$ *be a vector norm with Property II. Suppose that*

$$\frac{\|\mathbf{a}-\mathbf{b}\|}{\|\mathbf{a}\|} \leq \epsilon.$$

*where* $0 < \epsilon < 1$*. Let* $\epsilon_1 = \epsilon/(1-\epsilon)$*.Then*

$$|a_i - b_i| \leq \|\mathbf{b}\|\epsilon_1.$$

*If* $b_i \neq 0$ *set* $\epsilon_2 = \frac{\|\mathbf{b}\|}{|b_i|}\epsilon_1$.

*If* $\epsilon_2 < 1$*, then*

$$\frac{|a_i - b_i|}{|a_i|} \leq \epsilon_3$$

*where* $\epsilon_3 = \epsilon_2/(1-\epsilon_2)$.

Proof: Since $\epsilon < 1$, $\mathbf{b} \neq 0$. Applying Proposition 1.11 (with **a** and **b** interchanged) to the hypothesis we have

$$\frac{\|\mathbf{a}-\mathbf{b}\|}{\|\mathbf{b}\|} \leq \frac{\epsilon}{1-\epsilon} = \epsilon_1.$$

So $|a_i - b_i| \leq \|\mathbf{a} - \mathbf{b}\| \leq \|\mathbf{b}\|\epsilon_1$. If $|b_i|$ is nonzero, we have

$$\epsilon_1 \geq \frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{b}\|} \geq \frac{|a_i - b_i|}{\|\mathbf{b}\|} = \frac{|a_i - b_i|}{|b_i|} \frac{|b_i|}{\|\mathbf{b}\|}$$

hence

$$\frac{|a_i - b_i|}{|b_i|} \leq \frac{\|\mathbf{b}\|}{|b_i|}\epsilon_1 = \epsilon_2.$$

If $\epsilon_2 < 1$, an application of Proposition 1.4 shows

$$\frac{|a_i - b_i|}{|a_i|} \leq \frac{\epsilon_2}{1 - \epsilon_2} = \epsilon_3. \qquad \blacksquare$$

EXAMPLE 1.13. *The approximate vector* $\mathbf{b} = (1.2, -3.7, 4, 1.1 \times 10^{-8})$ *is returned by a software routine along with an estimate of the relative error in the approximation of* $10^{-5} = .1 \times 10^{-4}$ *using the 1-norm. Estimate the number of correct digits in the first component. Might the last component of the exact answer be zero?*

SOLUTION: Following the above recipe, $\|\mathbf{b}\|_1 = 8.900000011$ and $\epsilon_1 = 10^{-5}/(1 - 10^{-5}) = .100001 \times 10^{-4}$. Thus

$$\epsilon_2 = \frac{\|\mathbf{b}\|_1}{|b_1|}\epsilon_1 = \frac{8.900000011}{1.2}(.100001 \times 10^{-4}) = .741674 \times 10^{-4}$$

and $\epsilon_3 = \epsilon_2/(1 - \epsilon_2) = .741729 \times 10^{-4}$. Thus the first component is correct to at least 4 digits.

For the last component we have $|a_i - 1.1 \times 10^{-8}| \leq \|\mathbf{b}\|_1\epsilon_1 = .890009 \times 10^{-4}$ so $a_i$ might well be zero. ∎

In the above example the final esimate $\epsilon_3$ is the same order of magnitude as the original estimate $\epsilon$. This is often the case with the standard norms but in the above recipe it depends upon $|b_i|$ not being several orders of magnitude smaller than $\|\mathbf{b}\|$. (If $0 < \epsilon < .5$ then $\epsilon$ and $\epsilon/(1 - \epsilon)$ are close and usually have the same order of magnitude. See Exercise 6, Section 1.) The next example shows that this is a real effect and not simply due to the naiveté of the approximations used.

EXAMPLE 1.14. *Let* $\mathbf{a} = (1, 10000), \mathbf{b} = (3, 9996)$ *be vectors in* $\mathbf{R}^2$. *What is the relative error of* $\mathbf{b}$ *as an approximation to* $\mathbf{a}$ *using the Max-norm? What is the relative error for the individual components?*

Solution: The relative error for the vector approximation is

$$\begin{aligned}\frac{\|\mathbf{a}-\mathbf{b}\|_\infty}{\|\mathbf{a}\|_\infty} &= \frac{\|(1,10000)-(3,9996)\|_\infty}{\|(1,10000)\|_\infty} \\ &= \frac{\|(-2,4)\|_\infty}{\|(1,10000)\|_\infty} \\ &= \frac{4}{10000} \\ &= .4\times 10^{-3}.\end{aligned}$$

For the individual components the relative errors are

$$\frac{|1-3|}{|1|}=2 \quad \text{and} \quad \frac{|10000-9996|}{|10000|}=.4\times 10^{-3}.$$

Not even one digit of the first component is correct. ■

If a computation results in a vector with components of disparate orders of magnitude and the components represent physical quantities or measurements then the original data may often be rescaled so that the computation produces a vector all the components of which are about the same size.

As an illustration, suppose that the computation deals with population data. Assume that the data consists of pairs of numbers $(x, y)$ where $x$ is a year and $y$ is the population of the United States. In 1970 the U.S. population was 203.2 million individuals. The data vector is $(1973, 203200000) = (.1973\times 10^4, .2032\times 10^9)$. The two components differ by 5 orders of magnitude, which is undesirable. The population data can be rescaled by 5 orders of magnitude making the data point $(1970, 2032)$. Such rescalings are simply changes of units. The units for the two components are years and "individuals" for the original data vector and years and "hundreds-of-thousands of individuals" for the rescaled data vector.

The next proposition allows one to shift estimates from one norm to another.

Proposition 1.15 (Norm Equivalence). *Let $\mathbf{x}$ be a vector in $\mathbf{R}^n$.Then*

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

Proof: Suppose that $\|\mathbf{x}\|_\infty = |x_k|$. The first line of the conclusion follows from the two inequalites

$$|x_k|^2 \le |x_1|^2 + \cdots + |x_k|^2 + \cdots + |x_n|^2$$

and

$$|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2 \le |x_k|^2 + |x_k|^2 + \cdots + |x^k|^2 = n|x_k|^2.$$

The last line of the conclusion follows by noting that the above inequalites are valid with the exponents removed.

For the second line of inequalities note first that

$$\begin{aligned}
\|\mathbf{x}\|_1^2 = (|x_1| + |x_2| + \cdots + |x_n|)^2 &= |x_1|^2 + |x_2|^2 + \cdots + |x_n|^2 + |x_1|\,|x_2| + \cdots \\
&\ge |x_1|^2 + |x_2|^2 + \cdots + |x_n|^2 \\
&= \|\mathbf{x}\|_2^2
\end{aligned}$$

which is the first inequality. For the second inequality we have that for any real number $\alpha$, $(1 + \alpha|x_i|)^2 \ge 0$. Thus

$$\begin{aligned}
0 \le \sum_{i=1}^{n} (1 + \alpha|x_i|)^2 &= \sum_{i=1}^{n} (1 + 2\alpha|x_i| + \alpha^2|x_i|^2) \\
&= n + 2\alpha \sum_{i=1}^{n} |x_i| + \alpha^2 \sum_{i=1}^{n} |x_i|^2 \\
&= n + 2\alpha\|\mathbf{x}\|_1 + \alpha^2\|\mathbf{x}\|_2^2.
\end{aligned}$$

Now if a quadratic $c + b\alpha + a\alpha^2$ is never negative, then we must have $b^2 - 4ac \le 0$, which gives the desired result. ■

As an illustration suppose that the relative error of **b** as an approximation to **a** in the 2-norm is $\epsilon$. Then, using the Norm Equivalence inequalities, we have

$$\frac{\|\mathbf{a} - \mathbf{b}\|_\infty}{\|\mathbf{a}\|_\infty} \le \frac{\|\mathbf{a} - \mathbf{b}\|_2}{\|\mathbf{a}\|_\infty} \le \frac{\|\mathbf{a} - \mathbf{b}\|_2}{\|\mathbf{a}\|_2}\sqrt{n} \le \sqrt{n}\epsilon.$$

So, for a vector of length $n = 100$, if the estimate of the relative error in the 2-norm has order of magnitude $-s$ then, the order of magnitude of the relative error in the Max-norm is at most $-s + 1$.

Exercises

1. Graph the sphere $\|\mathbf{x}\|_1 = 1$ in the plane.
   Hint: $|a| = \pm a$ depending on the sign of $a$. Thus $\|\mathbf{x}\|_1 = 1$ is made up of portions of the four lines $\pm x \pm y = 1$.
2. Graph the set of points $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{a}\|_1 = 2$ where $\mathbf{a} = (1, 1)$.
3. Graph the sphere $\|\mathbf{x}\|_\infty = 1$ in the plane.

Hint: $\max(|x_1|, 1) = 1$ for all $x_1$ with $-1 \le x_1 \le 1$.

4. Graph the set of points $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{a}\|_\infty = 2$ where $\mathbf{a} = (1, 1)$.
5. The approximate vector $\mathbf{b} = (1.2, 10^{-3})$ is returned by a software routine along an estimate of the relative error in the Max-norm of $\epsilon = 10^{-4}$.
   a) Estimate the number of correct digits for each component. Might the last component of the exact answer be zero?
   b) Same problem with $\mathbf{b} = (1.2, 10^{-4})$, $\epsilon = 10^{-4}$.
   c) Same problem with $\mathbf{b} = (1.2, 10^{-5})$, $\epsilon = 10^{-4}$.
6. Redo Example 1.13 assuming that the error approximation is given in the 2-norm. The Max-norm.
7. Let $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$ be vectors in $\mathbf{R}^n$.
   a) Show that for the 1-, 2-, and Max-norms $|a_i| \le \|\mathbf{a}\|$ for all $i$.
   b) Show that for the 1-, 2-, and Max-norms, if $|a_i| \le |b_i|$ for all $i$ then $\|\mathbf{a}\| \le \|\mathbf{b}\|$.
8. Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$ and let $A$ be an $n \times n$ nonsingular matrix. Show that the function $\|\cdot\| : \mathbf{R}^n \to \mathbf{R}^n$ defined by $\|\mathbf{x}\|_A = \|A\mathbf{x}\|$ is a norm.
   Hint: Recall that a matrix is nonsingular if and only if $A\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$.
9. By Problem 8 the function $\|\mathbf{x}\| = \|A\mathbf{x}\|_2$ with

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

   is a norm on $\mathbf{R}^2$. Let $\mathbf{a} = (a_1, a_2)$ and let $\mathbf{b} = (b_1, b_2)$.
   a) Sketch some concentric spheres about the origin.
   b) Find a vector $\mathbf{a}$ with $\|\mathbf{a}\| < |a_i|$ for $i = 1$ or $i = 2$.
   c) Find vectors $\mathbf{a}$ and $\mathbf{b}$ with $|a_i| \le |b_i|$ for $i = 1, 2$ and $\|\mathbf{a}\| > \|\mathbf{b}\|$.

### 1.3. Norms and Condition Numbers for Matrices.

An $n \times n$ matrix represents a linear transformation from the vector space $\mathbf{R}^n$ to itself. Figure  shows the image of the unit sphere $\|\mathbf{x}\|_2 = 1$ under the linear transformation represented by the matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

The most obvious aspect of this image (it is an ellipse) is its elongation. The concentric circles in Figure  are .5 units apart so, by inspection, the points of the ellipse furthest from the origin are about 5.5 units and the points nearest the origin are about .4 units. We say that the 2-*norm* of the matrix is about 5.5 and the 2-*condition nuumber* is about $5.5/.4 = 13.75$. The norm is the maximum extension of the image and the condition number is the ratio of maximum extension to minimum extension. The formal definitions are slightly different.

DEFINITION 1.16 (NORM OF A MATRIX). *Let $V$ be a finite dimensional vector space with norm $\|\cdot\|$ and let $T$ be a linear transformation on $V$. The induced norm of $T$, denoted $\|T\|$, is defined by*

$$\|T\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|T(\mathbf{x})\|}{\|\mathbf{x}\|}.$$

*The induced norm of a square matrix is the induced norm of the associated linear transformation.*

Since the maximum is taken over an infinite set it is not obvious that the maximum exists—certainly $\max_{\mathbf{x}\neq\mathbf{0}} \|\mathbf{x}\|$ does not exist since there are vectors of arbitrary length. The proof that the norm exists is beyond the scope of this text.

The next proposition brings the definition back to Figure .

PROPOSITION 1.17. *Let $A$ be an $n \times n$ matrix and let $\mathbf{x}$ be any vector in $\mathbf{R}^n$. If $\|\cdot\|$ is a norm on $\mathbf{R}^n$ then the induced norm of $A$ is*

$$\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

PROOF: Since the maximum of the definition exists we may assume that there is a vector $\mathbf{x}_0$ such that $\|A\| = \|A\mathbf{x}_0\|/\|\mathbf{x}_0\|$. Then

$$\frac{\|A\mathbf{x}_0\|}{\|\mathbf{x}_0\|} = \frac{\|\|\mathbf{x}_0\|A(\mathbf{x}_0/\|\mathbf{x}_0\|)\|}{\|\|\mathbf{x}_0\|(\mathbf{x}_0/\|\mathbf{x}_0\|)\|} = \frac{\|A(\mathbf{x}_0/\|\mathbf{x}_0\|)\|}{\|\mathbf{x}_0/\|\mathbf{x}_0\|\|}$$

and $\|\mathbf{x}_0/\|\mathbf{x}_0\|\| = 1$. ∎

Norm equivalence for vector norms (Proposition 1.15) implies a similar set of identities for the induced matrix norms.

PROPOSITION 1.18 (MATRIX NORM EQUIVALENCE). *Let $A$ be an $n\times n$ matrix. Then*

$$\frac{\|A\|_2}{\sqrt{n}} \leq \|A\|_\infty \leq \|A\|_2\sqrt{n}$$
$$\frac{\|A\|_2}{\sqrt{n}} \leq \|A\|_1 \leq \|A\|_2\sqrt{n}$$
$$\frac{\|A\|_1}{n} \leq \|A\|_\infty \leq \|A\|_1 n$$

Proof: Exercise ■

If the condition number of $A$ is to be the ratio of the maximum extension to the minimum extension then the minimum extension should not be zero. That is there should not be a vector $\mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \mathbf{0}$. Recall that this condition is equivalent to the nonsingularity of $A$.

Definition 1.19 (Condition Number of a Matrix). *Let $A$ be a nonsingular $n \times n$ matrix, $\mathbf{x}$ any vector in $\mathbf{R}^n$, and $\|\cdot\|$ a norm on $\mathbf{R}^n$. The induced condition number, $\kappa(A)$ is defined by*

$$\kappa(A) = \frac{\|A\|}{m} \quad \textit{where} \quad m = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}.$$

As with the maximum, we merely state without proof that a nonzero minimum exists. The computation of the minimum may also be restricted to the unit sphere, giving the following proposition.

Proposition 1.20. *Let $A$ be a nonsingular $n \times n$ matrix, $\mathbf{x}$ any vector in $\mathbf{R}^n$ and $\|\cdot\|$ a norm on $\mathbf{R}^n$. The induced condition number is given by*

$$\kappa(A) = \frac{\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|}{\min_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|}.$$

Proof: Exercise. ■

The importance of the condition number of $A$ lies in its connection to estimates of the error arising when approximating $A^{-1}\mathbf{b}$, the solution of $A\mathbf{x} = \mathbf{b}$. The next proposition connects $\kappa(A)$ with $A^{-1}$.

Proposition 1.21. *Let $A$ be a nonsingular $n \times n$ matrix and let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. The induced condition number is given by*

$$\kappa(A) = \|A^{-1}\| \, \|A\|.$$

Proof: The matrix $A$ is assumed nonsingular and hence as $\mathbf{x}$ ranges over all of $\mathbf{R}^n$ so does $\mathbf{y} = A\mathbf{x}$ since, given any $\mathbf{y}$, $\mathbf{y} = A(A^{-1}\mathbf{y})$. Thus

$$m = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \min_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{y}\|}{\|A^{-1}\mathbf{y}\|} = \left( \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|A^{-1}\mathbf{y}\|}{\|\mathbf{y}\|} \right)^{-1}$$

and the conclusion follows. (Note that for any set $S$ of positive numbers with a smallest member, $\min_S(s) = \left[\max_S(\frac{1}{s})\right]^{-1}$.) ■

The use of the terms 'induced norm' and 'induced condition number' in the above propositions emphasizes the fact that the matrix norm and condition number change with the vector norm used. As usual, we use subscripts for particular norms. Thus the estimates from Figure  above are $\|A\|_2 \approx 5.5$ and $\kappa_2(A) \approx 13.75$. These estimates are evident from the figure but a general procedure for computing the 2-norm and 2-condition number is not evident. In fact, the computation is not easy. (A formula for $2 \times 2$ matrices appears in the exercises. The singular value decomposition of the matrix may be used in the general case.). The situation is different for the 1-norm and Max-norm (but not for the corresponding condition numbers). To see why, consider the Figures  and , which are the 1-norm and Max-norm versions of Figure .

The geometry of the situation is this: *because the unit spheres and their images are composed of straight lines, the maximum will occur at the image of a vertex of the unit sphere.* (This same geometry underlies the simplex method of linear programming.)

To compute these norms it is only necessary to apply the matrix to the vertices of the unit sphere and compute the norm of the result. We now describe how this procedure may be streamlined.

The vertices of the 1-norm unit sphere in $\mathbf{R}^2$ are $(\pm 1, 0)$ and $(0, \pm 1)$. Multiplying these vectors by a $2 \times 2$ matrix produces the columns of the matrix and their negatives. So the 1-norm of a $2 \times 2$ matrix is largest 1-norm of its columns considered as vectors.

The vertices of the Max-norm unit sphere are the four vectors $(\pm 1, \pm 1)$. Multiplying these vectors by a $2 \times 2$ matrix amounts to multiplying each element of a row by 1 or $-1$ and summing. Since all sign combinations occur, the sum of the absolute values of the elements of each row must occur as a component of some vertex. The Max-norm of a $2 \times 2$ matrix is the largest 1-norm of its rows considered as vectors.

These statements are true in general. The third statement of the next proposition is included for completeness. The proof may be found in the section on the Singular Value Decomposition. Note that an upper bound on $\|A\|_2$ may be found by computing $\|A\|_1$ or $\|A\|_\infty$ and using Proposition 1.18.

Proposition 1.22. *Let $A$ be an $n \times n$ matrix. Let $\mathbf{c}_i$ denote the $i^{\text{th}}$ column of $A$ and let $\mathbf{r}_i$ denote the $i^{\text{th}}$ row of $A$. Then*

(1) $\|A\|_1 = \max_i \|\mathbf{c}_i\|_1$
(2) $\|A\|_\infty = \max_i \|\mathbf{r}_i\|_1$.
(3) $\|A\|_2$ *is the first singular value of $A$.*

Proof: Let $A$ have entries $a_{ij}$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. For the first statement suppose that $\|\mathbf{c}_{i_0}\|_1 = \max_i \|\mathbf{c}_i\|_1$. Set $x_{i_0} = 1$ and $x_j = 0$ for $j \neq i_0$. Then $\|\mathbf{x}\|_1 = 1$ and so

$$\|\mathbf{c}_{i_0}\|_1 = \|A\mathbf{x}\|_1 \leq \max_{\|\mathbf{y}\|_1=1} \|A\mathbf{y}\|_1 = \|A\|_1.$$

For the opposite inequality we have

$$\begin{aligned}
\|A\|_1 = \max_{\|\mathbf{x}\|_1=1} \|A\mathbf{x}\|_1 &= \max_{\|\mathbf{x}\|_1=1} \|(\sum_j a_{1j}x_j, \sum_j a_{2,j}, \dots, \sum_j a_{nj})\|_1 \\
&= \max_{\|\mathbf{x}\|_1=1} \sum_i \left| \sum_j a_{ij}x_j \right| \leq \max_{\|\mathbf{x}\|_1=1} \sum_{i,j} |a_{ij}|\,|x_j| \\
&= \max_{\|\mathbf{x}\|_1=1} \sum_j \left( \sum_i |a_{ij}| \right) |x_j| = \max_{\|\mathbf{x}\|_1=1} \sum_j \|\mathbf{c}_j\|_1\,|x_j| \\
&\leq \max_{\|\mathbf{x}\|_1=1} \sum_j \|\mathbf{c}_{i_0}\|_1 |x_j| = \|\mathbf{c}_{i_0}\|_1 \max_{\|\mathbf{x}\|_1=1} \sum_j |x_j| \\
&= \|\mathbf{c}_{i_0}\|_1 \max_{\|\mathbf{x}\|_1=1} \|\mathbf{x}\|_1 = \|\mathbf{c}_{i_0}\|_1
\end{aligned}$$

and so the first statement is proved.

For the second statement suppose that $\|\mathbf{r}_{i_0}\|_1 = \max_i \|\mathbf{r}_i\|_1$. Choose $x_j = \pm 1$ so that $a_{i_0j}x_j = |a_{i_0j}|$. Then $\|\mathbf{x}\|_\infty = 1$ and

$$\begin{aligned}
\max_i \|\mathbf{r}_i\|_1 = \|\mathbf{r}_{i_0}\|_1 = \sum_j |a_{i_0j}| &= \left| \sum_j a_{i_0j}x_j \right| \\
&= \|A\mathbf{x}\|_\infty \leq \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{y}\|_\infty = \|A\|_\infty.
\end{aligned}$$

The reverse inequality is left as an exercise. ∎

Example 1.23. *Compute the 1-norm, Max-norm and the corresponding condition numbers for the matrix* $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

Solution: The 1-norms of the columns are 4 and 6 so $\|A\|_1 = 6$. The 1-norms of the rows are 3 and 7 so $\|A\|_\infty = 7$.

Recall that for a nonsingular $2 \times 2$ matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & d \end{pmatrix}.$$

Thus, using Proposition 1.21,

$$\kappa_1(A) = \|A^{-1}\|_1\|A\|_1 = \frac{7}{2}\cdot 6 = 21$$
$$\kappa_\infty(A) = \|A^{-1}\|_\infty\|A\|_\infty = 3\cdot 7 = 21$$

This is the matrix of Figure  which has an approximate 2-norm of 5.5 and an approximate 2-condition number of 13.75. ■

The next result shows that the induced norm on matrices is indeed a norm and gives two results on matrix multiplication that are important in making error estimates.

Proposition 1.24. *Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$, $\mathbf{x}$ any vector in $\mathbf{R}^n$ and $A$, $B$ any two $n\times n$ matrices.*

(1) $\|A\| \geq 0$ *and* $\|A\| = 0$ *if and only if $A$ is the zero matrix.*
(2) $\|\alpha A\| = |\alpha|\,\|A\|$ *for any real number $\alpha$.*
(3) $\|A+B\| \leq \|A\| + \|B\|$
(4) $\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$
(5) $\|AB\| \leq \|A\|\|B\|$

Proof: The first three statements follow from the corresponding statements for vectors (see exercises). The fourth statement is clearly true if $\mathbf{x} = \mathbf{0}$. Otherwise

$$\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\|$$

since the right hand side of this inequality is the maximum possible value of the left hand side. For the fifth statement note that

$$\|AB\mathbf{x}\| = \|A(B\mathbf{x})\| \leq \|A\|\|B\mathbf{x}\| \leq \|A\|\|B\|\|\mathbf{x}\|$$

so that the statement follows by maximizing over all $\mathbf{x}$ of norm one. ■

As far as addition and scalar multiplication are concerned, an $n\times n$ matrix is just a vector with $n^2$ components. The first three statements of Proposition 1.24 show that the induced norms are vector norms on this vector space. It follows that the defintion of relative and absolute error apply to matrices without change and Proposition 1.11 may be applied to matrices as well as vectors. This will be done below without comment.

Properties I and II may also be defined for matrix norms by considering matrices to be vectors (replace $a_i$ by $a_{ij}$, $\mathbf{a}$ by $A$, and so on).

Proposition 1.25. *The matrix norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ have Properties I and II. The matrix norm $\|\cdot\|_2$ has Property II but not Property I.*

Proof: Exercises ■

If a matrix has entries which differ significantly in size it is not necessarily possible to rescale the size difference away. In a linear system $A\mathbf{x} = \mathbf{b}$, rescaling the variables rescales the columns of $A$. Rescaling the rows is equivalent to multiplying the corresponding linear equations by a nonzero constant. The rescaling options for a matrix are muliplying each row by a constant and multiplying each column by a constant. These operations may interfere with each other. As an example consider the matrix

$$\begin{pmatrix} 1 & 10^5 \\ 10^5 & 10^5 \end{pmatrix}.$$

Example 1.26. *A software routine returns the matrix*

$$B = \begin{pmatrix} 1.5 & -2.9 & 4.1 \\ .4 & .3 \times 10^{-8} & 7.8 \\ -1.3 & 2.0 & 4.5 \end{pmatrix}$$

*and an error bound of* $10^{-5}$ *using one of the standard norms. Estimate the number of correct digits in the 1,1 entry. Might the 2,2 entry be zero?*

Solution: We make estimates for each of the standard norms. Using Proposition 1.12 we have $\epsilon_1 = .100001 \times 10^{-4}$ and

$$\epsilon_2 = \frac{\|B\|}{|b_{ij}|}\epsilon_1, \qquad \epsilon_3 = \epsilon_2/(1-\epsilon_2).$$

For $b_{ij} = b_{11} = 1.5$ we have $\|B\|_1 = 16.4$ and $\|B\|_\infty = 8.5$ giving $\epsilon_3$ values of $.109 \times 10^{-3}$ and $.567 \times 10^{-4}$ respectively. By Proposition 1.18, $\|B\|_2 \leq \sqrt{3}\|B\|_1, \sqrt{3}\|B\|_\infty$ Taking the smaller of these estimates gives $\epsilon_3 = .982 \times 10^{-4}$. Thus the number of digits correct is estimated at 3 using the 1-norm and at 4 using the Max- and 2-norms.

For $b_{ij} = b_{22} = 10^{-8}$ we get $\epsilon_2 > 1$ in all three norms and so no digits need be correct. However we have

$$|a_{22} - b_{22}| \leq \|B\|\epsilon_1$$

which gives absolute error bounds of $.85 \times 10^{-4}$, $.164 \times 10^{-3}$ and $.147 \times 10^{-3}$ for the Max-, 1-, and 2-norms respectively. Thus $a_{22}$ may well be zero. ■

A matrix with a large condition number is often said to be *nearly singular*. In fact, the condition number of a matrix gives the minimum relative error with which the matrix may be approximated by a singular matrix.

Proposition 1.27. *Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. Let $A$ be a nonsingular matrix and $B$ a singular matrix. Then*

$$\frac{\|A-B\|}{\|A\|} \geq \frac{1}{\kappa(A)}$$

*and there is a singular matrix $B$ for which equality holds.*

Proof: Let $m = \min_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$. Then $\kappa(A) = \|A\|/m$ so to show the inequality it is sufficient to show that $\|A-B\| \geq m$. Let $A-B = \Delta$. Since $B$ is singular, there is a vector $\mathbf{x} \neq \mathbf{0}$ such that $B\mathbf{x} = \mathbf{0}$. We may assume that $\|\mathbf{x}\| = 1$. Thus, by Proposition 1.24,

$$\|A-B\| = \|\Delta\| = \|\Delta\|\|\mathbf{x}\| \geq \|\Delta\mathbf{x}\| = \|A\mathbf{x}\| \geq m\|\mathbf{x}\| = m.$$

To demonstrate the converse for the three standard norms we can proceed as follows. We need a matrix $\Delta$ with $B = A + \Delta$ singular and $\|\Delta\| = m$. Since $\|A^{-1}\| = 1/m$ there is a vector $\mathbf{y}_m$ with $\|\mathbf{y}_m\| = 1$ and $\|A^{-1}\mathbf{y}_m\| = 1/m$. Set $\mathbf{x}_m = A^{-1}\mathbf{y}_m$. We find a vector $\mathbf{v}$ so that, considering vectors to be single column matrices, $\mathbf{v}^T\mathbf{x}_m = 1/m$ and $\|\mathbf{y}_m\mathbf{v}^T\| = 1$. Then for $\Delta = -m\mathbf{y}_m\mathbf{v}^T$ we have $\|\Delta\| = m$ and $(A+\Delta)\mathbf{x}_m = \mathbf{0}$ so $A + \Delta$ is singular. The choice of $\mathbf{v}$ is different for each of the three norms and is given in the exercises. ∎

Example 1.28. *An unknown matrix $A$ is approximated by the matrix*

$$B = \begin{pmatrix} 0.1472 & 1.3900 & 0.6840 \\ 0.5100 & 3.0100 & 1.7700 \\ 0.2690 & 1.0200 & 0.3340 \end{pmatrix}.$$

*The entries of $B$ are derived from physical measurements made with a relative accuracy bounded by $.5 \times 10^{-4}$. Is it possible that $A$ is singular?*

Solution: If we use the Max-norm, then the relative error in the norm is bounded by the relative error in the individual components and so

$$\frac{\|A-B\|_\infty}{\|A\|_\infty} \leq \epsilon = .5 \times 10^{-4}$$

hence

$$\frac{\|A-B\|_\infty}{\|B\|_\infty} \leq \frac{\epsilon}{1-\epsilon} = .50025 \times 10^{-4}.$$

The relative distance from $B$ to $A$ is at most $.50025 \times 10^{-4}$.

Using exact arithmetic software we find that

$$B^{-1} = \frac{1}{27317027}\begin{pmatrix} -200015000 & 58355000 & 100365000 \\ 76447500 & -33707800 & 22074000 \\ -72372500 & 55941500 & -66457000 \end{pmatrix}$$

Thus the relative distance from $B$ to the nearest singular matrix is

$$\frac{1}{\kappa_\infty(B)} = \frac{1}{\|B\|_\infty \|B^{-1}\|_\infty} = 0.014395 > .50025 \times 10^{-4}$$

and $A$ cannot be singular. ∎

This method can sometimes be used to prove a matrix nonsingular. It cannot be used to prove a matrix singular.

Exercises

1. Let
$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$
$F^2 = a^2 + b^2 + c^2 + d^2$, and $\Delta = \det A$. The *singular values* of $A$ are the numbers $\sigma_1 \geq \sigma_2 \geq 0$ defined by
$$\sigma_1^2 = \frac{F^2 + \sqrt{F^4 - 4\Delta^2}}{2}, \qquad \sigma_2^2 = \frac{F^2 - \sqrt{F^4 - 4\Delta^2}}{2}.$$
It can be shown that $\|A\|_2 = \sigma_1$ and, in the notation of Definition 1.19, $m = \sigma_2$.
   a) Show that $A$ is invertible if and only if $\sigma_2 \neq 0$ and then $\|A^{-1}\|_2 = 1/\sigma_2$ and $\kappa_2(A) = \sigma_1/\sigma_2$.
   b) Use these formulas to find $\|A\|_2$ and $\kappa_2(A)$ for
   $$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$
   the matrix for which the 2-norm and 2-condition number were estimated from Figure .
2. An unknown matrix $A$ is approximated by the matrix
$$B = \begin{pmatrix} 1 & 1 \\ 1 & 1.01 \end{pmatrix}.$$

The entries of $B$ are derived from physical measurements made with a relative accuracy bounded by $.5 \times 10^{-2}$. Is it possible that $A$ is singular? What if the relative accuracy of the measurements is bounded by $.1 \times 10^{-2}$?

3. An unknown matrix $A$ is approximated by the matrix

$$B = \begin{pmatrix} 3.429 & 4.357 & 4.150 \\ 1.869 & 2.339 & 3.254 \\ 1.607 & 1.964 & 4.100 \end{pmatrix}.$$

   The entries of $B$ are derived from physical measurements made with a relative accuracy bounded by $.5 \times 10^{-4}$. Is it possible that $A$ is singular?

4. Let $a_{ij}$ be an entry of the matrix $A$. Show that $|a_{ij}| \leq \|A\|$ for any of the three standard norms.
   Hint: For the 2-norm, note that if $\mathbf{e}_i$ is the vector with 1 in the $i^{\text{th}}$ place and zeros elsewhere then $\|\mathbf{e}_i\|_2 = 1$ and $A\mathbf{e}_i$ is the $i^{\text{th}}$ column of $A$.
5. Let $A$ and $B$ be square matrices of the same size.
   a) Suppose the norm $\|\cdot\|$ has the property that $|a_{ij}| \leq |b_{ij}|$ for all $i$, $j$ forces $\|A\| \leq \|B\|$. If $B$ is obtained from $A$ by changing the signs of some of the entries of $A$, then $\|A\| = \|B\|$.
   b) Using the formulas of Problem 1 show by example that the 2-norm does not have the property described in part a).
6. Show that $\kappa_1(A) = \kappa_\infty(A)$ for every nonsingular $2 \times 2$ matrix $A$. Show the statement false for $3 \times 3$ matrices.
7. Recall that matrices $A$ and $B$ are *similar* if there is a nonsingular matrix $P$ such that $A = P^{-1}BP$. Let $A$ and $B$ be similar.
   a) Show that if $\lim_{n\to\infty} \|B^n\|$ is zero then the same is true of $\lim_{n\to\infty} \|A^n\|$.
   b) Show that if $\lim_{n\to\infty} \|A^n\| = \infty$ then $\lim_{n\to\infty} \|B^n\| = \infty$

   Hint: Statement (5) of Proposition 1.24.
8. Prove Proposition 1.18
9. Prove Proposition 1.20
10. Finish the proof of Proposition 1.22
11. Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$ and let $A$ be an $n \times n$ matrix. Show that $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A$ is the zero matrix. (This is the first statement of Proposition 1.24).
    Hint: Show that $A\mathbf{x} = 0$ for every $\mathbf{x} \neq 0$. Then apply $A$ to the elements of a basis.

12. Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$ and let $A$ be an $n \times n$ matrix. Show that $\|\alpha A\| = |\alpha|\|A\|$ for every real number $\alpha$. (This is the second statement of Proposition 1.24).
13. Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$ and let $A$ and $B$ be two $n \times n$ matrices. Show that $\|A + B\| \leq \|A\| + \|B\|$. (This is the third statement of Proposition 1.24).
14. This exercise gives constructions for the vector $\mathbf{v}$ needed in the proof of Proposition 1.27. Show that the vector $\mathbf{v}$ satisfies $\mathbf{v}^T\mathbf{x}_m = 1/m$ and $\|\mathbf{y}_m\mathbf{v}^T\| = 1$.

    a) (Max-norm) The vector $\mathbf{y}_m$ will have components consisting of $\pm 1$'s. Some component of $\mathbf{x}_m$, say the $i^{\text{th}}$, must be $\pm m$. Take $\mathbf{v} = \mathbf{e}_i$, the vector with 1 in the $i^{\text{th}}$ component and zeros elsewhere.
    b) (1-norm) Since $\|\mathbf{x}_m\|_1 = 1/m$ there is a vector $\mathbf{v}$ with components $\pm 1$ such that $\mathbf{v}^T\mathbf{x}_m = 1/m$ (why?). Note that $\mathbf{y}_m$ will have one component equal to $\pm 1$ and the rest zeros.
    c) (2-norm) In this case take $\mathbf{v} = m\mathbf{x}_m$. To compute $\|\Delta\|_2$ you will need the formula $\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|_2\|\mathbf{y}\|_2 \cos\theta$, where $\theta$ is the angle between $\mathbf{x}$ and $\mathbf{y}$, to show $\max_{\|\mathbf{x}\|_2=1} \|\mathbf{y}_m\mathbf{v}^T\mathbf{x}\|_2 = 1$.
    d) Use the above constructions to find the singular matrix closest to

    $$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

    in the three standard norms.
    Note: To find the vector $\mathbf{y}_m$ for the 2-norm calculation you can proceed as follows. The vector $\mathbf{y}(t) = (\cos t, \sin t)$ sweeps out the unit sphere $\|\mathbf{x}\|_2 = 1$ as $t$ varies from 0 to $2\pi$. The maximum value of the function $f(t) = \|A^{-1}\mathbf{y}(t)\|_2^2$ may be found by standard calculus techniques.

### 1.4. Error Estimation for a System of Linear Equations.

Let $A$ be a nonsingular $n \times n$ matrix and $\mathbf{b}$ a vector in $\mathbf{R}^n$ and consider the problem of solving the equation $A\mathbf{x} = \mathbf{b}$ for the unknown vector $\mathbf{x}$ in the presence of error.

For the simplest case, assume that $A^{-1}$ is known exactly but $\mathbf{b}$ is known only approximately. How does the uncertainty in $\mathbf{b}$ translate into uncertainty in the answer $\mathbf{x} = A^{-1}\mathbf{b}$? The answer is that, in the worst case, the relative error in $\mathbf{b}$ may be multiplied by the condition number, $\kappa(A)$, of $A$.

Proposition 1.29 (Perturbing the Right Hand Side). *Let $A$ be nonsingular, $A\mathbf{x} = \mathbf{b}$ and $A\mathbf{x}_0 = \mathbf{b}_0$. Then*

$$\frac{\|\mathbf{x} - \mathbf{x}_0\|}{\|\mathbf{x}\|} \le \kappa(A) \frac{\|\mathbf{b} - \mathbf{b}_0\|}{\|\mathbf{b}\|}.$$

Proof: Let $A\mathbf{x} = \mathbf{b}$ and $A\mathbf{x}_0 = \mathbf{b}_0$. Note that $\|\mathbf{b}\| = \|A\mathbf{x}\| \le \|A\|\|\mathbf{x}\|$ by Proposition 1.24 and so $1/\|\mathbf{x}\| \le \|A\|/\|\mathbf{b}\|$. Now

$$\frac{\|\mathbf{x} - \mathbf{x}_0\|}{\|\mathbf{x}\|} = \frac{\|A^{-1}(\mathbf{b} - \mathbf{b}_0)\|}{\|\mathbf{x}\|} \le \frac{\|A^{-1}\| \, \|\mathbf{b} - \mathbf{b}_0\|}{\|\mathbf{x}\|} \le \|A\| \, \|A^{-1}\| \frac{\|\mathbf{b} - \mathbf{b}_0\|}{\|\mathbf{b}\|}$$

and so the result follows from Propositon 1.21. ∎

So if, for example, $\mathbf{b}_0$ agrees with $\mathbf{b}$ to 6 digits and $\kappa_\infty(A)$ has order of magnitude 4 then $\mathbf{x}_0$ might agree with $\mathbf{x}$ to only 2 digits.

The previous proposition is accurate but misleading in a probalistic sense. On the one hand, one may always find a $\mathbf{b}$ and a $\mathbf{b}_0$ for which the inequality is an equality. Choose $\mathbf{x}$ so that $\|A\mathbf{x}\| = \|A\| \, \|\mathbf{x}\|$ and choose $\mathbf{b} - \mathbf{b}_0$ for which $\|A^{-1}(\mathbf{b} - \mathbf{b}_0)\| = \|A^{-1}\|\|\mathbf{b} - \mathbf{b}_0\|$ See the exercises for details. Refering to Figure , this recipe calls for choosing $\mathbf{b}$ along the long axis of the ellipse and $\mathbf{b} - \mathbf{b}_0$ along the short axis. Because the ellipse is narrow, the choice of $\mathbf{b} - \mathbf{b}_0$ is not critical but the choice of $\mathbf{b}$ is more restricted. The narrower the ellipse, that is, the larger the condition number, the more critical the choice of $\mathbf{b}$.

The situation is illustrated in Figure . The actual ratio $k$ in

$$\frac{\|\mathbf{x} - \mathbf{x}_0\|_2}{\|\mathbf{x}\|_2} = k \frac{\|\mathbf{b} - \mathbf{b}_0\|_2}{\|\mathbf{b}\|_2}$$

is plotted as $\mathbf{b}$ varies through $\pi$ radians along the first axis and $\mathbf{b} - \mathbf{b}_0$ varies through $\pi$ radians along the second axis for the matrix of Figure . The values along the first two axes have been chosen to roughly center the peak. The matrix of Figure  has $\kappa_2(A) \approx 13.75$ but for most choices of $\mathbf{b}$ and $\mathbf{b} - \mathbf{b}_0$ the ratio $k$ is much smaller. Matrices with larger condition numbers have ridges that are higher but thinner.

Next consider the problem of solving $A\mathbf{x} = \mathbf{b}$ where $\mathbf{b}$ is known exactly but $A$ is known only approximately. The next proposition is the basic result for this case.

Proposition 1.30 (Perturbing the Left Hand Side). *Let $B$ be a nonsingular and $A$ the same shape as $B$. Let $A\mathbf{x} = \mathbf{b}$ and $B\mathbf{x}_0 = \mathbf{b}$. Then*

$$\frac{\|\mathbf{x} - \mathbf{x}_0\|}{\|\mathbf{x}\|} \le \kappa(B) \frac{\|B - A\|}{\|B\|}.$$

Proof:

$$\begin{aligned}\mathbf{b} = A\mathbf{x} &= (B + (A - B))\mathbf{x} \\ &= B\mathbf{x} + (A - B)\mathbf{x}\end{aligned}$$

Multiplying on the left by $B^{-1}$ gives

$$\begin{aligned}&\mathbf{x}_0 = \mathbf{x} + B^{-1}(A - B)\mathbf{x} \\ &\mathbf{x}_0 - \mathbf{x} = B^{-1}(A - B)\mathbf{x} \\ &\|\mathbf{x} - \mathbf{x}_0\| \le \|B^{-1}\| \, \|A - B\| \, \|\mathbf{x}\| \\ &\frac{\|\mathbf{x} - \mathbf{x}_0\|}{\|\mathbf{x}\|} \le \|B^{-1}\| \, \|A - B\|\end{aligned}$$

and the result follows since $\kappa(B) = \|B^{-1}\| \, \|B\|$ ∎

Example 1.31. *The matrix $B$ below, which comes from physical measurements made with a relative accuracy of $.5 \times 10^{-4}$, is used to solve a linear system $B\mathbf{x} = \mathbf{b}$. Find a bound on the accuracy of the answer obtained.*

$$B = \begin{pmatrix} 3.687 & 9.500 & 5.292 \\ 0.5330 & 0.8194 & 0.8044 \\ 1.500 & 3.278 & 2.194 \end{pmatrix}$$

Solution: The measurements define $B$, an approximation to an exact matrix $A$. If we use the Max-norm then

$$\frac{\|B - A\|_\infty}{\|B\|_\infty} \le .5 \times 10^{-4}.$$

Using exact arithmetic software we find

$$B^{-1} = \frac{1}{3539657} \begin{pmatrix} -2097649000 & -8739560000 & 8263838000 \\ 92995000 & 378195000 & -362967000 \\ 1295185000 & 5410035000 & 5105930500 \end{pmatrix}$$

and the condition number of $B$ is $\kappa_\infty(B) = \|B\|_\infty \, \|B^{-1}\|_\infty = 0.997182 \times 10^5$. It follows that the solution need have no digits correct.

The problem is that $\kappa_\infty(B)^{-1} = 0.10028 \times 10^{-4} < .5 \times 10^{-4}$ and so, by Proposition 1.27, $A$ might be singular given the accuracy of the measurements. For a singular matrix $A$ the system $A\mathbf{x} = \mathbf{b}$ might have no solution or an infinity of solutions with arbitrarily large norms. ∎

The above example assumes that the linear system will be solved using exact arithmetic. More often such systems are solved using approximate arithmetic. As a result further error is introduced during the computations. How does this affect the accuracy of the result?

Suppose then that we have a matrix $A$ known exactly and solve a system $A\mathbf{x} = \mathbf{b}$ using approximate arithmetic. The principle of *Backward Error Analysis*, which we will not attempt to justify, states that the solution obtained by the approximate computations is the *exact* solution to a system $B\mathbf{x}_0 = \mathbf{b}$. To apply Proposition 1.27 we need an estimate of $\kappa(B)$ and the relative error of $B$ as an approximation to $A$. For the most common solution method, Gaussian reduction with partial pivoting, an estimate of the relative error is known and the best software optionally returns an estimate of $\kappa(B)$ in one of the standard norms.

For Gaussian reduction with partial pivoting it is the consensus of numerical analysts that, using the 1- or Max-norms,

$$\frac{\|A - B\|}{\|A\|} \leq C\epsilon_{mach}$$

where $C$ is a constant that is rarely as large as $n$ for an $n \times n$ matrix—although examples can be constructed where $C$ is as large as $2^n$. Replacing the denominator $\|A\|$ by $\|B\|$ does not change the estimate significantly if $C$ is not much larger than $n$.

The estimates of the condition number of $B$ reported by numerical routines are underestimates but are rarely more than one or two orders of magnitude low.

If both the matrix $A$ and the left hand side $\mathbf{b}$ in $A\mathbf{x} = \mathbf{b}$ are approximate then the next propostion applies—$A$ and $\mathbf{b}$ are the unknown exact quantities and $B$ and $\mathbf{c}$ are the approximations. The esimates of $\kappa(A)$ reported by software routines are usually computed from $B$ and so are better estimates of $\kappa(B)$.

Proposition 1.32 (Perturbing Both Sides). *Let* $A\mathbf{x} = \mathbf{b}$ *and* $B\mathbf{y} = \mathbf{c}$ *where* $B$ *is nonsingular. Assume that*

$$\frac{\|A - B\|}{\|B\|} \leq \epsilon \quad \textit{and} \quad \frac{\|\mathbf{b} - \mathbf{c}\|}{\|\mathbf{b}\|} \leq \delta.$$

*Then*

$$\frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{x}\|} \leq \kappa(B)(\epsilon + \delta) + (\kappa(B)\epsilon)(\kappa(B)\delta).$$

Proof: Let $B\mathbf{z} = \mathbf{b}$. By Proposition 1.30,

$$\frac{\|\mathbf{x} - \mathbf{z}\|}{\|\mathbf{x}\|} \leq \kappa(B)\epsilon.$$

By Proposition 1.29

$$\frac{\|\mathbf{z}-\mathbf{y}\|}{\|\mathbf{z}\|} \leq \kappa(B)\delta.$$

By Proposition 1.11,

$$\frac{\|\mathbf{z}\|}{\|\mathbf{x}\|} \leq 1+\kappa(B)\epsilon.$$

Using these estimates in the inequality

$$\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{x}-\mathbf{z}+\mathbf{z}-\mathbf{y}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{x}-\mathbf{z}\|+\|\mathbf{z}-\mathbf{y}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{x}-\mathbf{z}\|}{\|\mathbf{x}\|} + \frac{\|\mathbf{z}\|}{\|\mathbf{x}\|}\frac{\|\mathbf{z}-\mathbf{y}\|}{\|\mathbf{z}\|}$$

gives the result. ∎

If $\epsilon$ and $\delta$ have the same order of magnitude and the order of magnitude of $\kappa(B)\epsilon$ is negative then the second term in the conculsion of this proposition will be negligible compared to the first and we arrive at the rule of thumb stated in the introduction: the order of magnitude of the condition number is the number of digits of accuracy that may be lost in solving a linear system.

The next example shows that the estimate of Proposition 1.32 is often too pessimistic. The example applies the estimate to an example for which the 'exact' answer is known.

Example 1.33. *The unique polynomial of degree 4 through the points (1/2,-121/16), (1/3,-91/27), (1/4,-213/256), (1/5,323/625), and (1/6,557/432) is*

$$p(x) = 3+2x-77x^2+123x^4.$$

*Suppose that these data points are known only with an accuracy of 6 digits. That is, we have the data (.5,-7.5625), (.333333,-3.37037), (.25,-0.832031), (.2,0.516800), (.166667,1.28935), and we estimate the polynomial. How does the accuracy of the the estimation compare with the estimate of Proposition 1.32?*

Solution: The general polynomial of degree 4 is $q(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$. Denoting the above data points by $(x_i, y_i), i = 1,2,3,4,5$, the 5 equations

$$a_0 + a_1x_i + a_2x_i^2 + a_3x_i^3 + a_4x_i^4 = y_i \qquad i = 1,2,3,4,5$$

determine the unknown coefficients. The corresponding matrix equation ($B\mathbf{y} = \mathbf{c}$) is

$$\begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ 1 & x_3 & x_3^2 & x_3^3 & x_3^4 \\ 1 & x_4 & x_4^2 & x_4^3 & x_4^4 \\ 1 & x_5 & x_5^2 & x_5^3 & x_5^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

The statement that the numbers are rounded to 6 digits means that the relative error is at most $.5 \times 10^{-6}$ for each component of the data points. Assume that the powers $x_i^j$ are computed with the same accuracy. it follows that, using either the Max-norm or the 1-norm, which satisfy Property I for norms,we may take $\epsilon = \delta = .5 \times 10^{-6}/(1 - .5 \times 10^{-6}) = .50000250 \ldots \times 10^{-6}$. Using exact arithmetic to compute $B^{-1}$ and $\mathbf{y} = B^{-1}\mathbf{c}$ shows that $\kappa_\infty(B) = .886987 \times 10^5$ and

$$\kappa_\infty(B)(\epsilon + \delta) + (\kappa_\infty(B)\epsilon)(\kappa_\infty(B)\delta) = .867$$

The value of $\mathbf{y}$ is

$$\mathbf{y} = (3.00011, 1.99804, -76.9875, -0.0332275, 123.03).$$

so, except for the coefficient of the cubic term, one or two digits are correct. The actual relative error is $.270 \times 10^{-3}$. ■

Interestingly enough, the graphs of the original polynomial and the computed polynomial agree to four digits in the interval containing the $x$ values of the data. This illustrates the fact that even if the computed solution $\mathbf{x}_0$ of a linear system $A\mathbf{x} = \mathbf{b}$ is not a very accurate approximation of $\mathbf{x}$, it still may be good enough for the purpose because $A\mathbf{x}_0$ is a good approximation to $\mathbf{b}$. This brings up the concept of the residuals of an approximate solution.

Residuals

Approximately solving a linear system $A\mathbf{x} = \mathbf{b}$ we obtain a vector $\mathbf{x}_0$ which is, by the principle of backward error analysis, the exact solution to a system $B\mathbf{x} = \mathbf{b}$. Instead of asking how accurate the approximation $\mathbf{x}_0$ is we can ask how well does it solve the system? That is, how accurate is $A\mathbf{x}_0$ as an approximation to $\mathbf{b}$? It is usually quite accurate. The difference $\mathbf{r} = \mathbf{b} - A\mathbf{x}_0$ is called the *vector of residuals* and it is usually small compared to the size of $\mathbf{b}$ or $\mathbf{b}_0$. Note that the condition number does not enter into the next result.

Proposition 1.34 (Residual error). *Let $A$ and $B$ be $n \times n$ matrices and let $\mathbf{x}$, $\mathbf{x}_0$, $\mathbf{b} \neq \mathbf{0}$, and $\mathbf{b}_0 \neq \mathbf{0}$ be vectors such that $A\mathbf{x} = \mathbf{b}$, $B\mathbf{x}_0 = \mathbf{b}$ and $A\mathbf{x}_0 = \mathbf{b}_0$. Then*

$$\frac{\|\mathbf{b} - \mathbf{b}_0\|}{\|\mathbf{b}\|} \leq \frac{\|B - A\|}{\|B\|} \left( \frac{\|B\|\,\|\mathbf{x}_0\|}{\|B\mathbf{x}_0\|} \right),$$

$$\frac{\|\mathbf{b} - \mathbf{b}_0\|}{\|\mathbf{b}_0\|} \leq \frac{\|B - A\|}{\|A\|} \left( \frac{\|A\|\,\|\mathbf{x}_0\|}{\|A\mathbf{x}_0\|} \right).$$

Proof:

$$\begin{aligned} \frac{\|\mathbf{b} - \mathbf{b}_0\|}{\|\mathbf{b}\|} &= \frac{\|B\mathbf{x}_0 - A\mathbf{x}_0\|}{\|\mathbf{b}\|} = \frac{\|(B - A)\mathbf{x}_0\|}{\|\mathbf{b}\|} \leq \frac{\|B - A\|\,\|\mathbf{x}_0\|}{\|\mathbf{b}\|} \\ &= \frac{\|B - A\|}{\|B\|} \frac{\|B\|\,\|\mathbf{x}_0\|}{\|\mathbf{b}\|} = \frac{\|B - A\|}{\|B\|} \left( \frac{\|B\|\,\|\mathbf{x}_0\|}{\|B\mathbf{x}_0\|} \right) \end{aligned}$$

and similarly for the second inequality. ∎

This proposition is not needed to estimate the size of the residuals since we usually have $\mathbf{b}$ and $\mathbf{b}_0$ at hand. Rather it helps explain why the residual vector is usually small.

First note that for any nonsingular matrix $A$ we have

$$1 \leq \frac{\|A\|\,\|\mathbf{x}\|}{\|A\mathbf{x}\|} = \frac{\|A\|}{\|A(\mathbf{x}/\|\mathbf{x}\|)\|} \leq \kappa(A).$$

In particular, to study this factor we may restrict our attention to unit vectors $\mathbf{x}$. It is possible to find $A$, $B$, and $\mathbf{b}$ so that the above inequality is an equality and the term in parenthesis in the proposition is $\kappa(A)$. This is, however, a rare occurence. Usualy this term is near 1. Figures  and  illustrate this point. Figure  is a graph of the ratio $\|A\|/\|A\mathbf{x}\|$ for the unit vectors $\mathbf{x} = (\cos t, \sin t)$ as $t$ varies from 0 to $\pi$. The matrix is that of Figure  which has a 2-condition number of 14.99 (see exercises). The ratio is above 10 for only 5% of the range of $t$. So in some sense the odds against the order of magnitude of the relative residuals being even two orders of magnitude larger than the relative error in the matrix are better than 10 to 1. Of course $A$ is a well conditioned matrix. Figure  is the corresponding graph for a matrix with a 2-condition number of one-million. It looks about the same. This time the ratio is above 10 for about 7% of the values of $t$. It can be shown that the ratio is never more than 7% for any 2 × 2 matrix.

Exercises

1. Verify the remark about the graphs of Example 1.33 by plotting the difference of the two polynomials on the the interval containing the $x$ data.
2. This exercise shows how to construct an example for which the inequality of Proposition 1.29 is an equality. It uses the fact that for any matrix the maximum $\|A\mathbf{x}\|/\|\mathbf{x}\|$ is attained for some value of $\mathbf{x}$. That is, there is always an $\mathbf{x} \neq \mathbf{0}$ such that $\|A\mathbf{x}\| = \|A\|\|\mathbf{x}\|$.
   a) Given a nonsingular matrix $A$, choose $\mathbf{x}$ such that $\|A\mathbf{x}\| = \|A\|\|\mathbf{x}\|$. Set $\mathbf{b} = A\mathbf{x}$. Choose a vector $\mathbf{y}$ such that $\|A^{-1}\mathbf{y}\| = \|A^{-1}\|\|\mathbf{y}\|$. Set $\mathbf{b}_0 = \mathbf{b} - \mathbf{y}$ and $\mathbf{x}_0 = A^{-1}\mathbf{b}_0$. Show that the inequality of Proposition 1.29 is an equality.
   b) Apply the construction to the matrix

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

   using the Max-norm.

3. Suppose the points $(x_1, y_1) = (1.234, 1)$ and $(x_2, y_2) = (1.235, 100)$ are known with a relative accuracy of $.5 \times 10^{-6}$ in the individual components. Find the straight line $y = a_0 + a_1 x$ through these points. Use Propositions 1.32 and 1.12 to estimate the number of correct digits of the coefficients $a_0$ and $a_1$ (use the Max-norm).