CPSC 303: WHAT THE CONDITION NUMBER DOES AND DOES NOT TELL US

JOEL FRIEDMAN

Contents

Research supported in part by an NSERC grant.

Disclaimer: The material may sketchy and/or contain errors, which I will elaborate upon and/or correct in class. For those not in CPSC 303: use this material at your own risk...

The main goals of this article is to motivate and define the *condition number* of a square matrix, and to explain the usefulness and shortcomings of this definition regarding interpolation. At the end of this article we will define what we call a *bi-normed condition number*, that involves two *norms*, making it much easier to understand what is going on with the examples regarding interpolation.

The condition number is studied in more detail in CPSC 302 and Section 5.8 of the course textbook [A&G] (by Ascher and Greif). It is defined in terms of *relative error*, measured in various ℓ^p*-norms* for $p \geq 1$ (see Section 4.2 of [A&G]). We will review these concepts after some motivating examples.

1. Motivating Examples from Interpolation

Consider fitting data (x_i, y_i) with $i = 0, \ldots, n$ to a polynomial $p(x)$ (we switch from $v(x)$ to $p(x)$ after Section 10.1 of [A&G]) of degree at most n, where $x_0, x_2, x_3, \ldots, x_n$ are fixed and distinct, and $x_1 = x_1(\epsilon) = x_0 + \epsilon$ where we think of $\epsilon > 0$ as very small; eventually we will take a limit as $\epsilon \to 0$.

1.1. An $n = 1$ Example: the Tangent Line of Calculus.

Consider case $n = 1$ in the example of data points $(x_0, y_0), (x_1, y_1)$ where $x_0 = 2$ and $x_1 = 2 + \epsilon$, which gives us the 2×2 system

$$\begin{bmatrix} 1 & 2 \\ 1 & 2+\epsilon \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}, \tag{1}$$

or equivalently

$$\begin{bmatrix} 1 & 2 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 - y_0 \end{bmatrix}, \tag{2}$$

or equivalently

$$\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} y_0 \\ (y_1 - y_0)/\epsilon \end{bmatrix}, \tag{3}$$

The *condition number* of the systems (1) and (2) turn out to be very large (i.e., very bad) for small ϵ, but that of (3) is reasonable. Of course, if $y_0 = f(x_0)$ and $y_1 = f(x_1)$ where f is a differentiable function, then

$$(y_1 - y_0)/\epsilon = \frac{f(2+\epsilon) - f(2)}{\epsilon},$$

whose limit as $\epsilon \to 0$ is $f'(2)$. Hence the $\epsilon \to 0$ limit of (3) when $y_i = f(x_i)$ is

$$\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \lim_{\epsilon \to 0} \begin{bmatrix} f(2) \\ \big(f(2+\epsilon) - f(2)\big)/\epsilon \end{bmatrix} = \begin{bmatrix} f(2) \\ f'(2) \end{bmatrix} \quad \Rightarrow \quad c_1 = f'(2),\ c_0 = f(2) - 2f'(2)$$

which is the line

$$p(x) = \big(f(2) - 2f'(2)\big) + f'(2)x = f(2) + (x-2)f'(2)$$

(you should recognize $f(2) + (x-2)f'(2)$ from calculus as the tangent line to f at $x = 2$).

1.2. **Degeneracy in Interpolation is Not Degenerate and Yields Derivatives.** The "degeneracy" above, i.e., $x_1 = x_0+\epsilon$ with x_0 fixed and $\epsilon \to 0$—and similar "degeneracies" where some of the x_i are infinitesimally close—will be extremely important when we study splines (Chapter 11 of [A&G]) and are an essential part of our discussion of interpolation (Chapter 10 of [A&G]). Furthermore, such degeneracies also arise and are an essential topic in linear interpolation more general than polynomial interpolation (such interpolation is briefly discussed in Section 10.1 of [A&G]).

The main point is that these degenaracies—such as $x_1 = x_0 + \epsilon$—become derivative conditions in the $\epsilon \to 0$ limit. Furthermore, even before we take the limit, the fact that we divide certain equations by ϵ makes the condition number go from bad to good.

It follows that the condition number does not tell us what we really want to know about degenerate interpolation and derivatives, at least until we divide certain equations by ϵ. It is extremely important—once we define and study the condition number—to understand this.

1.3. **An $n = 2$ Example With One Derivative.** A similar example to the last, except with $n = 2$ would be $x_0 = 2$, $x_1 = 2 + \epsilon$, $x_2 = 3$, which yields the system

$$\begin{bmatrix} 1 & 2 & 4 \\ 1 & 2+\epsilon & (2+\epsilon)^2 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}, \tag{4}$$

or equivalently,

$$\begin{bmatrix} 1 & 2 & 4 \\ 1 & \epsilon & 4\epsilon+\epsilon^2 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 - y_0 \\ y_2 \end{bmatrix}, \tag{5}$$

and both systems turn out to be "poorly conditioned" systems for small ϵ; by contrast, the equivalent system

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 2+\epsilon \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ (y_1 - y_0)/\epsilon \\ y_2 \end{bmatrix}$$

is well conditioned; if $y_i = f(x_i)$ then the $\epsilon \to 0$ limit of this system is

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 2 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} f(2) \\ f'(2) \\ f(3) \end{bmatrix}$$

which means that $p(x) = c_0+c_1x+c_2x^2$ is the (unique) polynomial with $p(2) = f(2)$, $p'(2) = f'(2)$, and $p(3) = f(3)$.

1.4. **A Doubly Degenate Example and Taylor Series.** Consider an example with $x_0 = 2$, $x_1 = 2 + \epsilon$, and $x_2 = 2 + 2\epsilon$. This gives the system

$$\begin{bmatrix} 1 & 2 & 4 \\ 1 & 2+\epsilon & (2+\epsilon)^2 \\ 1 & 2+2\epsilon & (2+2\epsilon)^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix};$$

subtracting the first row from the second row and the third row yields

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & \epsilon & 4\epsilon+\epsilon^2 \\ 0 & 2\epsilon & 8\epsilon+4\epsilon^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1-y_0 \\ y_2-y_0 \end{bmatrix};$$

subtracting 2 times the second row from the third row yields

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & \epsilon & 4\epsilon+\epsilon^2 \\ 0 & 0 & 2\epsilon^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1-y_0 \\ y_2-y_0-2(y_1-y_0) \end{bmatrix}; \tag{6}$$

all of the above systems are *extremely* poorly conditioned: once we define the condition number, we will see that all of the above systems in this subsection—which involve a "doubly degenerate"—have condition number of order $1/\epsilon^2$; the systems (1) and (2) (and (5)) will have condition number of order $1/\epsilon$. However, when we divide the second row by ϵ and the third row by ϵ^2, we get the system

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4+\epsilon \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ (y_1-y_0)/\epsilon \\ \big(y_2-y_0-2(y_1-y_0)\big)/\epsilon^2 \end{bmatrix} \tag{7}$$

which is "well-conditioned," i.e., has a moderate condition number. If $f=f(x)$ is twice differentiable at $x=2$, then as $\epsilon\to 0$ we have

$$\begin{bmatrix} y_0 \\ (y_1-y_0)/\epsilon \\ \big(y_2-y_0-2(y_1-y_0)\big)/\epsilon^2 \end{bmatrix} \xrightarrow{\epsilon\to 0} \begin{bmatrix} f(2) \\ f'(2) \\ L \end{bmatrix}$$

where

$$L = \lim_{\epsilon\to 0} \frac{f(2+2\epsilon)-2f(2+\epsilon)+f(2)}{\epsilon^2}.$$

We may compute L either by

(1) applying L'Hôpital's Rule twice:

$$L = \lim_{\epsilon\to 0} \frac{f(2+2\epsilon)-2f(2+\epsilon)+f(2)}{\epsilon^2} = \lim_{\epsilon\to 0} \frac{2f'(2+2\epsilon)-2f'(2+\epsilon)}{2\epsilon}$$
$$= \lim_{\epsilon\to 0} \frac{4f''(2+2\epsilon)-2f''(2+\epsilon)}{2} = \frac{4f''(2)-2f''(2)}{2} = f''(2);$$

or

(2) by using Talyor series:

$$f(2+2\epsilon) = f(2)+2\epsilon f'(2)+\frac{(2\epsilon)^2}{2}f''(2)+o(\epsilon^2)$$
$$f(2+\epsilon) = f(2)+\epsilon f'(2)+\frac{\epsilon^2}{2}f''(2)+o(\epsilon^2)$$

which gives (after some algebra)

$$L = \lim_{\epsilon\to 0} \frac{f(2+2\epsilon)-2f(2+\epsilon)+f(2)}{\epsilon^2} = \lim_{\epsilon\to 0} \frac{\epsilon^2\, f''(2)+o(\epsilon^2)}{\epsilon^2} = f''(2).$$

Hence (1.4) implies that the $\epsilon\to 0$ of (7) is

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} f(2) \\ f'(2) \\ f''(2) \end{bmatrix}.$$

After some algebra, we see that $p(x) = c_0 + xc_1 + x^2c_2$ can be written as

$$p(x) = f(2) + (x-2)f'(2) + \frac{(x-2)^2}{2}f''(2),$$

which you should recognize as the second order Taylor expansion of f about $x = 2$.

2. More Motivation (From General Linear Algebra) for the Condition Number

The condition number is an important consideration in any type of linear algebra, not merely interpolation. In this section we give additional motivation from linear algebra to study the condition number.

2.1. Sensitivity in General $n \times n$ Systems.

The idea behind the condition involves solving an $n \times n$ system of equations $A\mathbf{x} = \mathbf{b}$, where $\mathbf{x}$ is an $n \times 1$ ("column vector") of unknowns, $\mathbf{b}$ is a given $n \times 1$ (the "constants" of the equation or "system of equations"), and A is an $n \times n$ matrix (the "coefficients" of the equation); here we typically work over the real or complex numbers. We typically assume that A is invertible, so that for any $\mathbf{b}$ the system $Ax = b$ has a unique solution $\mathbf{x} = A^{-1}\mathbf{b}$.

In practice we are often interested in all possible values of $\widetilde{A}^{-1}\widetilde{\mathbf{b}}$ where the pairs $(\widetilde{A}, \widetilde{\mathbf{b}})$ range over some values that are "close to" $(A, \mathbf{b})$, and we hope that the values of $\widetilde{A}^{-1}\widetilde{\mathbf{b}}$ do not differ by much from one another. We may also be interested in the "typical range of values" value of $\widetilde{A}^{-1}\widetilde{\mathbf{b}}$ where $(\widetilde{A}, \widetilde{\mathbf{b}})$ range over some *probability distribution*.

Generally speaking (and rather imprecisely) we say that an $n \times n$ system is *sensitive* if a small change in its constants (i.e., in $\mathbf{b}$) and/or in its coefficients (i.e., in A) gives a large change in the solution. Analyzing such changes precisely can be very difficult, but the condition number will measure this to some extent.

Let us give some examples of question regarding sensitivity.

2.2. Sensitivity in General Interpolation.

Example 2.1. We measure three data points $(x_0, y_0), (x_1, y_1), (x_2, y_2)$ which we fit exactly with a polynomial $v(x) = c_0 + c_1x + c_2x^2$, which determine c_0, c_1, c_2 as:

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = A^{-1}\mathbf{b}, \qquad \text{where} \quad \begin{bmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}.$$

However, if there are errors in measuring the data (x_i, y_i), or the model is only an approximation, we may be interested in knowing all possible values of $A^{-1}\mathbf{b}$ defined as above, with each x_i and each y_i replaced by a range of possible values. We may also be interested in the same where each x_i and y_i vary over some probability distribution.

Knowing all values or a typical value in this first example is very difficult; we might settle for an approximate solution or a bound on how "bad" the solution can get.

2.3. **Optimal Sensitivity: Diagonal Systems.**

Example 2.2. We wish to determine all possible values of $\mathbf{x} = A^{-1}\mathbf{b}$ where

$$A = \begin{bmatrix} 10^7 & 0 \\ 0 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 4 \pm 0.04 \\ 3 \pm 0.03 \end{bmatrix}.$$

Therefore A is known exactly, and each value of $\mathbf{b}$ is known to within 1%. (Technically, 4 ± 0.04 refers to the set of real x with $3.96 \le x \le 4.04$.) It is easy to see that each value of $\mathbf{x}$ is known to within 1%, namely

$$x_1 = 10^{-7}(4 \pm 0.04), \quad x_2 = (1/3)(3 \pm 0.03).$$

A similar remark holds whenever A is a fixed diagonal matrix: if each component of $\mathbf{b}$ known to within 1%, or any percent, the solution $\mathbf{x}$ is known to within the same percent. This is an optimal situation; we shall soon see that if A is not diagonal—more precisely when A^{-1} has rows of mixed signs—then the situation is worse.

2.4. **Cancellation.** The next example involves a fact about cancellation: we easily check that the sum of two positive numbers known to within 1% is still known to within 1%, for example

$$(12 \pm 0.12) + (9 \pm 0.09) = 21 \pm 0.21.$$

(Technically the expression 9 ± 0.09 refers to the set of x in the closed interval $[8.91, 9.09]$, i.e., $8.91 \le x \le 9.09$.) However,

$$(12 \pm 0.12) - (9 \pm 0.09) = 3 \pm 0.21 \tag{8}$$

(since the left-hand-side can be as large as

$$(12 + 0.12) - (9 - 0.09) = 3 + 0.21$$

and, similarly, as small as $3 - 0.21$). Hence the difference of two positive quantities known to within 1% are not generally known to within 1% when the main terms have opposite signs.

2.5. **A Mixed Signs Example.**

Example 2.3. We wish to determine all possible values of $\mathbf{x} = A^{-1}\mathbf{b}$ where

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 4 \pm 0.04 \\ 3 \pm 0.03 \end{bmatrix}.$$

The formula for a 2×2 inverse gives

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix},$$

and hence

$$\mathbf{x} = A^{-1} \begin{bmatrix} 4 \pm 0.04 \\ 3 \pm 0.03 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix} \begin{bmatrix} 4 \pm 0.04 \\ 3 \pm 0.03 \end{bmatrix} = \begin{bmatrix} -5 \pm 0.11 \\ 9/2 \pm 0.075 \end{bmatrix},$$

where the -5 ± 0.11 results from the fact that

$$(-2)(4 \pm 0.04) + (1)(3 \pm 0.03)$$

can be as large as

$$(-2)(4 - 0.04) + (1)(3 + 0.03) = -5 + 0.11$$

and, similarly, as small as $-5-0.11$. Hence

$$x_1 = -5 \pm 0.11, \quad x_2 = 9/2 \pm 0.075 = 4.5 \pm 0.075.$$

Notice that in the above example, x_1, x_2 are not determined to within 1% because of the cancellation.

3. Cancellation, Awkwardness and Norms, and Relative Error

In this section we discuss sensivity and cancellation in $\mathbb{R}^n$ or $\mathbb{C}^n$ for $n \geq 2$; in the previous section we dealt with expressions like 9 ± 0.09, which is the case of $\mathbb{R}^n$ with $n = 1$.

The first step to note is that the sets in the previous section, such as

$$\begin{bmatrix} 4 \pm 0.04 \\ 3 \pm 0.03 \end{bmatrix} = \left\{ \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \;\middle|\; 3.96 \leq b_1 \leq 4.04,\ 2.97 \leq b_2 \leq 4.03 \right\} \tag{9}$$

tend to be very awkward to deal with (we'll explain why soon), and matters are worse with their analogs in $\mathbb{R}^n$ with n large. We can work with such sets, but it usually simpler—and almost as powerful—to introduce *norms* (also called *lengths* or *magnitudes*) on $\mathbb{R}^n$ or $\mathbb{C}^n$.

3.1. Common Norms and Awkward Norms.

We tend to work with the following *norms* for n-dimensional vectors $\mathbf{v} = [v_1 \ \dots \ v_n]^T$:

(1) the most common is the 2-*norm* (which coincides with and also called the L^2-norm or ℓ^2-norm):

$$\|\mathbf{v}\|_2 \stackrel{\text{def}}{=} \sqrt{|v_1|^2 + \cdots + |v_n|^2}$$

(writing $|v_1|^2$ instead of v_1^2 makes this expression valid for both $\mathbf{v} \in \mathbb{R}^n$ or $\mathbf{v} \in \mathbb{C}^n$); we use the 2-norm most often becuase is simplifies computations and has a nice "geometric interpretation," even though it is not always the most relevant norm to our given applications;

(2) the 1-*norm*

$$\|\mathbf{v}\|_1 \stackrel{\text{def}}{=} |v_1| + \cdots + |v_n|;$$

and

(3) the *max-norm* or ∞-*norm* (the textbook [A&G] tends to use ∞-norm)

$$\|\mathbf{v}\|_\infty = \|\mathbf{v}\|_{\max} \stackrel{\text{def}}{=} \max\big(|v_1|, \cdots, |v_n|\big).$$

In any norm $\|\cdot\|$, the *distance* between two vectors $\mathbf{v}$ and $\mathbf{u}$ is the length (or norm) of $\mathbf{v} - \mathbf{u}$, i.e., $\|\mathbf{v} - \mathbf{u}\|$.

Notice that the set in (9) is contained in the set

$$\big\{\mathbf{b} \ \big| \ \|\mathbf{b} - [4\ 3]^T\|_\infty \leq 0.04\big\},$$

i.e., the elements of ∞-distance at most 0.04 to $[4\ 3]^T$. Furthermore (9) contains the set

$$\big\{\mathbf{b} \ \big| \ \|\mathbf{b} - [4\ 3]^T\|_\infty \leq 0.03\big\};$$

for these reasons (9) is "pretty well" approximated using the $\|\cdot\|_\infty$ norm.

One could describe the set in (9) exactly by introducing a *weighted* ∞-norm, such as

$$\|\mathbf{v}\|_\infty^{\text{weird weight}} \stackrel{\text{def}}{=} \max\big(|v_1|, (4/3)|v_2|\big) \tag{10}$$

which gives a 4/3 weight to the v_2 component; however, this tends to be rather awkward to work with—each such set may require a different weignt and therefore a different norm.

Later in the course (and elsewhere in the literature) we may see examples of weighted 2-norms where the weights chosen for a good reason (and do not depend on the particular vector, such as $[4\ 3]^T$ in the case above).

3.2. **The L^p-Norm (or ℓ^p-Norm or p-Norm).** The textbook [A&G], Section 4.2, defines for any $p \geq 1$ the p-norm (also known as the L^p-norm or ℓ^p-norm):

$$\|\mathbf{v}\|_p = \big(|v_1|^p + \cdots + |v_n|^p\big)^{1/p}.$$

It also lists some inequalities between the $p = 1, 2, \infty$ norms. Since the ∞-norm is the simplest to calculate (you look at the largest absolute value of the components), one often uses the inequality

$$\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_p \leq n^{1/p}\|\mathbf{v}\|_\infty,$$

which is an easy exercise (namely, if the largest absolute value among the components of $\mathbf{v}$ is M, show that the smallest and largest possible values of $|v_1|^p+\cdots+|v_n|^p$ are, respectively, M^p and nM^p).

The more general equality of this type is that if $1 \leq p \leq r$, then

$$\|\mathbf{v}\|_r \leq \|\mathbf{v}\|_p \leq n^{1/s}\|\mathbf{v}\|_r,$$

where s satisfies $1/r + 1/s = 1/p$. For $p = 1$ and $r = 2$ this yields

$$\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1 \leq n^{1/2}\|\mathbf{v}\|_2,$$

3.3. **Cancellation and the Triangle Inequality.** The cancellation in (8) has an analog for higher dimension. Namely, the analogue of 12 ± 0.12 would be

$$\big\{\widehat{\mathbf{b}} \in \mathbb{R}^n \ \big|\ \|\widehat{\mathbf{b}} - \mathbf{b}\| \leq 0.12\big\}$$

where $\|\cdot\|$ denotes any of the basic norms (i.e., the p-norms with $p = 1, 2, \infty$); the special case $n = 1$ and $\mathbf{b} = 12$ is precisely the set 12 ± 0.12.

The fact that $\|\cdot\|_p$ with $p = 1, 2, \infty$ all satisfy the *triangle inequality*, namely

$$\|\mathbf{a} + \mathbf{b}\|_p \leq \|\mathbf{a}\|_p + \|\mathbf{b}\|_p$$

and all *scale*, i.e., $\|\beta\mathbf{b}\| = |\beta|\|\mathbf{b}\|$ for a scalar β, implies that the sum of any element in

$$\big\{\widehat{\mathbf{b}} \in \mathbb{R}^n \ \big|\ \|\widehat{\mathbf{b}} - \mathbf{b}\|_p \leq 0.12\big\}$$

with one in

$$\big\{\widehat{\mathbf{a}} \in \mathbb{R}^n \ \big|\ \|\widehat{\mathbf{a}} - \mathbf{a}\|_p \leq 0.09\big\}$$

must lie in

$$\big\{\widehat{\mathbf{c}} \in \mathbb{R}^n \ \big|\ \|\widehat{\mathbf{c}} - (\mathbf{a} + \mathbf{b})\|_p \leq 0.21\big\}$$

In (8) we have $n = 1$, $\mathbf{b} = 12$ and $\mathbf{a} = -9$ (subtracting 9 is the same as adding -9), and so the issue is how much cancellation there is in $\mathbf{a} + \mathbf{b}$.

In $\mathbb{R}^n$ with $n = 1$, two vectors are either pointing in the same direction or in opposite directions. For $n \geq 2$, vectors can also point in the same or opposite directions, but there is a range of angles between 0 and 180 degrees.

3.4. **Relative Error.** The condition number is based on *relative error*: In the two equalities,

$$(7 \pm 0.07) + (5 \pm 0.05) = 12 \pm 0.12 \quad \text{and} \quad (7 \pm 0.07) + (-5 \pm 0.05) = 2 \pm 0.12,$$

the two ± 0.12 represent the same *absolute error*, but a difference is in the *relative error*, namely 1% in 12 ± 0.12 and 6% in 2 ± 0.12. (See Section 1.2 of [A&G].)

Definition 3.1. Given two vectors, $\mathbf{b}, \widehat{\mathbf{b}}$ in $\mathbb{R}^n$ or $\mathbb{C}^n$, and fixed $p \geq 1$ (typically $p = 1, 2, \infty$), the *relative error of* $\widehat{\mathbf{b}}$ *with respect to* $\mathbf{b}$ *in the p-norm*, denoted $\mathrm{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b})$, is

$$\mathrm{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b}) = \frac{\|\widehat{\mathbf{b}} - \mathbf{b}\|_p}{\|\mathbf{b}\|_p}$$

For $n = 1$, the relative error does not depend on p; in this case $\mathbf{b} = [b_1]$ and $\|\mathbf{b}\|_p = |b_1|$. For example, regardless of the value of p,

$$\mathrm{Rel}_p(12.12, 12.00) = (1/100) = 1\%,$$
$$\mathrm{Rel}_p(12.00, 12.12) = (1/101) = 0.99009900\ldots\%,$$

and

$$\mathrm{Rel}_p(2.12, 2.00) = (6/100) = 6\%,$$
$$\mathrm{Rel}_p(2.00, 2.12) = (6/101) = 5.940594\ldots\%,$$

For $n \geq 2$, the relative error generally depends on p:

$$\mathrm{Rel}_1\left(\begin{bmatrix}1.01\\1.00\end{bmatrix}, \begin{bmatrix}1.00\\1.00\end{bmatrix}\right) = \left\|\begin{bmatrix}0.01\\0.00\end{bmatrix}\right\|_1 \Bigg/ \left\|\begin{bmatrix}1.00\\1.00\end{bmatrix}\right\|_1 = 0.01/2 = 0.5\%$$

$$\mathrm{Rel}_2\left(\begin{bmatrix}1.01\\1.00\end{bmatrix}, \begin{bmatrix}1.00\\1.00\end{bmatrix}\right) = \left\|\begin{bmatrix}0.01\\0.00\end{bmatrix}\right\|_2 \Bigg/ \left\|\begin{bmatrix}1.00\\1.00\end{bmatrix}\right\|_2 = 0.01/\sqrt{2} = 0.7071\ldots\%,$$

$$\mathrm{Rel}_\infty\left(\begin{bmatrix}1.01\\1.00\end{bmatrix}, \begin{bmatrix}1.00\\1.00\end{bmatrix}\right) = \left\|\begin{bmatrix}0.01\\0.00\end{bmatrix}\right\|_\infty \Bigg/ \left\|\begin{bmatrix}1.00\\1.00\end{bmatrix}\right\|_\infty = 0.01/1 = 1\%.$$

Similarly,

$$\mathrm{Rel}_p\left(\begin{bmatrix}1.01\\0.01\end{bmatrix}, \begin{bmatrix}1.00\\0\end{bmatrix}\right) = \left\|\begin{bmatrix}0.01\\0.01\end{bmatrix}\right\|_p \Bigg/ \left\|\begin{bmatrix}1.00\\0.00\end{bmatrix}\right\|_p = \begin{cases} 2/100 = 2\% & \text{if } p = 1, \\ \sqrt{2}/100 = 1.41\ldots\% & \text{if } p = 2, \text{ and} \\ 1/100 = 1\% & \text{if } p = \infty. \end{cases}$$

3.5. **Further Remarks on Norms.** Roughly speaking, a *norm* on $\mathbb{R}^n$ or $\mathbb{C}^n$ (or on any *vector space*) is a rule that assigns to each element a non-negative *length* (or *norm*) that

(1) *scales*, meaning $\|\alpha\mathbf{v}\| = |\alpha| \, \|\mathbf{v}\|$ for any vector, $\mathbf{v}$, in $\mathbb{R}^n$ or $\mathbb{C}^n$, and any scalar, α, in $\mathbb{R}$ or $\mathbb{C}$ respectively;
(2) satisfies the *triangle inequality*

$$\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$$

for all vectors $\mathbf{v}, \mathbf{w}$; and
(3) is *non-degenerate*, meaning that $\|\mathbf{v}\| = 0$ iff $\mathbf{v} = \mathbf{0} = (0, \ldots, 0)$.

For any real $p \neq 0$ we set

$$\|\mathbf{v}\|_p \stackrel{\text{def}}{=} \left(|v_1|^p + \cdots + |v_n|^p\right)^{1/p};$$

however, for $0 < p < 1$ this "p-norm" fails to satisfy the triangle inequality, so this "p-norm" is not truly a norm. We define the L^∞-norm or ∞-norm as the limit of the p-norm when $p \to \infty$, which turns out to be merely the max-norm. The limit of $\|\mathbf{v}\|_p$ for $p \to 0$ (where the above formula is not a norm) is the geometric-mean

$$\left(|v_1|\,|v_2|\,\ldots,\,|v_n|\right)^{1/n}$$

(one can easily check this using L'Hôpitals Rule and the fact that $(d/d\epsilon)a^\epsilon = \ln(a)a^\epsilon$). The p-norms are useful in theoretical work, but are not as prominent in practical work.

All these norms can generalized by having different positive weights assigned to different coordinates, as in (10). In some applications there is a particular set of weights of importance that weighs some components higher than others.

There are a very large set of possible norms on $\mathbb{R}^n$: in $\mathbb{R}^n$ it turns out that for any *closed, convex* subset, S, that contains $\mathbf{0}$ in its interior and is *symmetric under* $x \mapsto -x$, there is a unique norm $\|\cdot\|$ such that S is the set of $\mathbb{R}^n$ whose norm is at most 1 (and the converse is true). This theorem classifies all possible norms on $\mathbb{R}^n$, and shows that there are a lot of them (think of all the ways of drawing such an S in $\mathbb{R}^2$).

4. The L^2-Norm (or 2-Norm) of a Matrix

If A is an $n \times n$ matrix over the real or complex numbers, then we define the *L^2-norm* (or simply the *2-norm*) *of* A, denoted $\|A\|_2$, can be defined in two equivalent ways:

(1) the square root of the largest eigenvalue of A^*A, where A^* is the transpose of A when working over the reals, and, more generally, the conjugate transpose of A when working over the complex numbers; or, equivalently,

(2) the smallest real number $C > 0$ such that

$$\|A\mathbf{v}\|_2 \le C\|\mathbf{v}\|_2,$$

for any $\mathbf{v} \in \mathbb{R}^n$ or $\mathbf{v} \in \mathbb{C}^n$, where $\|\cdot\|_2$ is the usual 2-norm on vectors

$$\|\mathbf{v}\|_2 \stackrel{\text{def}}{=} \sqrt{|v_1|^2 + \cdots + |v_n|^2}$$

(writing $|v_1|^2$ instead of v_1^2 makes this expression valid for both $\mathbf{v} \in \mathbb{R}^n$ or $\mathbf{v} \in \mathbb{C}^n$).

There is no simple formula for the 2-norm of an $n \times n$ matrix, A, although you could write down a (rather unhelpful) formula for $n = 2$ using the quadratic equation for the characteristic polynomial of A^*A. However, it is not hard to approximate $\|A\|_2$. Here are two simple such approximations:

(1) if M is the maximum absolute value of an entry of an $n \times n$ matrix A, then

$$M \le \|A\|_2 \le nM;$$

(2) if M_2^{row} is the maximum 2-norm of all rows of an $n \times n$ matrix A, then

$$M_2^{\text{row}} \le \|A|_2 \le \sqrt{n}M_2^{\text{row}},$$

and similarly for columns instead of rows.

It turns out that the second approximation is always at least as good as the first, since we easily see that $M \le M_2^{\text{row}} \le \sqrt{n}M$.

Example 4.1. If

$$A = \begin{bmatrix} 6 & 1 \\ 8 & 1 \end{bmatrix},$$

then the simplest bound is

$$8 \le \|A\|_2 \le 2 \cdot 8 = 16$$

based on the fact that the maximum entry of A is 8. The largest 2-norm of a row of A is

$$\sqrt{8^2 + 1^2} = \sqrt{65},$$

and hence

$$\sqrt{65} \le \|A\|_2 \le \sqrt{2}\,\sqrt{65}.$$

Also, the largest 2-norm of a column of A is

$$\sqrt{6^2 + 8^2} = 10,$$

and so

$$10 \le \|A\|_2 \le 10\sqrt{2}.$$

Combining the row and column estimates we get

$$10 \le \|A\|_2 \le \sqrt{2}\,\sqrt{65} \approx 11.4$$

The exact value of $\|A\|$ is the square root of largest eigenvalue of

$$A^*A = \begin{bmatrix} 6 & 8 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 6 & 1 \\ 8 & 1 \end{bmatrix} = \begin{bmatrix} 100 & 14 \\ 14 & 2 \end{bmatrix}$$

which is given by solving

$$\det \begin{bmatrix} 100 - \lambda & 14 \\ 14 & 2 - \lambda \end{bmatrix} = 0$$

which gives $\|A\| = \sqrt{101.9607\ldots} = 10.0975\ldots$

Example 4.2. If A is a *diagonal* $n \times n$ matrix, i.e.,

$$A = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}$$

then there is a simple formula for $\|A\|_2$, namely

$$\|A\|_2 = \max\big(|d_1|, \ldots, |d_n|\big). \tag{11}$$

5. The L^p-Norm (or p-Norm) of a Matrix

More generally, for any value of $p \geq 1$, we define the p-norm (or L^p-norm) of an $n \times n$ matrix, A, denoted by $\|A\|_p$, as the smallest real $C > 0$ for which

$$\|A\mathbf{v}\|_p \leq C\|\mathbf{v}\|_p$$

for all $\mathbf{v} \in \mathbb{R}^n$ or $\mathbb{C}^n$. Hence $\|A\|_p$ is the maximum amount the p-norm of any vector is "stretched" in this norm. One can equivalently write

$$\|A\|_p = \max_{\mathbf{v}\neq\mathbf{0}} \frac{\|A\mathbf{v}\|_p}{\|\mathbf{v}\|_p}$$

(the fact that we write "max" and not "sup" is a bit subtle).

Similarly, if $\|\cdot\|$ is any *norm* (see Subsection 3.5) on $\mathbb{R}^n$ or $\mathbb{C}^n$, then there is a corresponding norm on $n \times n$ matrices A given as the smallest real number $C > 0$ such that

$$\|A\mathbf{v}\| \leq C\|\mathbf{v}\|,$$

for any $\mathbf{v} \in \mathbb{R}^n$ or $\mathbf{v} \in \mathbb{C}^n$.

If A happens to be a diagonal matrix with diagonal entries $d_1, \ldots, d_n$, then for any $p \geq 1$

$$\|A\|_p = \max\bigl(|d_1|, \ldots, |d_n|\bigr). \tag{12}$$

which is the analog of (11).

The easy upper and lower bounds on $\|A\|_p$, valid for any $p \geq 1$, are somewhat curious: the maximum entry bound is the same, but the row and column bounds involve the unique $q \geq 1$ such that $(1/p) + (1/q) = 1$; this q is often called the *conjugate of* p in this context; for example $p = 2$ gives $q = 2$ so 2 is its own conjugate, and $p = 1$ gives $q = \infty$, and vice versa, so 1 and ∞ are conjugates. Here are the bounds:

(1) if M is the maximum absolute value of an entry of an $n \times n$ matrix A, then
$$M \leq \|A\|_p \leq nM;$$
(2) if M_p^{col} is the maximum p-norm among all columns of an $n \times n$ matrix A, then
$$M_p^{\text{col}} \leq \|A\|_p \leq n^{1/q} M_p^{\text{col}},$$
and this is always at least as good as $M \leq \|A\|_2 \leq nM$ in view of the inequality $M \leq M_p^{\text{col}} \leq n^{1/p}M$;
(3) if M_q^{row} is the maximum q-norm among all columns of an $n \times n$ matrix A, then
$$M_q^{\text{row}} \leq \|A\|_p \leq n^{1/p} M_q^{\text{row}},$$
and this is always at least as good as $M \leq \|A\|_2 \leq nM$ in view of the inequality $M \leq M_q^{\text{row}} \leq n^{1/q}M$.

Since $p = 1$ and $q = \infty$ are conjugates, i.e., $(1/p)+(1/q) = 1$, and since $n^{1/\infty} = 1$ (to see this rigorously we should really take a limit of $p, q > 1$ and $p \to 1$ which takes $q \to \infty$, but we soon get used to what $q = \infty$ should mean) we have

$$\|A\|_\infty = A_1^{\text{row}}, \quad \|A\|_1 = A_1^{\text{col}}. \tag{13}$$

Hence it is generally easier to calculate $\|A\|_p$ exactly for $p = 1, \infty$ than for $p = 2$.

6. Some 2×2 Matrix Norm Formulas

To make formulas like (13) concrete, it is helpful to look at the $n = 2$ case,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

where a, b, c, d are real or complex numbers. In this case

$$\|A\|_\infty = \max(|a| + |b|, |c| + |d|), \quad \|A\|_1 = \max(|a| + |c|, |b| + |d|).$$

It is not hard to prove these when you consider that

$$\|A\|_p \stackrel{\text{def}}{=} \max_{\mathbf{v} \neq \mathbf{0}} \frac{\|A\mathbf{v}\|_p}{\|\mathbf{v}\|_p} = \max_{\|\mathbf{v}\|_p = 1} \|A\mathbf{v}\|_p = \max_{\|\mathbf{v}\|_p \leq 1} \|A\mathbf{v}\|_p$$

(the equality in the middle follows from the fact that if $\mathbf{v} \neq 0$, then one can scale $\mathbf{v}$ to make it a unit vector). [The proofs may appear in the exercises in the final draft of this article.]

Determining $\|A\|_2$ is generally much more difficult: when A has real entries we have

$$A^*A = \begin{bmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{bmatrix}$$

Then $\|A\|_2$ is the square root of the largest eigenvalue of the above matrix; the above matrix is of the form

$$\begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}$$

which leads us to solve

$$\det \begin{bmatrix} \alpha - \lambda & \beta \\ \beta & \gamma - \lambda \end{bmatrix} = 0$$

which gives the quadratic equation

$$(\alpha - \lambda)(\gamma - \lambda) - \beta^2 = 0$$

There is one special case where the above formula is simple, namely when $\gamma = \alpha$, and we find $\lambda = \alpha \pm \beta$; this special case occurs for A^*A iff $a^2 + c^2 = b^2 + d^2$; however, if $\alpha \neq \gamma$ we do not know a particularly nice formula.

To approximate $\|A\|_2$ one can use the inequality

$$M \leq \|A\|_p \leq 2M, \quad \text{where} \quad M = \max\big(|a|, |b|, |c|, |d|\big),$$

which is valid for any $p \geq 1$. There are two sharper inequalities, namely

$$M_2^{\text{row}} \leq \|A|_2 \leq \sqrt{2} M_2^{\text{row}},$$

where

$$M_2^{\text{row}} = \max\Big(\sqrt{a^2 + b^2}, \sqrt{c^2 + d^2}\Big),$$

and

$$M_2^{\text{col}} \leq \|A|_2 \leq \sqrt{2} M_2^{\text{col}},$$

where

$$M_2^{\text{col}} = \max\Big(\sqrt{a^2 + c^2}, \sqrt{b^2 + d^2}\Big).$$

7. The (L^2- and) L^p-Condition Number of a Matrix

If A is an $n \times n$ matrix, over $\mathbb{R}$ or $\mathbb{C}$ as usual, and $p \geq 1$, then the *L^p-conditiion number of A* is defined whenever A is invertible, and either of the equivalent real numbers (which is always at least as large as 1):

(1)
$$\mathrm{cond}_p(A) = \|A\|_p \left\|A^{-1}\right\|_p,$$
or equivalently,

(2) the smallest real $C > 0$ such that for all $\mathbf{b}, \widehat{\mathbf{b}}$ we have
$$\mathrm{Rel}_p(A^{-1}\widehat{\mathbf{b}}, A^{-1}\mathbf{b}) \leq C\, \mathrm{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b})$$

(3) the smallest real $C > 0$ such that for all $\mathbf{b}, \widehat{\mathbf{b}}$ we have
$$\mathrm{Rel}_p(\widehat{\mathbf{x}}, \mathbf{x}) \leq C\, \mathrm{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b}), \quad \text{where} \quad \widehat{\mathbf{x}} = A^{-1}\widehat{\mathbf{b}},\ \mathbf{x} = A^{-1}\mathbf{b}.$$

(4) the smallest real $C > 0$ such that for all $\mathbf{b}, \widehat{\mathbf{b}}$ we have
$$\mathrm{Rel}_p(\widehat{\mathbf{x}}, \mathbf{x}) \leq C\, \mathrm{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b}), \quad \text{where} \quad A\widehat{\mathbf{x}} = \widehat{\mathbf{b}},\ A\mathbf{x} = \mathbf{b}.$$

Of course, these last three formulations of the condition number are simple rewritings of one another; the last formulation is the most suggestive of our definition of the *doubly-normed condition number* that we describe in Section 9. The equivalence of the first formulation with the last three is given in the exercises.

Let us remark that $\mathrm{cond}_p(A) \geq 1$, and there are two simple reasons why this is true (for any norm, including all p-norms with $p \geq 1$):

(1) for any vector $\mathbf{b}$, the relative error of $(1.01)\mathbf{b}$ and $\mathbf{b}$ (in this order, since the order matters) is 1%, in any norm (including all the p-norms for $p \geq 1$). So setting $\widehat{\mathbf{b}} = (1.01)\mathbf{b}$ we have
$$\mathrm{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b}) = 1\%.$$
But since $A^{-1}\widehat{\mathbf{b}} = A^{-1}((1.01)\mathbf{b}) = (1.01)A^{-1}\mathbf{b}$, we have
$$\mathrm{Rel}_p(A^{-1}\widehat{\mathbf{b}}, A^{-1}\mathbf{b}) = \mathrm{Rel}_p\big((1.01)A^{-1}\mathbf{b}, A^{-1}\mathbf{b}\big) = 1\%.$$
Hence if the relative error in $\mathbf{b}$ is 1%, you cannot do better than a relative error of 1% in $\mathbf{x} = A^{-1}\mathbf{b}$.

(2) Another proof that $\mathrm{cond}_p(A) \geq 1$ for any $p \geq 1$ follows from the inequality
$$\|AB\|_p \leq \|A\|_p \, \|B\|_p$$
that can be proven by noting that for any $\mathbf{v}$,
$$\|(AB)\mathbf{v}\|_p \leq \|A(B\mathbf{v})\|_p \leq \|A\|_p \|B\mathbf{v}\|_p \leq \|A\|_p \, \|B\|_p \|\mathbf{v}\|_p.$$
Setting $B = A^{-1}$ in the above inequality, we have
$$\mathrm{cond}_p = \|A\|_p \, \|A^{-1}\|_p \geq \|AA^{-1}\|_p = \|I\|_p$$
where I is the identity matrix, whose norm is easily checked to equal 1 (since $I\mathbf{v} = \mathbf{v}$ for all $\mathbf{v}$ and/or since I is a diagnoal matrix whose entries are all 1's).

Notice that we are using p-norms and p-condition numbers, but these proofs hold for any norm on $\mathbb{R}^n$ or $\mathbb{C}^n$ (see Subsection 3.5), which give rise to a norm on $n \times n$ matrices, A (measuring by how much A stretches the norm), and therefore to condition numbers (either defined by the loss of relative error in solving $A\mathbf{x} = \mathbf{b}$, or as simply $\|A\| \, \|A^{-1}\|$.

Section 5.8 of the textbook [A&G] discusses condition numbers at length; typically one sees the values of $p = 1, 2, \infty$ for condition numbers, but this can vary.

Whenever A is a diagonal matrix with diagonal entries $d_1, \ldots, d_n$, we have that A^{-1} is the diagonal matrix with diagonal entries $1/d_1, \ldots, 1/d_n$, and

$$\text{cond}_p(A) = \frac{\max\big(|d_1|, \ldots, |d_n|\big)}{\min\big(|d_1|, \ldots, |d_n|\big)} \tag{14}$$

for any $p \geq 1$. Furthermore, the bounds $M \leq \|A\|_p \leq Mn$, where M is the largest absolute value of the entries of A (valid for any $n \times n$ matrix, A, and any p) imply that

$$M \, M' \leq \text{cond}_p(A) \leq n^2 M \, M' \tag{15}$$

where M' is the largest absolute value of the entries of A^{-1}. This is a crude bound, but sufficient in studying interpolation when n is (small and) fixed, and we have a degeneracy like $x_1 = x_0 + \epsilon$ and we consider $\epsilon \to 0$.

8. Interpolation: What The Condition Number Does and Does Not Tell Us

At this point we will compute the condition numbers of the matrices in Section 1 and understand why equivalent linear systems can have drastically different condition numbers.

8.1. **Formulas for the Inverse.**

It will be useful to recall the formula

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \Rightarrow \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

It will also be helpful to see that if

$$M = \max\big(|a|, |b|, |c|, |d|\big)$$

(i.e., the largest absolute value among the entries of A), then the largest absolute value among the entries of A^{-1} is just

$$M' = M/\big|ad - bc\big| = M/\big|\det(A)\big|$$

This should affirm our intuition that if, roughly speaking, $\det(A)$ is "very small" compared to M, then A^{-1} will have "very large" entries compared to those of A.

Notice that the generalization of such formulas to $n \times n$ matrices with $n \geq 3$ is not as simple: the general formula is

$$A^{-1} = \frac{1}{\det(A)} \, \text{adjugate}(A),$$

where the *adjugate* of A is formed by the determinants of the $(n-1) \times (n-1)$ minors of A, placed in the appropriate positions and given appropriate signs (i.e., ± 1).

Fortunately, we can get a lot of intuition about condition numbers and "degenerate" interpolation from the case of $n = 2$ and the tangent line.

8.2. **The 2×2 Interpolation for a Tangent Line.**

Example 8.1. Let A be the matrix in (1), namely

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 2+\epsilon \end{bmatrix}$$

Then

$$A^{-1} = \frac{1}{\epsilon}\begin{bmatrix} 2+\epsilon & -2 \\ -1 & 1 \end{bmatrix}$$

Hence the largest entries of A and A^{-1} are respectively

$$M = M(A) = 2+\epsilon, \quad M' = M'(A) = (1/\epsilon)(2+\epsilon),$$

and hence

$$MM' = (1/\epsilon)(2+\epsilon)^2.$$

To simplify this expression, notice that when $\epsilon \to 0$ we have

$$MM' = 4/\epsilon + 4 + \epsilon = 4/\epsilon + O(1),$$

where $O(1)$ refers to a term bounded by 1 times some constant when ϵ is sufficiently small (you can take any constant strictly greater than 4, such as 5 or 4.000001). You could also write

$$MM' = 4/\epsilon + 4 + O(\epsilon)$$

which is more precise than $4/\epsilon + O(1)$, but for now let's just look at the simpler expression $4/\epsilon + O(1)$.

It follows from (15) that

$$MM' \le \mathrm{cond}_p(A) \le 2^2 MM'$$

(which is valid for all $p \ge 1$!), and so

$$4/\epsilon + O(1) \le \mathrm{cond}_p(A) \le 16/\epsilon + O(1).$$

It follows that as $\epsilon \to 0$, the condition number of A *grows as* (meaning within a multiplicative factor of) $1/\epsilon$. In particular, the $\mathrm{cond}_p(A)$ becomes infinite as $\epsilon \to 0$.

Example 8.2. Let A be the matrix in (2), namely

$$A = \begin{bmatrix} 1 & 2 \\ 0 & \epsilon \end{bmatrix}$$

Here $\det(A) = 1/\epsilon$ just like in the example above, and $M = M(A) = 2$ (exactly). Since this M is also equal to $2 + O(\epsilon)$, which is all that we needed above, we get the condition number of A again grows like $1/\epsilon$.

Example 8.3. Let A be the matrix in (3), namely

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

Then

$$A^{-1} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix},$$

and hence $M(A) = M'(A) = 2$. Hence for all $p \ge 1$, (15) implies that

$$4 \le \mathrm{cond}_p(A) \le 16.$$

Here is one way to understand the difference between the last example (with a bounded condition number as $\epsilon \to 0$) and the first two examples: the condition number is the smallest C such that

$$\mathrm{Rel}_p(A^{-1}\widehat{\mathbf{b}}, A^{-1}\mathbf{b}) \le C\,\mathrm{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b})$$

When $x_0 = 2$ and $x_1 = 2+\epsilon$, and $y_i = f(x_i)$ for $i = 1, 2$, then the constants of the equation in (3) are

$$\begin{bmatrix} y_0 \\ (y_1 - y_0)/\epsilon \end{bmatrix} = \begin{bmatrix} f(2) \\ \big(f(2+\epsilon) - f(2)\big)/\epsilon \end{bmatrix} \approx \begin{bmatrix} f(2) \\ f'(2) \end{bmatrix}$$

So when we expect a small relative error in

$$\mathbf{b} = \begin{bmatrix} y_0 \\ (y_1 - y_0)/\epsilon \end{bmatrix}$$

we get a bounded C. On the other hand, the relative error of

$$\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \approx \begin{bmatrix} f(2) \\ f(2) + \epsilon f'(2) \end{bmatrix} \approx \begin{bmatrix} f(2) \\ f(2) \end{bmatrix}$$

is of percentage 1% (for example), then you are fitting a line to the two points

$$\big(2, f(2) \pm f(2)/100\big), \quad \big(2+\epsilon, f(2) + O(\epsilon) \pm f(2)/100\big).$$

In this case the $\pm f(2)/100$ dominates $O(\epsilon)$ for small ϵ (if $f(2) \neq 0$), and for ϵ small you cannot confidently know even if the slope you get for the interpolating line is positive or negative.

In the next section we describe how to rectify this problem in a conceptually simpler way.

8.3. **A 3×3, Doubly Degenerate Example.** We remark that in the 2×2 cases above, looking at M and $\det(A)$ tells us the whole story. This is not true of 3×3 systems.

Example 8.4. Let A be the matrix in (6), namely

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 0 & \epsilon & 4\epsilon + \epsilon^2 \\ 0 & 0 & 2\epsilon^2 \end{bmatrix}.$$

Since A is upper triangular, $\det(A)$ is simply the product of its diagonal entries, and so

$$\det(A) = 1 \cdot \epsilon \cdot 2\epsilon^2 = 2\epsilon^3;$$

however, the condition number grows only like $1/\epsilon^2$, not $1/\epsilon^3$: to see this we can

(1) compute A^{-1} (not necessarily the easiest way);
(2) look at how each entry of the adjugate grows (i.e., look at how the determinant of each 2×2 minor of A—obtained by crossing off one row and one column—grows).

The adjugate method works as follows: crossing off the bottom row of A gives

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & \epsilon & 4\epsilon + \epsilon^2 \end{bmatrix}$$

and we see that all determinants formed by crossing out one column grow like ϵ; crossing out any other row leaves at least one row whose entries grow like ϵ or ϵ^2, we see that largest entry of the adjugate grows like ϵ. Hence M' grows like $\det(A)$

times ϵ, which grows like $1/\epsilon^2$. Hence, in rough terms, MM' grows like $1/\epsilon^2$, and hence so does $\mathrm{cond}_p(A)$ for any $p \geq 1$.

Alternatively, an exact computation of A^{-1} shows that

$$A^{-1} = \begin{bmatrix} 1 & -2/\epsilon & 2/\epsilon^2 + 1/\epsilon \\ 0 & 1/\epsilon & -2/\epsilon^2 + (1/2)(1/\epsilon) \\ 0 & 0 & (1/2)(1/\epsilon^2) \end{bmatrix}$$

(at least, this is what the software `Maple` gives as the inverse...), showing that $M' = 2/\epsilon^2 + O(1/\epsilon)$ as $\epsilon \to 0$.

9. The Doubly-Normed Condition Number

One way to understand all the interpolation examples that we have given is that when we write $Ax = b$, and then the $\mathrm{cond}_p(A)$ (for some $p \geq 1$) is the smallest C such that

$$\mathrm{Rel}_p(\widehat{\mathbf{x}}, \mathbf{x}) \leq C \, \mathrm{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b}), \quad \text{where} \quad A\widehat{\mathbf{x}} = \widehat{\mathbf{b}}, \ A\mathbf{x} = \mathbf{b}.$$

Although $\widehat{\mathbf{x}}, \mathbf{x}$ and $\widehat{\mathbf{b}}, \mathbf{b}$ both refer to n-dimensional vectors, sometimes we want to one norm on the $\mathbf{x}$ vectors and one on the $\mathbf{b}$ vectors. Hence it can be better to define the condition number, $\mathrm{cond}(A)$, as the smallest C such that

$$\mathrm{Rel}_{\mathrm{dom}}(\widehat{\mathbf{x}}, \mathbf{x}) \leq C \, \mathrm{Rel}_{\mathrm{range}}(\widehat{\mathbf{b}}, \mathbf{b}), \quad \text{where} \quad A\widehat{\mathbf{x}} = \widehat{\mathbf{b}}, \ A\mathbf{x} = \mathbf{b},$$

where we have two norms, $\|\mathbf{x}\|_{\mathrm{dom}}$, $\|\mathbf{b}\|_{\mathrm{range}}$, one appropriate for vectors in the domain of A (viewed as a map), and another one appropriate for vectors in the range (i.e., codomain) of A and the relative errors are measured in these two norms, i.e.,

$$\mathrm{Rel}_{\mathrm{dom}}(\widehat{\mathbf{x}}, \mathbf{x}) = \frac{\big\|\widehat{\mathbf{x}} - \mathbf{x}\big\|_{\mathrm{dom}}}{\big\|\mathbf{x}\big\|_{\mathrm{dom}}}$$

and similarly for $\mathrm{Rel}_{\mathrm{range}}$.

In the above sitation one can prove that the condition number is also given by the formula:

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

where the norms are computed with respect to the two norms, i.e.,

$$\|A\| = \max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_{\mathrm{range}}}{\|\mathbf{x}\|_{\mathrm{dom}}} = \max_{\|\mathbf{x}\|_{\mathrm{dom}}=1} \|A\mathbf{x}\|_{\mathrm{range}},$$

$$\|A^{-1}\| = \max_{\mathbf{b} \neq 0} \frac{\|A\mathbf{b}\|_{\mathrm{dom}}}{\|\mathbf{b}\|_{\mathrm{range}}} = \max_{\|\mathbf{b}\|_{\mathrm{range}}=1} \|A\mathbf{b}\|_{\mathrm{dom}}.$$

The point is that merely because the domain and range consist of n-dimensional vectors, we do not necessarily want to measure their "magnitude" or "length" (regarding errors in their measurements, or *perturbations*) in the same way. So if $\mathbf{b} = (y_0, y_1)$ are measured with the same relative error, then any of the usual ℓ^p-norms (or p-norms) may be appropriate. In this case the condition number tends to infinity as $\epsilon \to 0$ for $x_1 = x_0 + \epsilon$ and $x_0 = 2$ (or any fixed value).

On the other hand, if we are working with information on $f(2)$ and $f'(2)$, the more appropriate norm is the one where y_0, y_1 are obtained as

$$y_0 = b_1, \ y_1 = b_1 + \epsilon b_2$$

with b_1, b_2 representing the values of $f(2), f'(2)$ with some small relative errors in these values. Hence the appropriate norm would be

$$\big\|(b_1, b_2)\big\|_p$$

for some p, which equals

$$\Big\|\big(y_0, (y_1 - y_0)/\epsilon\big)\Big\|_p$$

For example, for $p = \infty$ we get the norm

$$\max\Big(|y_0|, |y_1 - y_0|/\epsilon\Big).$$

Another way of seeing the problem with the condition number is that the set of (y_0, y_1) that are of a given magnitude in a p-norm is a larger set than those such that $(y_0, (y_1 - y_0)/\epsilon)$ are of a given p-norm (for ϵ "very small"). So although the condition number refers to the worst case over all $\widehat{\mathbf{b}}, \mathbf{b}$ (see the formulas at the beginning of Section 7), in the "degenerate interpolation" we are only interested in a small subset of them (which we measure differently).

A more general principle is that the condition number speaks of the worst case, whereas a specific application may involve cases that are not so bad; the diagonal Example 2.2 is a good example: its condition number is $10^7/3$, whereas there is no loss in the relative error in the solution (i.e., in $\mathbf{x}$) over those in the constants (i.e., $\mathbf{b}$).

In this way, the doubly-normed condition number gives us a better overall understanding of the situtation:

(1) as always, $\mathrm{cond}_p(A)$, is a way of indicating *possible* problems with the system $A\mathbf{x} = \mathbf{b}$, although it is giving the worst case $\mathbf{b}, \widehat{\mathbf{b}}$;
(2) when there are more appropriate norms of the domain and range of A, you get a better idea of the worst case situation of relative error (or perturbation) in exact arithmetic;
(3) if your domain and/or range norms are not closely related to the usual p-norms, for numerical computations you should write an equivalent linear systems where the domain and range norms are comparable with the usual p-norms (which will change the condition number); and
(4) [something we haven't mentioned yet, but is related to the above remarks:] if you are only interested in a limited set of cases of $\mathbf{b}, \widehat{\mathbf{b}}$, any condition number may be unduly pessimistic and not particularly appropriate.

Furthmore—as we have vaguely mentioned—for some applications there is a natural different norm, such as a *weighted p-norm*, e.g.,

$$\|\mathbf{v}\|_p = \big(|v_1|^p w_1 + \cdots + |v_n|^p w_p\big)^{1/p}$$

for some real $w_1, \ldots, w_p$, generally positive but not the same. Also, for some applications there may be no single norm that actually describes what you want to study.

Exercises

(1) Use the formulas in Section 6 to find the 1-norm, 2-norm, and ∞-norm of the matrix

$$A = \begin{bmatrix} 3 & 4 \\ 4 & 3 \end{bmatrix}.$$

(2) Use the formulas in Section 6 to find the 1-norm, 2-norm, and ∞-norm of any matrix of the form

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix}.$$

[Hint: it turns out that these norms are all the same, and equal to $|a|+|b|$.]

(3) Use the formula in Exercise 2 to find the 1-norm, 2-norm, and ∞-condition number of the matrix

$$A = \begin{bmatrix} 3 & 4 \\ 4 & 3 \end{bmatrix}.$$

(4) Consider the system in Example 2.2: $A\mathbf{x} = \mathbf{b}$ with

$$A = \begin{bmatrix} 10^7 & 0 \\ 0 & 3 \end{bmatrix}.$$

Show your work in the following calculations:

(a) If $\mathbf{b} = [1\ 0]^T$, what is $\mathbf{x} = A^{-1}\mathbf{b}$?

(b) Let $y \in \mathbb{R}$ be any number. If $\widehat{\mathbf{b}} = [1\ y]^T$, what is $\widehat{\mathbf{x}} = A^{-1}\widehat{\mathbf{b}}$?

(c) Show that the relative error between $\widehat{\mathbf{b}}$ and $\mathbf{b}$ (see Definition 3.1) is

$$\mathrm{Rel}_\infty(\widehat{\mathbf{b}}, \mathbf{b}) = |y|.$$

(d) Show that

$$\mathrm{Rel}_\infty(\widehat{\mathbf{x}}, \mathbf{x}) = |y|10^7/3.$$

(e) How does the ratio of the relative errors in the previous two parts relate to the ∞-condition number of A?

(f) Show that

$$\mathrm{Rel}_\infty(\mathbf{b}, \widehat{\mathbf{b}}) = \frac{|y|}{\max(1, |y|)}$$

(g) Show that

$$\mathrm{Rel}_\infty(\mathbf{x}, \widehat{\mathbf{x}}) = \frac{|y|/3}{\max(10^{-7}, |y|/3)}$$

(h) Compute the ratio of the relative errors in the previous two parts for the four values $y = 10, 1, 10^{-1}, 10^{-8}$; how do these values relate to the ∞-condition number of A?

(5) Let $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$, and let $M = \|\mathbf{v}\|_\infty$, i.e.,

$$M = \max\big(|v_1|, \dots, |v_n|\big).$$

Let p be any real number with $p \neq 0$.

(a) **In at most 20 words,** explain why

$$M^p \leq |v_1|^p + \cdots + |v_n|^p.$$

(b) **In at most 20 words,** explain why

$$|v_1|^p + \cdots + |v_n|^p \leq M^p n.$$

(c) **In at most 20 words plus one or two formulas,** using parts (a) and (b), explain why

$$\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_p \leq n^{1/p}\|\mathbf{v}\|_\infty.$$

(6) Let $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$, and let $\mathbf{1}$ denote the vector $(1, 1, \ldots, 1) \in \mathbb{R}^n$. Let $x \in \mathbb{R}$ be a variable.
   (a) Write an expression for
   $$f(x) \stackrel{\text{def}}{=} \|\mathbf{v} - x\mathbf{1}\|_2^2$$
   as a polynomial of degree 2 in x: what are the coefficients of this polynomial?
   (b) Differentiate the above expression in x to determine for which x we have $f'(x) = 0$.
   (c) **In at most 20 words**, explain why the x for which $f'(x) = 0$ is a *(global) minimum* for x.

(7) Let $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ such that
$$v_1 \leq v_2 \leq \ldots \leq v_n,$$
and let $\mathbf{1}$ denote the vector $(1, 1, \ldots, 1) \in \mathbb{R}^n$. Let $x \in \mathbb{R}$ be a variable. Let $v_{\text{middle}} = (v_1 + v_n)/2$, and $r = (v_n - v_1)$.
   (a) **In at most 20 words plus one or two formulas**, show that if $x < v_{\text{middle}}$, then
   $$|x - v_n| > r/2.$$
   (b) **In at most 20 words plus one or two formulas**, show that if $x > v_{\text{middle}}$, then
   $$|x - v_1| > r/2.$$
   (c) **In at most 20 words plus one or two formulas**, show that if $x = v_{\text{middle}}$, then
   $$|x - v_1| = |x - v_n| = r/2$$
   (d) **In at most 20 words plus one or two formulas**, find the value of $x \in \mathbb{R}$ at which
   $$f(x) = \|\mathbf{v} - x\mathbf{1}\|_\infty$$
   attains its minimum.

(8) Let $n \geq 3$ be an odd integer. Let $n_{\text{mid}} = (n + 1)/2$. Let $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ such that
$$v_1 \leq v_2 \leq \ldots \leq v_n,$$
and let $\mathbf{1}$ denote the vector $(1, 1, \ldots, 1) \in \mathbb{R}^n$. Let $x \in \mathbb{R}$ be a variable.
   (a) **In at most 30 words plus one to three formulas**, show that if for some integer $i \leq n_{\text{mid}} - 1$ we have $v_i \leq x < v_{i+1}$, then if $x' = v_{i+1}$ we have
   $$\|\mathbf{v} - x'\mathbf{1}\|_\infty < \|\mathbf{v} - x\mathbf{1}\|_\infty.$$
   [Hint: It may be helpful to first consider a special case, such as $n = 5$, $n_{\text{mid}} = 3$. In this case either $v_1 \leq x < v_2$ or $v_2 \leq x < v_3$.]
   (b) **In at most 30 words plus one to three formulas**, show that if $x < v_1$ and $x' = v_1$, then
   $$\|\mathbf{v} - x'\mathbf{1}\|_\infty < \|\mathbf{v} - x\mathbf{1}\|_\infty.$$

(c) **In at most 20 words**, use the previous two parts to show that if $x_{\text{median}} = v_{n_{\text{mid}}}$ and $x < x_{\text{median}}$, then

$$\|\mathbf{v} - x_{\text{median}}\mathbf{1}\|_\infty < \|\mathbf{v} - x\mathbf{1}\|_\infty.$$

(d) Show that if $x > x_{\text{median}}$, then

$$\|\mathbf{v} - x_{\text{median}}\mathbf{1}\|_\infty < \|\mathbf{v} - x\mathbf{1}\|_\infty.$$

(9) Let $n \geq 2$ be an even integer. Let $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ such that

$$v_1 \leq v_2 \leq \ldots \leq v_n,$$

and let $\mathbf{1}$ denote the vector $(1, 1, \ldots, 1) \in \mathbb{R}^n$. At which values of $x \in \mathbb{R}$ does

$$f(x) = \|\mathbf{v} - x\mathbf{1}\|_\infty$$

attain its minimum values? Explain you answer by modifying the argument in the previous problem (where n is odd).

(10) Let A be an $n \times n$ matrix with real entries, and let $\widehat{\mathbf{b}}, \mathbf{b} \in \mathbb{R}^n$ with $\widehat{\mathbf{b}} \neq \mathbf{b}$. For any $p \geq 1$, show that

$$\frac{\|A^{-1}\widehat{\mathbf{b}} - A^{-1}\mathbf{b}\|_p}{\|\widehat{\mathbf{b}} - \mathbf{b}\|_p} \frac{\|\mathbf{b}\|_p}{\|A^{-1}\mathbf{b}\|_p} \leq \|A^{-1}\|_p \|A\|_p.$$

Then use this to show that

$$\text{Rel}_p(A^{-1}\widehat{\mathbf{b}}, A^{-1}\mathbf{b}) \leq C \,\text{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b})$$

where $C = \|A^{-1}\|_p \|A\|_p$.

(11) Let A be an $n \times n$ real matrix, $p \geq 1$, and $\mathbf{b}_1, \mathbf{b}_2$ satisfy

$$\frac{\|A\mathbf{b}_1\|_p}{\|\mathbf{b}_1\|_p} = \|A\|_p, \quad \frac{\|A^{-1}\mathbf{b}_2\|_p}{\|\mathbf{b}_2\|_p} = \|A^{-1}\|_p.$$

Show that

$$\lim_{\epsilon \to 0} \frac{\text{Rel}_p(A^{-1}(A\mathbf{b}_1 + \epsilon\mathbf{b}_2), A^{-1}(A\mathbf{b}))}{\text{Rel}_p(A\mathbf{b}_1 + \epsilon\mathbf{b}_2, A\mathbf{b}_1)} = \|A^{-1}\|_p \|A\|_p.$$

Then show that if

$$\text{Rel}_p(A^{-1}\widehat{\mathbf{b}}, A^{-1}\mathbf{b}) \leq C \,\text{Rel}_p(\widehat{\mathbf{b}}, \mathbf{b})$$

for all $\widehat{\mathbf{b}} \neq \mathbf{b}$, then $C \geq \|A^{-1}\|_p \|A\|_p$.

(12) Show that for any $n \times n$ matirx A (over $\mathbb{R}$ or $\mathbb{C}$) and any $p \geq 1$:

(a)

$$\max_{\mathbf{v} \neq \mathbf{0}} \frac{\|A\mathbf{v}\|_p}{\|\mathbf{v}\|_p} = \max_{\|\mathbf{v}\|_p = 1} \|A\mathbf{v}\|_p,$$

and

(b)

$$\max_{\|\mathbf{v}\|_p = 1} \|A\mathbf{v}\|_p = \max_{\|\mathbf{v}\|_p \leq 1} \|A\mathbf{v}\|_p.$$

Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, CANADA.

E-mail address: jf@cs.ubc.ca

URL: http://www.cs.ubc.ca/~jf