Lecture `cond2`: Matrix condition number

October 3, 2023

Summary: Vector and matrix norms. Inputs and outputs for the problem of solving a linear system. Conditioning of linear systems.

References: Section 2.3 of *Numerical Analysis* by T. Sauer.

Norms

Vector norms

How do we measure the "size" of vectors and matrices? When is a vector or matrix "big" or "small"? You are, no doubt, already familiar with the *Euclidean norm* or *2-norm* of a vector,

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2},$$

which is just the Pythagorean notion of length. However, there are many other vector norms. Indeed, the 2-norm is a particular case of the p-norm

$$\|\mathbf{x}\|_p = \big(|x_1|^p + |x_2|^p + \cdots + |x_n|\big)^{1/p},$$

where p is an integer obeying $1 \le p < \infty$. For example, the 1-norm is

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|.$$

The norm we shall mostly focus on is the "infinity norm"

$$\|\mathbf{x}\|_\infty = \max_{1 \le k \le n} |x_k|,$$

so called because, with proper interpretation, we may view it as the $p \to \infty$ limit of the p-norm. For example,

$$\mathbf{x} = \begin{pmatrix} 2 \\ -10.1 \\ 3.4 \\ 7 \end{pmatrix}, \qquad \|\mathbf{x}\|_\infty = 10.1.$$

All norms obey the following conditions.

- $\|\mathbf{x}\| \ge 0$, with $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$ (zero vector)
- $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for any scalar number α
- $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$

The last property is the *triangle inequality.* Here, and often, we just write $\|\cdot\|$ when we refer to a generic norm (it might be the 2-norm, the 8-norm, whatever).

Matrix norms

How do we measure the size of a matrix? We need a *matrix norm.*[1] One way is to use the *Frobenius norm*,

$$\|A\|_F = \sqrt{\sum_{j=1}^{n}\sum_{k=1}^{n}|a_{jk}|^2},$$

which stems from treating the matrix like a vector and using the Euclidean length. While this is sometimes a good norm to use, here we consider *operator norms.* Given a vector norm $\|\cdot\|$ defined on $\mathbb{R}^n$ vectors, define the corresponding or *induced matrix norm* defined on $\mathbb{R}^{n\times n}$ matrices as follows:

$$\|A\| = \max_{\mathbf{x}\neq\mathbf{0}}\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \quad \text{or equivalently} \quad \|A\| = \max_{\|\mathbf{x}\|=1}\|A\mathbf{x}\|. \tag{1}$$

For each vector norm, this defines a matrix norm. For example,

$$\|A\|_2 = \max_{\mathbf{x}\neq\mathbf{0}}\frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \quad \text{or} \quad \|A\|_\infty = \max_{\mathbf{x}\neq\mathbf{0}}\frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}. \tag{2}$$

Be careful here. The $\|\cdot\|$ appearing on the right side of (1) is a vector norm (both $\mathbf{x}$ and $A\mathbf{x}$ are length-n vectors), while the $\|\cdot\|$ appearing on the right side is a matrix norm. Induced matrix norms are useful because they are consistent with the vector norm used to define them. That is,

$$\boxed{\|A\mathbf{w}\| \leq \|A\|\|\mathbf{w}\|}$$

always holds, with the identity involving two vector norms (on $A\mathbf{w}$ and $\mathbf{w}$) and one matrix norm (on A alone). The proof is as follows. The inequality is clearly true if $\mathbf{w} = \mathbf{0}$ is the zero vector, so assume $\mathbf{w} \neq \mathbf{0}$. Then

$$\frac{\|A\mathbf{w}\|}{\|\mathbf{w}\|} \leq \max_{\mathbf{x}\neq\mathbf{0}}\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\|.$$

So $\|A\mathbf{w}\|/\|\mathbf{w}\| \leq \|A\|$ which gives the result.

Infinity matrix norm

We will work with the infinity matrix norm, defined above in (2). The reason is that this norm is also given by the following explicit expression.

$$\boxed{\|A\|_\infty = \max_{1\leq j\leq n}\sum_{k=1}^{n}|a_{jk}|} \tag{3}$$

The infinity norm of a matrix is the *maximum row sum in absolute value.* For example,

$$A = \begin{pmatrix} -1 & -2.1 & 4 & 5 \\ 1.4 & 2.1 & 1.02 & 0.5 \\ 2 & 1 & -3 & 2.2 \\ -0.01 & -0.1 & 0 & 10 \end{pmatrix}, \quad \text{row sums in absolute value} = \begin{cases} 12.1 \\ 5.02 \\ 8.2 \\ 10.11 \end{cases}, \quad \|A\|_\infty = 12.1.$$

Moreover, since it is an induced matrix norm, we also have the aforementioned inequality.

$$\boxed{\|A\mathbf{w}\|_\infty \leq \|A\|_\infty\|\mathbf{w}\|_\infty} \tag{4}$$

[1] As you have learned in Math 314 or Math 321, the set of matrices of a fixed size (here $n \times n$) is a vector space. Therefore, matrix norms are actually vector norms too, but let's not get confused by this. We will call a norm for matrices a *matrix norm*, and reserve the term *vector norm* for norms on vectors in $\mathbb{R}^n$.

Let us prove that the second definition in (2) indeed yields (3). You might skip this if you're OK just using the formula. But if you're curious, read on. First, note that the second definition in (2) is equivalent to

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

Let j_0 be the row index where the maximum row sum occurs. That is, assume

$$\sum_{k=1}^{n} |a_{jk}| \leq \sum_{k=1}^{n} |a_{j_0 k}| \text{ for all } j = 1, 2, \ldots, n.$$

To show that $\|A\|_\infty = \sum_{k=1}^{n} |a_{j_0 k}|$, we will show

$$\|A\|_\infty \leq \sum_{k=1}^{n} |a_{j_0 k}| \quad \text{and} \quad \sum_{k=1}^{n} |a_{j_0 k}| \leq \|A\|_\infty. \tag{5}$$

Let's prove the left-hand inequality in (5) first. All components $|x_k| \leq 1$ for any vector $\mathbf{x}$ which is unit in the infinity norm, $\|\mathbf{x}\|_\infty = 1$. Therefore,

$$\|A\mathbf{x}\|_\infty = \max_{1 \leq j \leq n} \Big| \sum_{k=1}^{n} a_{jk} x_k \Big| \leq \max_{1 \leq j \leq n} \sum_{k=1}^{n} |a_{jk}||x_k| \leq \max_{1 \leq j \leq n} \sum_{k=1}^{n} |a_{jk}|.$$

This implies

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty = 1} \|A\mathbf{x}\|_\infty \leq \sum_{k=1}^{n} |a_{j_0 k}|.$$

To get the right-hand inequality in (5), first write the matrix A as follows:

$$A = \begin{pmatrix} (-1)^{s_{11}}|a_{11}| & (-1)^{s_{12}}|a_{12}| & \cdots & (-1)^{s_{1n}}|a_{1n}| \\ (-1)^{s_{21}}|a_{21}| & (-1)^{s_{22}}|a_{22}| & \cdots & (-1)^{s_{2n}}|a_{2n}| \\ \vdots & \vdots & & \vdots \\ (-1)^{s_{n1}}|a_{n1}| & (-1)^{s_{n2}}|a_{n2}| & \cdots & (-1)^{s_{nn}}|a_{nn}| \end{pmatrix}, \tag{6}$$

where the $s_{jk} = 0$ or 1, and choose $\mathbf{x}_j = ((-1)^{s_{j1}}, (-1)^{s_{j2}}, \ldots, (-1)^{s_{jn}})^T$. Then $\|\mathbf{x}_j\|_\infty = 1$ and

$$A\mathbf{x}_j \text{ has } j\text{th component } |a_{j1}| + |a_{j2}| + \cdots + |a_{jn}|.$$

Therefore, we indeed find

$$\sum_{k=1}^{n} |a_{j_0 k}| = \|A\mathbf{x}_{j_0}\|_\infty \leq \|A\|_\infty \|\mathbf{x}_{j_0}\|_\infty = \|A\|_\infty.$$

Matrix condition number and error magnification

Residual and error magnification

Consider an *invertible* linear system $A\mathbf{z} = \mathbf{b}$, and assume that the "exact solution" is $\mathbf{z} = \mathbf{x}$. That is, $\mathbf{x}$ is the unique vector obeying $A\mathbf{x} = \mathbf{b}$ (solving the system). Suppose we are given an "approximate solution" $\mathbf{x}_A$ to the system.[2] For now, $\mathbf{x}_A$ is just some given vector; we may pick it to be anything we like. Later, it will be what the computer returns when we solve the linear system numerically.

[2] The subscript A on $\mathbf{x}_A$ stands for "approximation" and is only coincidentally the same letter used for the matrix A.

Let us assume that $\mathbf{b} \neq \mathbf{0}$ (zero vector), so that $\mathbf{x} \neq \mathbf{0}$ also. Define the *residual vector*

$$\mathbf{r} = \mathbf{b} - A\mathbf{x}_A, \tag{7}$$

where the word *residual* stems from Latin for "something that's left over". Indeed, if $\mathbf{x}_A$ is close to the solution, then $\mathbf{r}$ should be a small vector, with $\mathbf{r} = \mathbf{0}$ if $\mathbf{x}_A = \mathbf{x}$. The *backward error* is $\|\mathbf{r}\| = \|\mathbf{b} - A\mathbf{x}_A\|$, and

$$\boxed{\text{relative backward error } = \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.}$$

Note that the relative backward error can computed with $\mathbf{x}_A$ *even if we don't know the exact solution* $\mathbf{x}$. It measures *how well* $\mathbf{x}_A$ *solves the linear system.* If we do know the exact solution, we may also compute

$$\boxed{\text{relative forward error } = \frac{\|\mathbf{x}_A - \mathbf{x}\|}{\|\mathbf{x}\|}.}$$

The *error magnification* is defined as

$$\boxed{\text{error magnification } = \frac{\text{relative forward error}}{\text{relative backward error}} = \frac{\|\mathbf{x}_A - \mathbf{x}\|}{\|\mathbf{x}\|}\frac{\|\mathbf{b}\|}{\|\mathbf{r}\|}.} \tag{8}$$

If the error magnification is large, then despite $\mathbf{x}_A$ nearly solving the system, it is far from the exact solution $\mathbf{x}$. As we shall see, this scenario can happen.

The follow examples are from T. Sauer's textbook, page 86. First, consider

$$\begin{pmatrix} 1 & 1 \\ 3 & -4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix},$$

for which the exact solution is clearly $\mathbf{x} = (2,1)^T$. Let us postulate that $\mathbf{x}_A = (1,1)^T$. Then the residual vector is

$$\mathbf{r} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 3 & -4 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \Longrightarrow \|\mathbf{r}\|_\infty = 3,$$

and the relative backward error is $\|\mathbf{r}\|_\infty / \|\mathbf{b}\|_\infty = 3/3 = 1$. However, $\mathbf{x}_A - \mathbf{x} = (-1,0)^T$, so the forward error is $\|\mathbf{x}_A - \mathbf{x}\|_\infty = 1$, and the relative forward error is $1/\|\mathbf{x}\|_\infty = \frac{1}{2}$. Then

$$\text{error magnification} = \tfrac{1}{2}.$$

For this example, the relative forward and backward errors are of about the same magnitude.

Second, consider the system

$$\begin{pmatrix} 1 & 1 \\ 1.0001 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2.0001 \end{pmatrix}, \tag{9}$$

for which the exact solution is clearly $\mathbf{x} = (1,1)^T$. Let us pick $\mathbf{x}_A = (-1, 3.0001)^T$. Now,

$$\mathbf{r} = \begin{pmatrix} 2 \\ 2.0001 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1.0001 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ 3.0001 \end{pmatrix} = \begin{pmatrix} -0.0001 \\ 0.0001 \end{pmatrix} \Longrightarrow \|\mathbf{r}\|_\infty = 0.0001 = 10^{-4},$$

and the relative backward error $= 10^{-4}/\|\mathbf{b}\|_\infty = 10^{-4}/2.0001$. On the other hand, the relative forward error equals the absolute forward error since $\|\mathbf{x}\|_\infty = 1$. So the relative forward error $= \|\mathbf{x}_A - \mathbf{x}\|_\infty = \|(-2, -2.0001)^T\|_\infty = 2.0001$. We have then found

$$\text{error magnification } = \frac{2.0001}{10^{-4}/2.0001} = 40004.0001.$$

This is a somewhat large error magnification. The relative forward error is more than four orders of magnitude (powers of 10) larger than the relative backward error.

Matrix condition number

Can we interpret $\mathbf{x}_A$ as the "exact solution" to a perturbed system $(A + \delta A)\mathbf{z} = \mathbf{b} + \delta\mathbf{b}$? The answer is yes! Here is one possibility, with $\delta A = O$ (zero matrix), but $\delta\mathbf{b} \neq \mathbf{0}$ (zero vector). Recall that for the problem of solving a linear system, the inputs are the set $(A, \mathbf{b})$ and the output is the solution "$\mathbf{x} = \mathbf{h}(A, \mathbf{b})$" which we view as a (vector-valued) function of the inputs. Therefore, our aim is to view

$$\mathbf{x}_A = \mathbf{h}(A + \delta A, \mathbf{b} + \delta\mathbf{b}) = \mathbf{h}(A, \mathbf{b} + \delta\mathbf{b}).$$

Often, when examining the backward stability of specific algorithms for solving a linear system in floating-point arithmetic, it proves necessary to allow for the possibility that $\delta A \neq O$. However, to keep the discussion simple here, we'll keep $\delta A = O$.

First, define $\mathbf{b}_A = A\mathbf{x}_A$. Then clearly $\mathbf{x}_A$ is the exact solution to $A\mathbf{z} = \mathbf{b}_A$. We then write $\mathbf{b}_A = \mathbf{b} + \delta\mathbf{b}$, as is always possible since $\delta\mathbf{b}$ is just defined as $\mathbf{b}_A - \mathbf{b}$. Then $A\mathbf{x}_A = \mathbf{b}_A$ becomes

$$A\mathbf{x}_A = \mathbf{b} + \delta\mathbf{b}.$$

We want to use the last expression to derive the condition number for solving a linear system, i.e. *a measure of how sensitive the solution* $\mathbf{h}(A, \mathbf{b})$ *is to small changes in the input* $\mathbf{b}$. To this end, also write $\mathbf{x}_A = \mathbf{x} + \delta\mathbf{x}$, and express the last equation as

$$A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}.$$

Therefore, upon using $A\mathbf{x} = \mathbf{b}$,

$$A\delta\mathbf{x} = \delta\mathbf{b} \implies \delta\mathbf{x} = A^{-1}\delta\mathbf{b}.$$

The formula on the right relates the change in the solution (output) to the change in the inhomogeneity (input). This is a good start. Again, we will alway work with the infinity norm, but to keep the calculation general, let's assume that $\|\cdot\|$ is any p-norm, both for the vector norm and the induced matrix norm. Taking the vector norm of each side of the last expression, and then using the consistency between the vector and induced matrix norm, we find

$$\|\delta\mathbf{x}\| \leq \|A^{-1}\|\|\delta\mathbf{b}\| \implies \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A^{-1}\|\frac{\|\mathbf{b}\|}{\|\mathbf{x}\|}\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Owing to the fact that $\|\mathbf{b}\| = \|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$, we have shown that

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Define[3] $\kappa(A) = \|A\|\|A^{-1}\|$. Then we have shown

$$\frac{\|\delta\mathbf{x}\|/\|\mathbf{x}\|}{\|\delta\mathbf{b}\|/\|\mathbf{b}\|} \leq \kappa(A) \text{ for all } \delta\mathbf{b} \implies \max_{\delta\mathbf{b}\in\mathbb{R}^n} \frac{\|\delta\mathbf{x}\|/\|\mathbf{x}\|}{\|\delta\mathbf{b}\|/\|\mathbf{b}\|} \leq \kappa(A). \tag{10}$$

This formula tells us that the expression $\kappa(A)$ is an upper bound on the condition number (for the problem of solving a linear system, given a fixed A). In fact, $\kappa(A)$ is equal to the condition number which follows upon showing that the inequality above can be realized as an equality for a certain choice of $\delta\mathbf{b}$. This isn't hard to show, but we will omit the details. Notice that

$$\mathbf{r} \equiv -\delta\mathbf{b} = \mathbf{b} - \mathbf{b}_A = \mathbf{b} - A\mathbf{x}_A$$

is the residual. So the backward error is $\|\mathbf{r}\| = \|\delta\mathbf{b}\|$, and (10) can also be written as

$$\text{error magnification} = \frac{\|\mathbf{x} - \mathbf{x}_A\|/\|\mathbf{x}\|}{\|\mathbf{r}\|/\|\mathbf{b}\|} \leq \kappa(A).$$

[3] The expression depends on the choice of norm. We shall focus on the case $\kappa_\infty(A) = \|A\|_\infty\|A^{-1}\|_\infty$.

That is, $\kappa(A)$ *is the largest possible error magnification* (the inequality can be realized as an equality). Let's at least confirm this for the second example above, where we observed

$$\text{error magnification } = 40004.0001.$$

The coefficient matrix from (9) and its inverse are

$$A = \begin{pmatrix} 1 & 1 \\ 1.0001 & 1 \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} -10000 & 10000 \\ 10001 & -10000 \end{pmatrix},$$

as can be straightforwardly checked. Therefore, using the infinity norm, as in the analysis above for the example,

$$\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = 2.0001 \cdot 200001 = 400004.0001.$$

So for this example, the error magnification is the largest possible value.

A notoriously bad, yet invertible, matrix

The Hilbert matrix H_n has entries

$$h_{ij} = 1/(i+j-1),$$

so, for example,

$$H_2 = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}, \qquad H_3 = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}.$$

As n gets larger, H_n is notoriously ill-conditioned; it has a large condition number. We consider the following experiment. We declare the exact solution to be the length-n vector of all 1's: $\mathbf{x} = (1, 1, \ldots, 1)^T$. Given this exact solution, we define $\mathbf{b} = H_n\mathbf{x}$, and then solve the system $H\mathbf{x} = \mathbf{b}$. We should of course get the vector of all ones back again. In Matlab the experiment looks like this.

```
n = 6           % pick n
H = hilb(n);    % create Hilbert matrix
x = ones(n,1); % the "exact solution"
b = H*x;        % the right-hand side belonging to x.
xA = H\b;       % solve the system using Matlab's backslash
ferr = norm(xA-x,inf)/norm(x,inf); % forward error, note ||x||_inf = 1
berr = norm(b-H*xA)/norm(b,inf);   % backward error
errormag = ferr/berr;              % error magnification
condH = cond(H,inf);               % infinity-norm condition number
```

We tabulate the results of this experiment for different value of n in the following table.

```
-----------------------------------------------------------
|  n  |  f.errors  |  b.errors  |  m.factor  |  cond.num  |
-----------------------------------------------------------
|   6 | 5.4048e-10 | 9.0630e-17 | 5.9636e+06 | 2.9070e+07 |
|   7 | 1.7371e-08 | 8.5637e-17 | 2.0284e+08 | 9.8519e+08 |
|   8 | 5.0750e-07 | 8.1698e-17 | 6.2118e+09 | 3.3873e+10 |
|   9 | 7.1554e-06 | 7.8490e-17 | 9.1164e+10 | 1.0996e+12 |
|  10 | 3.3731e-04 | 7.5810e-17 | 4.4495e+12 | 3.5351e+13 |
|  11 | 6.8060e-03 | 7.3528e-17 | 9.2564e+13 | 1.2279e+15 |
|  12 | 2.9735e-03 | 1.0017e-15 | 2.9684e+12 | 3.8273e+16 |
-----------------------------------------------------------
f.errors, b.errors: relative forward and backward errors
m.factor, cond.num: magnification and inf-norm condition #.
```

Notice that the condition number is always greater than the magnification factor, consistent with its interpretation as the largest possible magnification factor over all right-hand sides $\mathbf{b}$. Nevertheless, for these solves the magnification factor is close to the largest possible one, and becomes very large with increased n. For n large, $\mathbf{x}_A$ is of poor quality, despite yielding a small residual $\mathbf{r} = \mathbf{b} - H\mathbf{x}_A$, and H is ill-conditioned.

Rule of thumb. We can work with any condition number $\kappa(A)$, be it $\kappa_\infty(A)$ or another. If $\kappa(A) = 10^p$, then when solving $A\mathbf{x} = \mathbf{b}$, we expect that the relative backward error $\|\mathbf{r}\|/\|\mathbf{b}\|$ is of size about $\varepsilon_{\text{mach}} \simeq 10^{-16}$. This means that $\|\mathbf{x}_A - \mathbf{x}\|/\|\mathbf{x}\| \lesssim 10^{p-16}$. Quite possibly then, $\|\mathbf{x}_A - \mathbf{x}\|/\|\mathbf{x}\|$ is as large as 10^{p-16}. In a relative sense we should therefore expect to lose p digits of accuracy in the computed solution. Although the relative forward error may be much smaller than 10^{p-16}, in practice this size for the relative forward error is often spot-on.