Notes on Conditioning

Robert A. van de Geijn
The University of Texas
Austin, TX 78712

October 6, 2014

NOTE: I have not thoroughly proof-read these notes!!!

1 Motivation

Correctness in the presence of error (e.g., when floating point computations are performed) takes on a different meaning. For many problems for which computers are used, there is one correct answer, and we expect that answer to be computed by our program. The problem is that, as we will see later, most real numbers cannot be stored exactly in a computer memory. They are stored as approximations, floating point numbers, instead. Hence storing them and/or computing with them inherently incurs error.

Naively, we would like to be able to define a program that computes with floating point numbers as being "correct" if it computes an answer that is close to the exact answer. Unfortunately, some problems that are computed this way have the property that a small change in the input yields a large change in the output. Surely we can't blame the program for not computing an answer close to the exact answer in this case. The mere act of storing the input data as a floating point number may cause a completely different output, even if all computation is exact. We will later define *stability* to be a property of a program. It is what takes the place of correctness. In this note, we instead will focus on when a problem is a "good" problem, meaning that in exact arithmetic a "small" change in the input will always cause at most a "small" change in the output, or a "bad" problem if a "small" change may yield a "large" A good problems will be called *well-conditioned*. A bad problem will be called *ill-conditioned*.

Notice that "small" and "large" are vague. To some degree, norms help us measure size. To some degree, "small" and "large" will be in the eyes of the beholder (in other words, situation dependent).

2 Notation

Throughout this note, we will talk about small changes (perturbations) to scalars, vectors, and matrices. To denote these, we attach a "delta" to the symbol for a scalar, vector, or matrix.

- A small change to scalar $\alpha \in \mathbb{C}$ will be denoted by $\delta\alpha \in \mathbb{C}$;
- A small change to vector $x \in \mathbb{C}^n$ will be denoted by $\delta x \in \mathbb{C}^n$; and
- A small change to matrix $A \in \mathbb{C}^{m\times n}$ will be denoted by $\Delta A \in \mathbb{C}^{m\times n}$.

Notice that the "delta" touches the α, x, and A, so that, for example, δx is not mistaken for $\delta \cdot x$.

3 The Prototypical Example: Solving a Linear System

Assume that $A \in \mathbb{R}^{n \times n}$ is nonsingular and $x, y \in \mathbb{R}^n$ with $Ax = y$. The problem here is the function that computes x from y and A. Let us assume that no error is introduced in the matrix A when it is stored, but that in the process of storing y a small error is introduced: $\delta y \in \mathbb{R}^n$ so that now $y + \delta y$ is stored. The question becomes by how much the solution x changes as a function of δy. In particular, we would like to quantify how a relative change in the right-hand side y ($\|\delta y\|/\|y\|$ in some norm) translates to a relative change in the solution x ($\|\delta x\|/\|x\|$). It turns out that we will need to compute norms of matrices, using the norm induced by the vector norm that we use.

Since $Ax = y$, if we use a consistent (induced) matrix norm,

$$\|y\| = \|Ax\| \leq \|A\|\|x\| \text{ or, equivalently, } \frac{1}{\|x\|} \leq \|A\| \frac{1}{\|y\|}. \tag{1}$$

Also,

$$\left.\begin{array}{rcl} A(x + \delta x) & = & y + \delta y \\ Ax & = & y \end{array}\right\} \text{ implies that } A\delta x = \delta y \text{ so that } \delta x = A^{-1}\delta y.$$

Hence

$$\|\delta x\| = \|A^{-1}\delta y\| \leq \|A^{-1}\|\|\delta y\|. \tag{2}$$

Combining (1) and (2) we conclude that

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

What does this mean? It means that the relative error in the solution is at worst the relative error in the right-hand side, amplified by $\|A\|\|A^{-1}\|$. So, if that quantity is "small" *and* the relative error in the right-hand size is "small" *and* exact arithmetic is used, then one is guaranteed a solution with a relatively "small" error.

The quantity $\kappa_{\|\cdot\|}(A) = \|A\|\|A^{-1}\|$ is called the *condition number* of nonsingular matrix A (associated with norm $\|\cdot\|$).

Are we overestimating by how much the relative error can be amplified? The answer to this is **no**. For every nonsingular matrix A, there exists a right-hand side y and perturbation δy such that, if $A(x + \delta x) = y + \delta y$,

$$\frac{\|\delta x\|}{\|x\|} = \|A\|\|A^{-1}\| \frac{\|\delta y\|}{\|y\|}.$$

In order for this equality to hold, we need to find y and δy such that

$$\|y\| = \|Ax\| = \|A\|\|x\| \text{ or, equivalently, } \|A\| = \frac{\|Ax\|}{\|x\|}$$

and

$$\|\delta x\| = \|A^{-1}\delta y\| = \|A^{-1}\|\|\delta y\|. \text{ or, equivalently, } \|A^{-1}\| = \frac{\|A^{-1}\delta y\|}{\|\delta y\|}.$$

In other words, x can be chosen as a vector that maximizes $\|Ax\|/\|x\|$ and δy should maximize $\|A^{-1}\delta y\|/\|\delta y\|$. The vector y is then chosen as $y = Ax$.

What if we use the 2-norm? For this norm, $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \sigma_0/\sigma_{n-1}$. So, the ratio between the largest and smallest singular value determines whether a matrix is well-conditioned or ill-conditioned.

To show for what vectors the maximal magnification is attained, consider the SVD

$$A = U\Sigma V^T = \left(\begin{array}{c|c|c|c} u_0 & u_1 & \cdots & u_{n-1} \end{array} \right) \begin{pmatrix} \sigma_0 & & & \\ & \sigma_1 & & \\ & & \ddots & \\ & & & \sigma_{n-1} \end{pmatrix} \left(\begin{array}{c|c|c|c} v_0 & v_1 & \cdots & v_{n-1} \end{array} \right)^H.$$

Recall that

- $\|A\|_2 = \sigma_0$, v_0 is the vector that maximizes $\max_{\|z\|_2=1} \|Az\|_2$, and $Av_0 = \sigma_0 u_0$;
- $\|A^{-1}\|_2 = 1/\sigma_{n-1}$, u_{n-1} is the vector that maximizes $\max_{\|z\|_2=1} \|A^{-1}z\|_2$, and $Av_{n-1} = \sigma_{n-1}u_{n-1}$.

Now, take $y = \sigma_0 u_0$. Then $Ax = y$ is solved by $x = v_0$. Take $\delta y = \beta\sigma_1 u_1$. Then $A\delta x = \delta y$ is solved by $x = \beta v_1$. Now,

$$\frac{\|\delta y\|_2}{\|y\|_2} = \frac{|\beta|\sigma_1}{\sigma_0} \text{ and } \frac{\|\delta x\|_2}{\|x\|_2} = |\beta|.$$

Hence

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\sigma_0}{\sigma_{n-1}} \frac{\|\delta y\|_2}{\|y\|_2}$$

This is depicted in Figure 1 for $n = 2$.

The SVD can be used to show that A maps the unit ball to an ellipsoid. The singular values are the lengths of the various axes of the ellipsoid. The condition number thus captures the eccentricity of the ellipsoid: the ratio between the lengths of the largest and smallest axes. This is also illustrated in Figure 1.

Number of accurate digits Notice that for scalars $\delta\psi$ and ψ, $\log_{10}(\frac{\delta\psi}{\psi}) = \log_{10}(\delta\psi) - \log_{10}(\psi)$ roughly equals the number leading decimal digits of $\psi + \delta\psi$ that are accurate, relative to ψ. For example, if $\psi = 32.512$ and $\delta\psi = 0.02$, then $\psi + \delta\psi = 32.532$ which has three accurate digits (highlighted in read). Now, $\log_{10}(32.512) - \log_{10}(0.02) \approx 1.5 - (-1.7) = 3.2$.

Now, *if*

$$\frac{\|\delta x\|}{\|x\|} = \kappa(A) \frac{\|\delta y\|}{\|y\|}.$$

then

$$\log_{10}(\|\delta x\|) - \log_{10}(\|x\|) = \log_{10}(\kappa(A)) + \log_{10}(\|\delta y\|) - \log_{10}(\|y\|)$$

so that

$$\log_{10}(\|x\|) - \log_{10}(\|\delta x\|) = [\log_{10}(\|y\|) - \log_{10}(\|\delta y\|)] - \log_{10}(\kappa(A)).$$

In other words, if there were k digits of accuracy in the right-hand side, then it is possible that (due only to the condition number of A) there are only $k - \log_{10}(\kappa(A))$ digits of accuracy in the solution. If we start with only 8 digits of accuracy and $\kappa(A) = 10^5$, we may only get 3 digits of accuracy. If $\kappa(A) \geq 10^8$, we may not get *any* digits of accuracy...

Exercise 1. *Show that, for a consistent matrix norm, $\kappa(A) \geq 1$.*

We conclude from this that we can generally only expect as much relative accuracy in the solution as we had in the right-hand side.

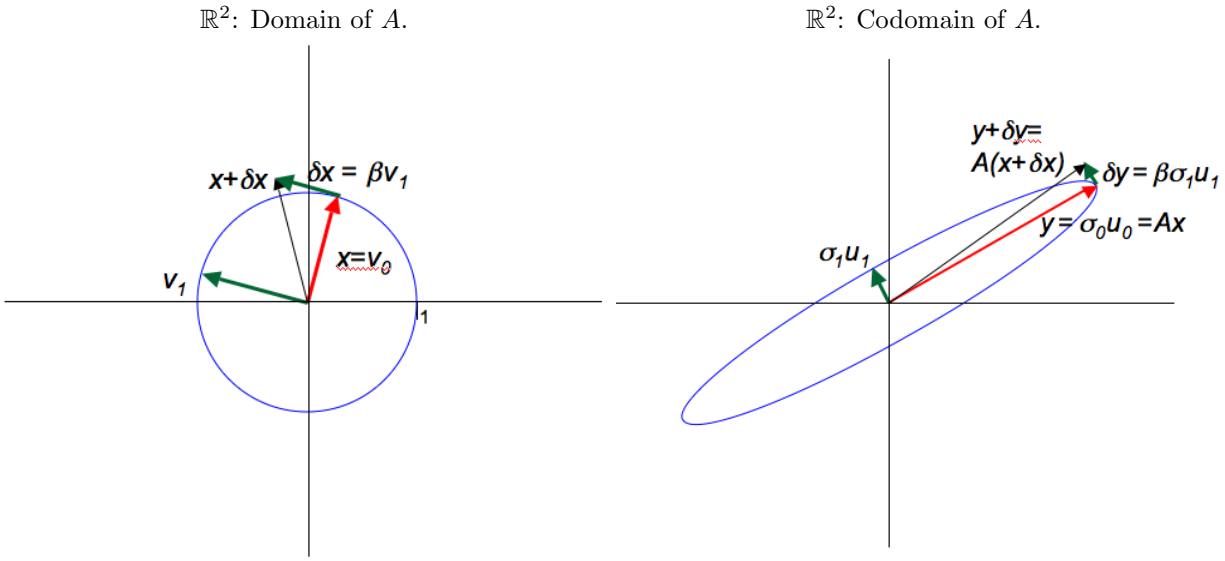

Figure 1: Illustration for choices of vectors y and δy that result in $\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\sigma_0}{\sigma_{n-1}} \frac{\|\delta y\|_2}{\|y\|_2}$. Because of the eccentricity of the ellipse, the relatively small change δy relative to y is amplified into a relatively large change δx relative to x. On the right, we see that $\|\delta y\|_2/\|y\|_2 = \beta\sigma_1/\sigma_0$ (since $\|u_0\|_2 = \|u_1\|_2 = 1$). On the left, we see that $\|\delta x\|_2/\|x\| = \beta$ (since $\|v_0\|_2 = \|v_1\|_2 = 1$).

Alternative exposition Note: the below links conditioning of matrices to the relative condition number of a more general function. For a more thorough treatment, you may want to read Lecture 12 of "Trefethen and Bau". That book discusses the subject in much more generality than is needed for our discussion of linear algebra. Thus, if this alternative exposition baffles you, just skip it!

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a continuous function such that $f(y) = x$. Let $\|\cdot\|$ be a vector norm. Consider for $y \neq 0$

$$\kappa^f(y) = \lim_{\delta \to 0} \sup_{\|\delta y\| \leq \delta} \left(\frac{\|f(y + \delta y) - f(y)\|}{\|f(y)\|} \right) / \left(\frac{\|\delta y\|}{\|y\|} \right).$$

Letting $f(y + \delta y) = x + \delta x$, we find that

$$\kappa^f(y) = \lim_{\delta \to 0} \sup_{\|\delta y\| \leq \delta} \left(\frac{\|x + \delta x\|}{\|x\|} \right) / \left(\frac{\|\delta y\|}{\|y\|} \right).$$

(Obviously, if $\delta y = 0$ or $y = 0$ or $f(y) = 0$, things get a bit hairy, so let's not allow that.)

Roughly speaking, $\kappa^f(y)$ equals the maximum that a(n infitessimally) small relative error in y is magnified into a relative error in $f(y)$. This can be considered the **relative condition number** of function f. A large relative condition number means a small relative error in the input (y) can be magnified into a large relative error in the output ($x = f(y)$). This is bad, since small errors will invariable occur.

Now, if $f(y) = x$ is the function that returns x where $Ax = y$ for a nonsingular matrix $A \in \mathbb{C}^{n \times n}$, then via an argument similar to what we did earlier in this section we find that $\kappa^f(y) \leq \kappa(A) = \|A\|\|A^{-1}\|$, the condition number of matrix A:

$$\lim_{\delta \to 0} \sup_{\|\delta y\| \leq \delta} \left(\frac{\|f(y + \delta y) - f(y)\|}{\|f(y)\|} \right) / \left(\frac{\|\delta y\|}{\|y\|} \right)$$

$$
\begin{aligned}
&= \lim_{\delta \to 0} \sup_{\|\delta y\| \le \delta} \left(\frac{\|A^{-1}(y+\delta y) - A^{-1}(y)\|}{\|A^{-1}y\|} \right) / \left(\frac{\|\delta y\|}{\|y\|} \right) \\
&= \lim_{\delta \to 0} \max_{\substack{\|z\|=1 \\ \delta y = \delta \cdot z}} \left(\frac{\|A^{-1}(y+\delta y) - A^{-1}(y)\|}{\|A^{-1}y\|} \right) / \left(\frac{\|\delta y\|}{\|y\|} \right) \\
&= \lim_{\delta \to 0} \max_{\substack{\|z\|=1 \\ \delta y = \delta \cdot z}} \left(\frac{\|A^{-1}\delta y\|}{\|\delta y\|} \right) / \left(\frac{\|A^{-1}y\|}{\|y\|} \right) \\
&= \lim_{\delta \to 0} \max_{\substack{\|z\|=1 \\ \delta y = \delta \cdot z}} \left(\frac{\|A^{-1}\delta y\|}{\|\delta y\|} \right) \left(\frac{\|y\|}{\|A^{-1}y\|} \right) \\
&= \left[\lim_{\delta \to 0} \max_{\substack{\|z\|=1 \\ \delta y = \delta \cdot z}} \left(\frac{\|A^{-1}\delta y\|}{\|\delta y\|} \right) \right] \left[\left(\frac{\|y\|}{\|A^{-1}y\|} \right) \right] \\
&= \lim_{\delta \to 0} \max_{\substack{\|z\|=1 \\ \delta y = \delta \cdot z}} \left(\frac{\|A^{-1}(\delta \cdot z)\|}{\|\delta \cdot z\|} \right) \left(\frac{\|y\|}{\|A^{-1}y\|} \right) \\
&= \max_{\|z\|=1} \left(\frac{\|A^{-1}z\|}{\|z\|} \right) \left(\frac{\|y\|}{\|A^{-1}y\|} \right) \\
&= \max_{\|z\|=1} \left(\frac{\|A^{-1}z\|}{\|z\|} \right) \left(\frac{\|Ax\|}{\|x\|} \right) \\
&\le \left[\max_{\|z\|=1} \left(\frac{\|A^{-1}z\|}{\|z\|} \right) \right] \left[\max_{x \ne 0} \left(\frac{\|Ax\|}{\|x\|} \right) \right] \\
&= \|A\| \|A^{-1}\|,
\end{aligned}
$$

where $\|\cdot\|$ is the matrix norm induced by vector norm $\|\cdot\|$.

4 Condition Number of a Rectangular Matrix

Given $A \in \mathbb{C}^{m \times n}$ with linearly independent columns and $y \in \mathbb{C}^m$, consider the linear least-squares (LLS) problem

$$\|Ax - y\|_2 = \min_w \|Aw - y\|_2 \tag{3}$$

and the perturbed problem

$$\|A(x + \delta x) - y\|_2 = \min_{w+\delta w} \|A(w + \delta w) - (y + \delta y)\|_2. \tag{4}$$

We will again bound by how much the relative error in y is amplified.

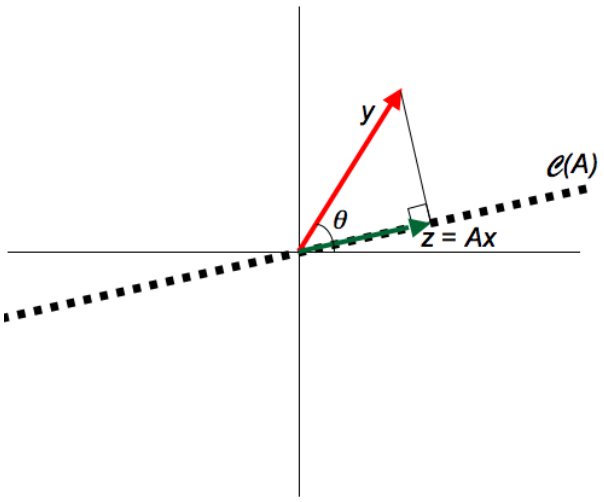

Figure 2: Linear least-squares problem $\|Ax - y\|_2 = \min_v \|Av - y\|_2$. Vector z is the projection of y onto $\mathcal{C}(A)$.

Notice that the solutions to (3) and (4) respectively satisfy

$$\begin{aligned} A^H A x &= A^H y \\ A^H A(x + \delta x) &= A^H (y + \delta y) \end{aligned}$$

so that $A^H A \delta x = A^H \delta y$ (subtracting the first equation from the second) and hence

$$\|\delta x\|_2 = \|(A^H A)^{-1} A^H \delta y\|_2 \leq \|(A^H A)^{-1} A^H\|_2 \|\delta y\|_2.$$

Now, let $z = A(A^H A)^{-1} A^H y$ be the projection of y onto $\mathcal{C}(A)$ and let θ be the angle between z and y. Let us assume that y is not orthogonal to $\mathcal{C}(A)$ so that $z \neq 0$. Then $\cos\theta = \|z\|_2 / \|y\|_2$ so that

$$\cos\theta \|y\|_2 = \|z\|_2 = \|Ax\|_2 \leq \|A\|_2 \|x\|_2$$

and hence

$$\frac{1}{\|x\|_2} \leq \frac{\|A\|_2}{\cos\theta \|y\|_2}$$

We conclude that

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\|A\|_2 \|(A^H A)^{-1} A^H\|_2}{\cos\theta} \frac{\|\delta y\|_2}{\|y\|_2} = \frac{1}{\cos\theta} \frac{\sigma_0}{\sigma_{n-1}} \frac{\|\delta y\|_2}{\|y\|_2}$$

where σ_0 and σ_{n-1} are (respectively) the largest and smallest singular values of A, because of the following result:

Exercise 2. *If A has linearly independent columns, show that $\|(A^H A)^{-1} A^H\|_2 = 1/\sigma_{n-1}$, where σ_{n-1} equals the smallest singular value of A. Hint: Use the SVD of A.*

The condition number of $A \in \mathbb{C}^{m \times n}$ with linearly independent columns is $\kappa_2(A) = \sigma_0 / \sigma_{n-1}$.

Notice the effect of the $\cos\theta$. When y is almost perpendicular to $\mathcal{C}(A)$, then its projection z is small and $\cos\theta$ is small. Hence a small relative change in y can be greatly amplified. This makes sense: if y is almost perpendical to $\mathcal{C}(A)$, then $x \approx 0$, and any small $\delta y \in \mathcal{C}(A)$ can yield a *relatively* large change δx.

5 Why Using the Method of Normal Equations Could be Bad

Exercise 3. *Let A have linearly independent columns. Show that $\kappa_2(A^H A) = \kappa_2(A)^2$.*

Exercise 4. *Let $A \in \mathbb{C}^{n\times n}$ have linearly independent columns.*

- *Show that $Ax = y$ if and only if $A^H Ax = A^H y$.*
- *Reason that if the method of normal equations is used to solve $Ax = y$, then the condition number of the matrix is unnecessarily squared.*

Let $A \in \mathbb{C}^{m\times n}$ have linearly independent columns. If one uses the Method of Normal Equations to solve the linear least-squares problem $\min_x \|Ax-y\|_2$, one ends up solving the square linear system $A^H Ax = A^H y$. Now, $\kappa_2(A^H A) = \kappa_2(A)^2$. Hence, using this method squares the condition number of the matrix being used.

6 Why Multiplication with Unitary Matrices is a Good Thing

Next, consider the computation $C = AB$ where $A \in \mathbb{C}^{m\times m}$ is nonsingular and $B, \Delta B, C, \Delta C \in \mathbb{C}^{m\times n}$. Then

$$\begin{array}{rcl} (C + \Delta C) & = & A(B + \Delta B) \\ C & = & AB \\ \Delta C & = & A\Delta B \end{array}$$

Thus,

$$\|\Delta C\|_2 = \|A\Delta B\|_2 \leq \|A\|_2 \|\Delta B\|_2.$$

Also, $B = A^{-1}C$ so that

$$\|B\|_2 = \|A^{-1}C\|_2 \leq \|A^{-1}\|_2 \|C\|_2$$

and hence

$$\frac{1}{\|C\|_2} \leq \|A^{-1}\|_2 \frac{1}{\|B\|_2}.$$

Thus,

$$\frac{\|\Delta C\|_2}{\|C\|_2} \leq \|A\|_2 \|A^{-1}\|_2 \frac{\|\Delta B\|_2}{\|B\|_2} = \kappa_2(A) \frac{\|\Delta B\|_2}{\|B\|_2}.$$

This means that the relative error in matrix $C = AB$ is at most $\kappa_2(A)$ greater than the relative error in B.

The following exercise gives us a hint as to why algorithms that cast computation in terms of multiplication by unitary matrices avoid the buildup of error:

Exercise 5. *Let $U \in \mathbb{C}^{n\times n}$ be unitary. Show that $\kappa_2(U) = 1$.*

This means is that the relative error in matrix $C = UB$ is no greater than the relative error in B when U is unitary.