Numerical Mathematics 1: Numerical Analysis

Dr. Zahra Lakdawala[1]

[1] ***Associate Professor of Mathematics, LUMS University, zahra.lakdawala@lums.edu.pk***

These notes have been compiled from various sources.

1 Wellposeness, conditioning and stability

We discuss three important concepts in numerical analysis which are related but important to keep apart. While well-posedness and conditioning refer to the mathematical problem one wants to tackle, stability is a property of the algorithm one uses to solve the mathematical problem. For example, let us formalize polynomial root-finding. So the problem is: Given a polynomial $p(x) = \sum_{i=0}^{n} a_i x_i$ (which can be entirely described by its coefficients), find its roots. The map F given by

$$F : \text{coefficient vector } (a_n, \ldots, a_0)^T \mapsto \text{ vector of all real roots}$$

For example,

$$F : (1, 2)^T \mapsto -2$$

and $F : (1, 0, 1)^T \mapsto \{\}$. Note that the map does not always yield a solution, since we look for real roots. In the above example, the function F maps from $\mathbb{R}^{n+1}$ to $\mathbb{R}^k$, with $k \leq n$, depending on the number and algebraic mutilplicity of (real) roots.

In general, a mathematical problem is not restricted to either Euclidean spaces or even finite-dimensional ones. For the input and output spaces we are free to choose any arbitrary normed (function) space such as the space of polynomials of degree not bigger than n as well as the space of continuous or integrable functions. So in general, we study a mathematical problem

$F : V \mapsto W$ for some normed real vector spaces V and W . We call V the input data space and W the output data space. The corresponding norms are denotes with $||\cdot||_V$ and $||\cdot||_W$. Thus, $F(x)$ denotes the output from the mathematical problem F with input x.

1.1 Wellposedness

Hadamard postulated three properties which a mathematical problem should satisfy in order to accurately describe our physical reality. He called a problem well-posed

- if there exists a solution to it,
- the solution is unique
- and depends continuously on the input data.

In this sense looking for the roots of $x+2$ is a well-posed problem whereas looking for the (real) roots of $x^2 + 1$ is not. If the mathematical problem cannot be solved, it makes no sense to devise a root-finding algorithm.The best algorithm in the world is doomed to fail if the problem is ill-posed.

The final property means that if we have two input data sets which are close then also the mathematical problems should produce similar solutions. Notice, however, that the word close in this context is a bit vague. Despite its continuous dependence, it can still be true that varying the input a little bit, results in a relative large change in the output. The condition number helps to quantify exactly how large that change is. There are two common types of condition numbers: The normwise and componentwise condition numbers.

1.2 Condition numbers

1.2.1 Normwise condition number

For some norm given norms $||\cdot||_V$ and $||\cdot||_W$ on the linear spaces V and W which have to be specified for each application the absolute normwise condition number $\kappa_a \geq 0$ is defined as the smallest number such that

$$||F(\tilde{x} - F(x)||_W \leq \kappa_a ||\tilde{x} - x||_V \tag{1}$$

in the limit of $\tilde{x}$ approaching x. Note, that if the mathematical problem does not depend continuously on the data (i.e. it is ill-posed) formally the condition number becomes infinite, $\kappa_a = \infty$. Similarly, one can define the relative normwise condition number $\kappa_r \geq 0$ to be the smallest number such that

$$\frac{||F(\tilde{x} - F(x)||_W}{||F(x)||_W} \leq \kappa_r \frac{||\tilde{x} - x||_V}{||x||_V} \tag{2}$$

in the limit of $\tilde{x}$ approaching x. Both κ_a and κ_r measure how small perturbations to the input data (e.g. the polynomial coefficients) impact the solution to the mathematical problem (the roots). That is,

$$\frac{||\text{change in output}||_W}{||\text{output}||_W} \leq \kappa_r \frac{||\text{change in input}||_V}{||\text{input}||_V}.$$

In other words the condition numbers describe how sensitive the output is with respect to the input. If the function F is differentiable, its derivative help to measure this sensitivity. In fact, we will see below that we can think of the condition numbers as some generalized derivative. For "large" condition numbers the mathematical problem is called ill-conditioned. On the other hand, for condition numbers close to zero or one, the problem is called well-conditioned.

We can obtain equivalent definitions for Equations 1 and 2. By setting $\Delta x = \tilde{x} - x$ we have

$$\kappa_a = \lim_{\epsilon \to 0^+} sup_{\|\Delta x\|_V \leq \epsilon} \frac{\|F(x + \Delta x) - F(x)\|_W}{||\Delta x||_V} \tag{3}$$

and

$$\kappa_r = \lim_{\epsilon \to 0^+} sup_{\|\Delta x\|_V \leq \epsilon} \frac{\|F(x + \Delta x) - F(x)\|_W}{||F(x)||_W} / \frac{\|\Delta x\|_V}{\|x\|_V} \tag{4}$$

From these definitions it become clear that for differentiable F , we obtain

$$\kappa_a = \|DF(x)\|_{V,W} \text{ and } \kappa_r = \frac{\|DF(x)\|_{V,W}}{\|F(x)\|_W \,/\, \|x\|_V},$$

where $DF(x) = \frac{\partial F_i(x)}{\partial x_j}$ denotes the Jacobian of F at x and

$$\|DF(x)\|_{V,W} = sup_{\Delta x \neq 0} \left\{ \frac{\|DF(x)\Delta x\|_W}{\|\Delta x\|_V} \right\}$$

the induced operator norm.

This looks rather cumbersome. What happens if we simplify the problem and study some (differentiable) function $f : \mathbb{R} \mapsto \mathbb{R}$, that is $V = W = \mathbb{R}$? This case corresponds to studying the conditioning of evaluating the function f at x. For real numbers the standard norm is the absolute value $|\cdot|$. In this case, the Jacobian becomes simply a derivative and the operator norm is its absolute value,

$$|f'(x)|_{\mathbb{R},\mathbb{R}} = sup_{\Delta x \neq 0} \left\{ \frac{|f'(x)\Delta x|}{|\Delta x|} \right\} = |f'(x)|.$$

Then the absolute and relative condition numbers simplify to

$$\kappa_a = |f'(x)| \text{ and } \kappa_r = \left| \frac{f'(x)x}{f(x)} \right|.$$

The above expressions one can also extend to functions defined on finite domains.

We finish with another example. Polynomial root finding can become highly ill-conditioned -even ill-posed. Consider the polynomial $p_\epsilon(x) = x^2 - \epsilon$ and its perturbation $p(x) = x^2$ with roots $\pm\sqrt{\epsilon}$ and double root at $x = 0$, respectively. Then the absolute condition number κ_a according to (2.1) would need to satisfy in the discrete maximum norm ($V = \mathbb{R}^3$ and $W = \mathbb{R}^2$), for example

$$\sqrt{\epsilon} = \left\| \begin{bmatrix} \sqrt{\epsilon} \\ -\sqrt{\epsilon} \end{bmatrix} \right\|_\infty = \left\| F \begin{bmatrix} 1 \\ 0 \\ -\epsilon \end{bmatrix} - F \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\|_\infty \leq K_a \left\| \begin{bmatrix} 1 \\ 0 \\ -\epsilon \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\|_\infty = \kappa_a \epsilon.$$

For small ϵ this is impossible since $\sqrt{\epsilon}$ dominates ϵ. Hence, $\kappa_a = \infty$.

1.2.2 The condition number of a linear operator

The previously discussed condition numbers have a special form if we make the additional assumption that the mapping (the operator) $F : V \mapsto W$ is linear, that is for all $x, y \in V$ and $\alpha, \beta \in \mathbb{R}$ we have

$$F(\alpha x + \beta y) = \alpha F(x) + \beta F(y)$$

If we consider the vector space of all linear operators, we can supply it with the induced operator norm

$$\|F\|_{V,W} := sup_{x \neq 0} \frac{\|F(x)\|_W}{\|x\|_V} = sup_{\|x\|_V = 1} \|F(x)\|_W \, .$$

We say the linear operator is bounded if $\|F\|_{V,W} < \infty$. For bounded linear operators the operator norm becomes actually a norm as one can easily verify. From the definition we deduce the important property

$$\|F(x)\|_W \leq \|F\|_{V,W} \|x\|_V$$

for all $x \in V$. From the definition it is clear that $\|F\|_{V,W}$ is the smallest constant for which the above inequality holds. Due to linearity we deduce

$$\|F(\tilde{x}) - F(x)\|_W = \|F(\tilde{x} - x)\|_W \leq \|F\|_{V,W} \|\tilde{x} - x\|_V$$

for all $x, \tilde{x} \in V$.

Theorem: Let $F : V \mapsto W$ be a linear operator. Then

$$\kappa_a = \|F\|_{V,W} \in [0, \infty]$$

and

$$\kappa_r \leq \frac{\|F\|_{V,W}}{inf_{\|x\|_V=1} \|F(x)\|_W} \in [0, \infty].$$

If F is bijective, we have

$$\kappa_r \leq \|F\|_{V,W} \left\|F^{-1}\right\|_{W,V}$$

We point out that if the linear operator F is bounded, its absolute condition number is finite. If it is additionally injective, also its relative condition number is finite. In other words, for linear operators the operator norm is equivalent to the absolute condition number. In fact, the following theorem holds, where we allow the condition numbers to become infinite (indicating an ill-posed problem.

A word on notation: Matrices are a class of very well known linear operators. Since it is somewhat unusual to write $A(x)$ for Ax, for linear operators F we will often drop the parentheses and write Fx instead of $F(x)$.

1.2.3 Componentwise condition number

The previously introduced of normwise condition numbers is sometimes too restrictive. Consider for example the linear system $Ax = b$ with the diagonal matrix

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix} \text{ which implies } A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\epsilon} \end{bmatrix}$$

for $\epsilon > 0$. Intuitively, solving this linear system with a diagonal matrix should be well-conditioned since the linear system decouples into two equations which can be solved separately. However, using the normwise condition number we have in the discrete maximum norm

$$\kappa_r(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty = \frac{1}{\epsilon}$$

So for small ϵ this type of condition number grows!

If we assume for simplicity that $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, it is instructive to study the more sensitive componentwise error

$$F_i(\tilde{x}) - F_i(x) = F_i(x + \Delta x) - F_i(x)$$

for $1 \leq i \leq m$. If we assume that F is differentiable and apply the mean-value theorem to the scalar function $g(t) = F(x + t\Delta x)$, then we may estimate for some $\tau \in [0, 1]$

$$\begin{aligned}
\left|\frac{F_i(x+\Delta x) - F_i(x)}{F_i(x)}\right| &= \frac{1}{|F_i(x)|}\left|\sum_{j=1}^{n} \frac{\partial F_i(x+\tau\Delta x)}{\partial x_j}\Delta x_j\right| \\
&= \frac{1}{|F_i(x)|}\left|\sum_{j=1}^{n} \frac{\partial F_i(x+\tau\Delta x)}{\partial x_j} x_j \frac{\Delta x_j}{x_j}\right| \\
&\leq \frac{1}{|F_i(x)|}\left(\sum_{j=1}^{n} \left|\frac{\partial F_i(x+\tau\Delta x)}{\partial x_j}\right| |x_j|\right) max_{1\leq j\leq n}\left|\frac{\Delta x_j}{x_j}\right| \\
&= \frac{|\nabla F_i(x+\tau\Delta x)|^T |x|}{|F_i(x)|} max_{1\leq j\leq n}\left|\frac{\Delta x_j}{x_j}\right|.
\end{aligned}$$

In the last expression the absolute value sign is applied componentwise to the gradient and the vector x. Similar to before, we rearrange this to

$$\left|\frac{F_i(x+\Delta x) - F_i(x)}{F_i(x)}\right| / max_{1\leq j\leq n}\left|\frac{\Delta x_j}{x_j}\right| \leq \frac{|\nabla F_i(x+\tau\Delta x)|^T |x|}{|F_i(x)|}$$

where again the absolute value sign is applied componentwise to the Jacobian and the vector x. If we understand the division componentwise, the right-hand side becomes

$$\left\|\frac{|DF(x+\tau\Delta x)||x|}{|F(x)|}\right\|_\infty$$

For vanishing Δx this motivates our definition of the componentwise condition number

$$\kappa_r^c = \left\| \frac{|DF(x)||x|}{|F(x)|} \right\|_\infty .$$

Let us compute the componentwise condition number for the linear system in the introductory example. In this case

$$\kappa_r^c = \left\| \frac{\left|\begin{bmatrix}1 & 0\\ 0 & \epsilon\end{bmatrix}\right| \left|\begin{bmatrix}x_1\\ x_2\end{bmatrix}\right|}{\left|\begin{bmatrix}1 & 0\\ 0 & \epsilon\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}\right|} \right\|_\infty = \left\| \frac{\begin{bmatrix}|x_1|\\ \epsilon|x_2|\end{bmatrix}}{\left|\begin{bmatrix}x_1\\ x_2\end{bmatrix}\right|} \right\|_\infty = max\left\{\frac{|x_1|}{|x_1|}, \frac{|x_2|}{|x_2|}\right\} = 1$$

So the componentwise condition number is actually ideal, reflecting our initial intuition.

Let us look at another example. Consider the multiplication of two real numbers, defined by

$$f : \mathbb{R}^2 \mapsto \mathbb{R},\ f(x,y) = xy.$$

The Jacobian is given by

$$Df(x,y) = \begin{bmatrix} y & x \end{bmatrix}.$$

Hence, the normwise condition number in the discrete maximum norm yields

$$\begin{aligned}\kappa_r = \frac{\|Df(x,y)\|_\infty \left\|\begin{bmatrix}x\\ y\end{bmatrix}\right\|_\infty}{|f(x,y)|} &= \frac{(|x|+|y|)max\{|x|,|y|\}}{|xy|}\\ &= \begin{cases} \frac{|x|^2+|x||y|}{|xy|} = \frac{|x|}{|y|}+1, & \text{for } |x|>|y| \\ \frac{|y|^2+|x||y|}{|xy|} = \frac{|y|}{|x|}+1, & \text{for } |x|\le|y| \end{cases}\end{aligned}$$

This means that the normwise condition number implies that multiplication is only well defined if $|x| \approx |y|$. However, if the absolute values of both factors differ by a lot, then the normwise condition number becomes large. On the other hand, the componentwise condition number is given by

$$\kappa_r^c = \left\| |Df(x)| \left|\begin{bmatrix}x\\ y\end{bmatrix}\right| / |f(x,y)| \right\|_\infty = \frac{2|x||y|}{|xy|} = 2$$

implying that in the componentwise sense multiplication is well-conditioned regardless of the magnitude of x and y.

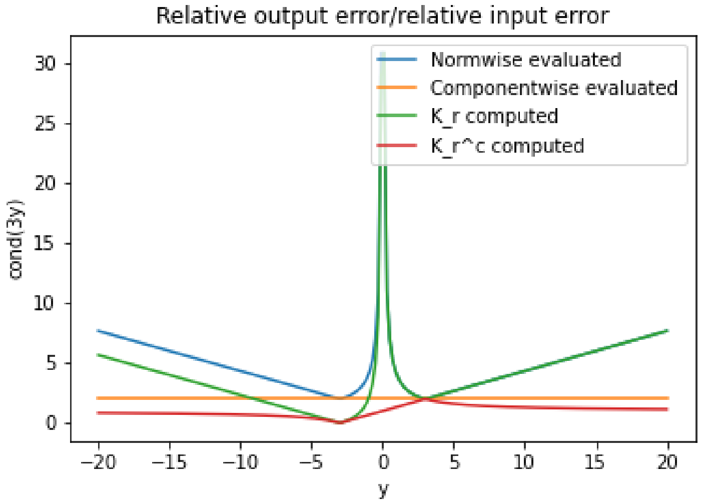

Figure 1: Comparison between relative output error over relative input error for multiplication xy with $x = 3$ and $\epsilon_1 = \epsilon_2 = 10^{-8}$, using normwise (green) and componentwise (orange) errors

So which condition number mirrors the "correct" behavior? Suppose we perturb some $x, y \in \mathbb{R}$ of possibly quite different magnitudes by $\tilde{x} = x(1+\epsilon_1)$ and $\tilde{y} = y(1+\epsilon_2)$ for some $\epsilon_1, \epsilon_2 \in \mathbb{R}$. Then the output error is given by

$$\frac{xy - \tilde{x}\tilde{y}}{xy} = \epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2 = \frac{\tilde{x} - x}{x} + \frac{\tilde{y} - y}{y} + \frac{\tilde{x} - x}{x}\frac{\tilde{y} - y}{y}$$

This means that the multiplication error is directly bounded by the errors in the input, independently from any size difference between x and y. So the componentwise condition number wins. Figure 1 demonstrates that even the componentwise condition number could be improved to accurately reflect the ratio of relative output to relative input error.

Even though the componentwise condition number is a very precise tool, in practice it might be more difficult to compute than the normwise condition number.

1.3 Stability

Well-posedness and the condition number(s) only apply to the mathematical problem. When solving the mathematical problem numerically on a computer, we need to be aware that due to floating point arithmetic rounding errors are introduced. For example, the irrational number π is represented in double precision by these 16 digits:

$$3.141592653589793$$

Again, the question is how does the algorithm propagate these rounding errors? This is addressed by the concept of stability. There are several different concept of stability. We focus on one concept related to the error in the output and one related to the error in the input.

Analogously, to the mathematical problem, we will denote the algorithm with $\tilde{F} : V \mapsto W$, i.e. another map between the same input and output space as for the mathematical problem.

The absolute and relative **forward errors** are defined by

$$\left\|F(x) - \tilde{F}(x)\right\|_W \text{ and } \frac{\left\|F(x) - \tilde{F}(x)\right\|_W}{\|F(x)\|_W}$$

respectively.

The absolute and relative **backward errors** are defined by

$$\beta \text{ and } \frac{\beta}{\|x\|_V}$$

where $\beta = inf\{\|x - \tilde{x}\|_V \,|\tilde{F}(x) = F(\tilde{x})\}$

The backward error is the error between the original input x and some perturbed input $\tilde{x}$. This perturbation is determined by asking which perturbed input one needs to supply to the exact problem to obtain the same output like the algorithm. This question can possibly not be answered in a unique way. Hence, we take the infimum of all such possible input errors.

Now an algorithm is called **forward stable** if the forward error divided by the condition number is ”small”. It is called **backward stable** if the backward error is small for all inputs x.

Key takeaway:
While well-posedness and the condition number relates to the underlying mathematical problem (and are thus independent from the numerical methods ones uses to solve it), stability is inherent to the choice of the algorithm.