Lecture Notes:
Condition Numbers and Geometry

Paul Breiding and Elima Shehu

April 26, 2021

This document contains notes for the lecture *Geometry and Condition* held during the summer 2021 at the Max-Planck-Institute for Mathematics in the Sciences Leipzig.

The lecture is in parts based on the book *Condition: The Geometry of Numerical Algorithms* by Bürgisser and Cucker [BC13].

Both authors have been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 445466444.

Author's addresses:
Paul Breiding, MPI MiS, paul.breiding@mis.mpg.de.
Elima Shehu, MPI MiS, elima.shehu@mis.mpg.de.

Contents

Lecture 1

Motivation and definition of condition numbers

We start with a quote by [Dem96]

> "The correct answers produced by numerical algorithms are seldom exactly correct. There are two sources of error. First, there may be errors in the input data to the algorithm, caused by prior calculations or perhaps measurements errors. Second, there are errors caused by the algorithm itself, due to approximations made within the algorithm. In order to estimate the errors in the computed answers from both these sources, we need to understand how much the solution of a problem is changed, if the input data is slightly perturbed."

The first source of error that Demmel describes is a property of data. The second source is a property of algorithms.
Any algorithm has to cope with the first source of errors!

Example 1.1 (Exact algorithm)**.** Consider the following computational problem: on input $(A, b) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^2$ with $\det(A) \neq 0$ find $x \in \mathbb{R}^2$, such that $Ax = b$.
We consider two different inputs:
Input 1:

$$A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

An error in the data could have given us the input

$$\tilde{A} = \begin{pmatrix} 1 & 1 \\ -1+\epsilon & 1 \end{pmatrix}, \quad \text{and} \quad \tilde{b} = \begin{pmatrix} 2 \\ 0 \end{pmatrix},$$

where $\epsilon > 0$ is small. The *exact* solutions for the equations $Ax = b$ and $\tilde{A}\tilde{x} = \tilde{b}$ are

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \tilde{x} = \tilde{A}^{-1}\tilde{b} = \begin{pmatrix} 1 + \frac{\epsilon}{2-\epsilon} \\ 1 - \frac{\epsilon}{2-\epsilon} \end{pmatrix}.$$

Comparing the errors we find

$$\frac{\|x - \tilde{x}\|}{\left\|A - \tilde{A}\right\|} = \frac{1}{\epsilon} \cdot \frac{\sqrt{2}\epsilon}{2-\epsilon} = \frac{\sqrt{2}}{2-\epsilon} \approx \frac{1}{\sqrt{2}} \tag{1.1}$$

This shows that the error in the input $\left\|A - \tilde{A}\right\|$ is amplified in the out $\|x - \tilde{x}\|$ by a factor of $\frac{1}{\sqrt{2}}$. Next, we consider a second input.

Input 2:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1+10^{-8} \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} 0 \\ 10^{-8} \end{pmatrix}.$$

Consider the following peturbation for a small $\epsilon > 0$.

$$\tilde{A} = \begin{pmatrix} 1 & 1 \\ 1+\epsilon & 1+10^{-8} \end{pmatrix}, \quad \text{and} \quad \tilde{b} = \begin{pmatrix} 0 \\ 10^{-8} \end{pmatrix}.$$

The exact solutions for the equations $Ax = b$ and $\tilde{A}\tilde{x} = \tilde{b}$ are

$$x = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad \text{and} \quad \tilde{x} = \begin{pmatrix} -1 - \frac{\epsilon}{10^{-8}-\epsilon} \\ 1 + \frac{\epsilon}{10^{-8}-\epsilon} \end{pmatrix}.$$

This implies

$$\frac{\|x - \tilde{x}\|}{\left\|A - \tilde{A}\right\|} = \frac{1}{\epsilon} \frac{\mid \epsilon \mid \sqrt{2}}{\mid 10^{-8} - \epsilon \mid} = \frac{\sqrt{2}}{\mid 10^{-8} - \epsilon \mid}. \tag{1.2}$$

This shows that, if $\epsilon \leq 10^{-8}$, then we have $\frac{\|x-\tilde{x}\|}{\|A-\tilde{A}\|} > 10^{8}$.
Even though we applied an exact algorithm to the problem we got different quantities in the output:

$$\text{Output 1: close to the exact solution}$$
$$\text{Output 2: is far from the exact solution}$$

The theory of condition numbers explains these different behaviours of data with respect to perturbations. A general theory for the condition numbers was given by [Ric66]. But, What is a condition number? What a condition number measure? How is it defined? A condition number of a problem measures the sensitivity of the solution to small perturbations in the input data. The condition number depends on the problem and the input data, on the norm used to measure size, and on whether perturbations are measured in an absolute or a relative sense.

Definition 1.2. A computational problem is a function $f : I \longrightarrow O$ from a space of inputs I to a space of outputs O.

For the example from the above: the input space is $I = \{(A, b) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid \det(A) \neq 0\}$, the output space is $O = \mathbb{R}^2$, and $f(A, b) = A^{-1}b$.

Definition 1.3 (Classic definition of condition number; see, e.g., [TB97])**.** Let I and O be finite dimensional normed vector spaces. The (absolute) condition number of f at $x \in I$ is

$$\kappa[f](x) := \lim_{\epsilon \longrightarrow 0} \sup_{y \in I,\, \|x-y\| \leqslant \epsilon} \frac{\|f(x) - f(y)\|}{\|x-y\|} \tag{1.3}$$

We have the following properties of $\kappa[f]$.

Lemma 1.4. *Suppose that $I = \mathbb{R}^n$ and $O = \mathbb{R}^m$.*

1. *If f is differentiable, we have $\kappa[f](x) = \|J(x)\|$ where $J(x) = \left(\frac{\partial f_i}{\partial x_i}\right)$ is the Jacobian of the partial derivatives of $f = (f_1, f_1, \dots, f_m)$ and $\|A\| := \max_{\|x\|=1} \|Ax\|$ is the operator norm.*

2. In the limit $\|x-y\| \to 0$ *we have* $\|f(x)-f(y)\| \leqslant \kappa[f]\,\|x-y\|$*.*

3. $\kappa[f\circ g](x) \leqslant \kappa[f](g(x)) \cdot \kappa[g](x)$ *for* $\kappa[g](x) \leqslant \infty$*.*

What does "small" means? If $\|x\| = 10^{-4}$, is $\|x-y\| = 10^{2}$ small or large?
The ambiguity motivates the following definition

Definition 1.5. The error between x and y relative to x is

$$\mathrm{RelError}(x,y) = \frac{\|x-y\|}{\|x\|}$$

Definition 1.6. The relative condition number is defined

$$\kappa_{\mathrm{REL}}[f](x) := \lim_{\epsilon\to 0} \sup_{\mathrm{RelError}(x,y)\leqslant\epsilon} \frac{\mathrm{RelError}(f(x), f(y))}{\mathrm{RelError}(x,y)} \tag{1.4}$$

Lemma 1.7. *If* f *is differentiable and* $x \neq 0$*, then*

$$\kappa_{\ \mathrm{REL}}[f](x) = \|J(x)\| \frac{\|(x)\|}{\|f(x)\|} \tag{1.5}$$

The condition numbers κ and κ_{REL} are called normwise condition numbers.

Definition 1.8. The componentwise relative condition number is defined as

$$\mathrm{CW}[f](x) := \max_{j} \lim_{\epsilon\to 0} \sup_{\max_i \mathrm{RelError}(x_i,y_i)} \frac{\mathrm{RelError}(f_j(x), f_j(y))}{\max_i \mathrm{RelError}(x_i,y_i)} \tag{1.6}$$

where $x = (x_i)$ and $f = (f_j)$.

Trefethen and Bau [TB97, p. 91]:

> "Both absolute and relative condition numbers have their uses, but the latter are more important in numerical analysis. This is ultimately because the floating point arithmetic used by computers introduces relative errors rather than absolute ones."

Floating point arithmetic

On a computer we can represent numbers using only a finite amount of information so we must work with approximations of real numbers.

The most commonly used number system on a computer are floating point numbers.

Definition 1.9. 1. A floating point number system $F \subseteq \mathbb{R}$ is a subset of the reals of the form: $F = \left\{ \pm \beta^e \sum\limits_{i=1}^{t} \frac{d_i}{\beta_i} \mid 0 \leq d_i \leq \beta - 1, e_{\min} \leq e \leq e_{\max} \right\}$, where $\beta, t, e_{\min}, e_{\max}$ are integers.

- β is called the base
- t is called precision
- $[e_{\min}, e_{\max}]$ is called exponential range.

2. For $G := \left\{ \pm \beta^e \sum\limits_{i=1}^{t} \frac{d_i}{\beta_i} \mid 0 \leq d_i \leq \beta - 1 \right\}$ we put

$$\mathrm{fl} : \mathbb{R} \to G, \quad x \mapsto \underset{y \in G}{\arg\min} \, |x - y|.$$

This is called the *rounding map.*

3. The range of F is $\operatorname{range}(F) := \{x \in \mathbb{R} \mid \beta^{e_{\min}-1} \leq |x| \leq \beta^{e_{\max}} (1 - \beta^{-1})\}$. All numbers in $\operatorname{range}(F)$ are approximated by relative precision $u := \frac{1}{2}\beta^{1-t}$.

Theorem 1.10. *For all* $x \in \operatorname{range}(F)$*, then* $\mathrm{fl}(x) = x(1 + \delta) \in F, \quad |\, \delta \,| \leqslant u$

This theorem shows that every real number lying in $\operatorname{range}(F)$ can be approximated by an element of F with the relative error no larger than $u = \frac{1}{2}\beta^{1-t}$. So, we have:

$$\operatorname{RelError}(x, \mathrm{fl}(x)) = \frac{\|x - x(1 + \delta)\|}{\|x\|} = \|\delta\| \leqslant u,$$

where u is called machine precision and $\epsilon_{\mathrm{MACH}} := \beta^{1-t}$ is called machine epsilon. For more details see [Ste97],[Knu98], and [Hig96].

The bound of the relative error can be refined, so that $\operatorname{RelError}(x, \mathrm{fl}(x)) \leqslant \frac{u}{1+u}$

Remark 1. In [JR18] optimal bounds on relative errors are established for each five basic operations. Each of these bounds is attained for some explicit input values in F and rounding functions under some mild conditions on β and t .

The IEEE 754 standart defines a floating point arithmetic system with $\mathrm{fl}\left(x\circ y\right)=\left(x\circ y\right)\left(1+\delta\right)$, $\mid\delta\mid\leqslant u$ where $\circ\in\left\{+,\ -,\ \times,\diagup,\ \sqrt{\cdot}\right\}$ and formats.

	β	t	e_{min}	e_{max}	u
half (16 bit)	2	11	-14	$16=2^4$	$\approx 5\cdot 10^{-4}$
single (32 bit)	2	24	-125	$128=2^7$	$\approx 6\cdot 10^{-8}$
double (64 bit)	2	53	-1021	$1024=2^{10}$	$\approx 10^{-16}$

For instance, 64-bit floating point number system can approximate any real number within its range with a relative error of at most $u\approx 10^{-16}$.

Bibliography

[BC13] P. Bürgisser and F. Cucker. *Condition: The Geometry of Numerical Algorithms*, volume 349 of *Grundlehren der mathematischen Wissenschaften.* Springer, Heidelberg, 2013.

[Dem96] J. Demmel. *Applied Numerical Linear Algebra.* SIAM, 1996.

[Hig96] N. J. Higham. *Accuracy and stability of numerical algorithms.* Society for Industrial and Applied Mathematics, second edition, 1996.

[JR18] C.-P. Jeannerod and S. M. Rump. On relative errors of floating-point operations: optimal bounds and applications. *Mathematics of Computation*, 87:803–819, 2018.

[Knu98] D. E. Knuth. *The Art of Computer Programming*, volume 2. SIAM, 3 edition, 1998.

[Ric66] John R. Rice. A theory of condition. *SIAM J. Numer. Anal*, 3:287–310, 1966.

[Ste97] P. H. Sterbenz. *Floating-Point Computation.* SIAM, 1997.

[TB97] L. N. Trefethen and D. Bau. *Numerical Linear Algebra.* SIAM, 1997.