Lecture Notes:
Condition Numbers and Geometry

Paul Breiding and Elima Shehu

May 31, 2021

This document contains lecture notes for the course *Geometry and Condition* held during the summer 2021 at the Max-Planck-Institute for Mathematics in the Sciences Leipzig. Each chapter contains material for a 90 minutes lecture.

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

We do not provide proofs for all results, but focus on proving the central theorems. If we don't give a proof for a lemma, a proposition or a theorem, we give a reference instead.

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

1.1 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

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

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

Output 1: close to the exact solution

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 $I = \{(A, b) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \mid \det(A) \neq 0\}$, 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 $f = (f_1, f_1, \ldots, f_m)$ and $\|A\| := \max_{\|x\|=1} \|Ax\|$ is the operator norm.*

2. *We have* $\|f(x) - f(y)\| \leqslant \kappa[f]\|x - y\| + o(\|x - y\|)$ *for small* $\|x - y\|$.

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

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

Definition 1.6. The relative condition number is defined

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

The first part of Lemma 1.4 implies the following formula for the relative condition number.

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

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

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

Definition 1.8. The componentwise relative condition number is defined as

$$\text{CW}[f](x) := \max_j \lim_{\epsilon \to 0} \sup_{\max_i \text{RelError}(x_i, y_i) \leq \epsilon} \frac{\text{RelError}(f_j(x), f_j(y))}{\max_i \text{RelError}(x_i, y_i)}$$

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."

1.2 Floating point arithmetic

The material in this section is based on [Hig96, Sections 2.1–2.3].

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: $F = \left\{ \pm \beta^e \sum_{i=1}^{t} \frac{d_i}{\beta_i} \mid 0 \leq d_i \leq \beta - 1, e_{\min} \leq e \leq e_{\max} \right\}$, 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 $G := \left\{ \pm \beta^e \sum_{i=1}^{t} \frac{d_i}{\beta_i} \mid 0 \leq d_i \leq \beta - 1 \right\}$ we put

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

This is called the *rounding map.*

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

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

This theorem shows that every real number lying in $\mathrm{range}\,(F)$ 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:

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

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 $\mathrm{RelError}\,(x, \mathrm{fl}\,(x)) \leqslant \frac{u}{1+u}$

Remark 1.11. 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 $\mathrm{fl}\,(x \circ y) = (x \circ y)\,(1+\delta)$, $|\,\delta\,| \leqslant u$ where $\circ \in \{+,\,-,\,\times, /,\,\sqrt{\cdot}\}$ 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 $u \approx 10^{-16}$.

Lecture 2

2.1 The loss of precision

Computing with finite-precision arithmetic such as floating-point arithmetic means we will face the negative effect of loss of precision. Recall that u is the machine precision, which the smallest relative difference between two numbers that the computer recognizes. For instance, 64-bit floating point arithmetic is based on $u \approx 10^{-16}$.

Definition 2.1 (Loss of precision)**.** Let β be the base for a floating point number system. The loss of precision in the computation of f is

$$\mathrm{LOP}\left(f, x\right) := \log_\beta \frac{\mathrm{RelError}\left(f\left(x\right), f\left(y\right)\right)}{u},$$

where $y = \mathrm{fl}\left(x\right)$.

The loss of precision tells us how many digits are lost in the computation of f in floating point arithmetic with base β. For instance, of $\mathrm{LOP}\left(f, x\right) = k$ it means that the the first k terms in β-adic expansions of $f(\mathrm{fl}(y))$ and $f(x)$ do not necessarily coincide.
Rule of thumb: $\mathrm{LOP}\left(f, x\right) \approx \log_\beta \kappa_{\mathrm{REL}}[f]\left(x\right)$.
Note that the loss of precision is a property of the problem f and the data x, but not of an algorithm computing $f(x)$. Any algorithm has to cope with loss of precision. An algorithm that for an input x computes accurately $f\left(y\right)$, where $\mathrm{RelError}(x, y)$ is small, is called backward stable.

2.2 Matrix-vector multiplication

The material in this section is based on [BC13, Section O.4] and [Hig96, Section 3.5].
Let us illustrate condition numbers and floating point arithmetic with an example. The problem of matrix-vector multiplication is modelled as on the following three maps:

1. $f : \mathbb{R}^{n\times n} \times \mathbb{R}^n \to \mathbb{R}^n$, $(A, x) \mapsto Ax$ (both A and x are variables);
2. $g : \mathbb{R}^{n\times n} \to \mathbb{R}^n$, $A \mapsto Ax$ (A is a variable, x is fixed);
3. $h : \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto Ax$ (A is fixed, x is a variable).

Depending on where we allow perturbations, we have to choose one of these maps for modelling matrix-vector multiplication. If we allow perturbations in both A and x, we choose f. If we only allow perturbations in A, we choose g. If we only allow perturbations in x, we choose h. It seems reasonable to choose f for modelling the problem. However, the next theorem shows that we should actually choose g!
Regardless on whether or not we allow perturbations on A or x, let us assume that (A, x) is given exactly up to machine precision. Thus, we have the following theorem:

Theorem 2.2. *There is a finite precision, which on input $(A, x) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ computes $\tilde{b}$. If $(\lceil \log_2 n \rceil + 2)^2 u < 1$, then $\tilde{b} = \tilde{A}x$ with*

$$\frac{|\, \tilde{a}_{ij} - a_{ij} \,|}{|\, a_{ij} \,|} \leq (\lceil \log_2 n \rceil + 2)\, u.$$

Before we prove this theorem, let us first discuss its implications: there exists a backward stable algorithm for computing matrix-vector multiplication, and that this algorithm only takes into account errors in A. Even if the input data has errors in both A and x, the algorithm will produce the correct output for $(\tilde{A}, x)$ where $\mathrm{RelError}(A, \tilde{A})$ is bounded by $(\lceil \log_2 n \rceil + 2)\, u$ provided that $(\lceil \log_2 n \rceil + 2)^2 u < 1$. For 64-bit floating point arithmetic we have $u \approx 10^{-16}$ and so the range for n in this case is roughly $n < 2^{(10^8)}$. For instance, if $n = 10^4$, then the bound implied by the theorem is $\approx 2 \cdot 10^{-15}$.
The theorem justifies studying the condition of matrix-vector multiplication through g.

We need an auxiliary lemma in the proof of the theorem.

Lemma 2.3. *Let m be an integer with $mu < 1$. Let $\delta^{(1)}, \ldots, \delta^{(m)}$ be real numbers with $| \delta_i | \leq u$ for each i. Then, we have $\prod_{t=1}^{m} \left(1 + \delta^{(t)}\right) = (1 + \theta_m)$ with $| \theta_m | < \frac{mu}{1-mu}$.*

Proof. Let $\delta := \max |\delta_i|$. Then, by assumption $|\theta_m| \leq (1+\delta)^m - 1$, and so

$$|\theta_m| \leq \sum_{k=1}^{m} \binom{m}{k} \delta^k \leq \sum_{k=1}^{m} (mu)^k = \frac{\left(\sum_{k=1}^{m} (mu)^k\right)(1-mu)}{1-mu} = \frac{mu(1-mu^m)}{1-mu} \leq \frac{mu}{1-mu},$$

the last inequality because $mu < 1$. This finishes the proof. □

Now, we can prove Theorem 2.2.

Proof of Theorem 2.2. Let $b = Ax$, then $b_i = \sum_{j=1}^{n} a_{ij} x_j$.
By definition of the rounding map fl we have $\mathrm{fl}\,(a_{ij}x_{ij}) = (a_{ij}x_{ij})\,(1+\delta_{ij})$, where $|\delta_{ij}| < u$. Adding the first to summands in the expansion of b_i in floating point arithmetic we get

$$\begin{aligned} \mathrm{fl}\,(a_{i1}x_1 + a_{i2}x_2) &= (a_{i1}x_1\,(1+\delta_{i1}) + a_{i2}x_2\,(1+\delta_{i2}))\,(1+\delta_{i12}) \\ &= a_{i1}x_1\,(1+\delta_{i1})\,(1+\delta_{i2}) + a_{i2}x_2\,(1+\delta_{i2})\,(1+\delta_{i12})\,, \end{aligned}$$

where $|\delta_{i12}| < u$. We can group the summands in the expansion of b_i in pairs of two. There, $\lceil \log_2 n \rceil$ many such summands. Thus, if we add the pairs, then add pairs of pairs and so on, we get an algorithm based on a binary tree that computes b_i. Let us put $m := \lceil \log_2 n \rceil + 1$. This is the number of subsequent floating point operators of our algorithm. The $+1$ is due to the fact that we also have to compute each $a_{ij}x_j$ in floating point arithmetic. By construction, we get

$$\mathrm{fl}\left(\sum_{j=1}^{n} a_{ij}x_j\right) = \sum_{j=1}^{n} a_{ij}x_j \prod_{t=1}^{m} \left(1 + \delta_{ij}^{(t)}\right)$$

for some numbers $\delta_{ij}^{(t)}$ with $|\delta_{ij}^{(t)}| < u$. Here, we have freely relabelled the indices of the $\delta_{ij}^{(t)}$ in comparison to above. In the following only the number of factors will be important.
Let $\tilde{a}_{ij} := a_{ij} \prod_{t=1}^{m} (1 + \delta_{ij}^{(t)})$ be the entries of the matrix $\tilde{A} := (\tilde{a}_{ij})$. Our algorithm computes the matrix-vector product $\tilde{b} := \tilde{A}x$.

By assumption, $(m+1)^2u < 1$ and so $mu < 1$. We can apply Lemma 2.3 to get

$$\frac{\mid a_{ij} - \tilde{a}_{ij} \mid}{\mid a_{ij} \mid} = \frac{mu}{1-mu}.$$

It remains to show that $\frac{mu}{1-mu} < (m+1)u$. By assumption, we have $(m+1)^2u < 1$. This implies $0 < 1 - m(m+1)u = 1 - mu - m^2u$, which is equivalent to $m < (1+m)(1-mu)$. Multiplying both sides by $\frac{u}{1-mu}$ gives $\frac{mu}{1-mu} < (m+1)u$ as desired. This finishes the proof. □

The theorem tells us that for matrix-vector multiplication we can focus on the case where there are only errors in the matrix A but not in x. We therefore consider the condition number of the map $g : \mathbb{R}^{n\times n} \to \mathbb{R}^n, A \mapsto Ax$ for fixed x. We study the normwise condition number and componentwise condition number of this map.

2.2.1 Normwise condition number of matrix-vector multiplication

Proposition 2.4. *Measuring the error in the in- and output space of g with the Euclidean norm we get the normwise absolute condition number $\kappa[g](A) = \|x\|$. The normwise relative condition number is $\kappa_{\text{REL}}[g](A) = \frac{\|x\|\|A\|}{\|Ax\|}$*

Proof. We use the formula $\kappa[g](A) = \|Jg(A)\|$, where Jg is the jacobian matrix of first order partial derivatives of g. Let's look at the derivative of

$$g(A) = Ax = \begin{pmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_n^T x \end{pmatrix}$$

where a_i are rows of A. We can also write

$$g(A) = \begin{pmatrix} x^T a_1 \\ \vdots \\ x^T a_n \end{pmatrix}$$

Then, the partial derivative of g with respect to a_{ij} is x_j- This shows the following formula for the Jacobian:

$$Jg\left(A\right) = \begin{pmatrix} x^T & \cdots & 0 \\ 0 & x^T \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x^T \end{pmatrix} \in \mathbb{R}^{n \times n^2}$$

And by definition $\|Jg\left(A\right)\| = \max_{y \in \mathbb{R}^{n^2}, \|y\|=1} \|Jg\left(A\right) y\|$. We partition the $\mathbb{R}^{n^2}$ into blocks of size n and write $y = (y_1, y_2, \cdots, y_n)$, where $y_i \in \mathbb{R}^n$. Then:

$$\|Jg\left(A\right) y\| = \left\| \begin{pmatrix} x^T y_1 \\ \vdots \\ x^T y_n \end{pmatrix} \right\| = \sqrt{\left(x^T y_1\right)^2 + \cdots + \left(x^T y_n\right)^2}$$

We know that $\mid x^T y_i \mid \leq \|x\| \, \|y\|$ by Cauchy-Schwartz and that this inequality is sharp. This fact implies:

$$\max_{y_i, \|y\|=1} \sqrt{\|x\|^2 \|y_1\|^2 + \cdots + \|x\|^2 \|y_n\|^2} = \|x\| \cdot \max_{\|y\|=1} \|y\| = \|x\|$$

This shows the claim for $\kappa[g]\left(A\right)$. For the relative condition number we use the formula $\kappa_{\text{REL}}[g]\left(A\right) = \kappa[g]\left(A\right) \cdot \frac{\|A\|}{\|Ax\|}$. This finishes the proof. □

2.2.2 Componentwise condition number of matrix-vector multiplication

Proposition 2.5. *The relative componentwise condition number satisfies*

$$\text{CW}[g]\left(A\right) \leq \max_i \frac{1}{\mid \cos \sphericalangle \left(a_i, x\right) \mid},$$

where $a_1, \ldots, a_n$ are the rows of A and $\sphericalangle$ denotes the angle between two vectors.

Proof. Let $\tilde{A} = A + E$, where E is the error in A. Let us write $b := Ax$ and $\tilde{b} := \tilde{A}x$. Recall

from Definition 1.8 that the relative componentwise condition number of g at A is

$$\mathrm{CW}\left[g\right](A) = \max_{k} \lim_{\epsilon \to 0} \sup_{\max_{i,j} \mathrm{RelError}(a_{ij}, \tilde{a}_{ij}) \leq \epsilon} \frac{\mathrm{RelError}(b_k, \tilde{b}_k)}{\max_{i,j} \mathrm{RelError}\left(a_{ij}, \tilde{a}_{ij}\right)}, \tag{2.1}$$

where $A = (a_{ij})$, $\tilde{A} = (\tilde{a}_{ij})$ and $b = (b_k)$, $\tilde{b} = (\tilde{b}_k)$. Then, by definition,

$$\mathrm{RelError}\left(a_{ij}, \tilde{a}_{ij}\right) = \frac{\mid e_{ij} \mid}{\mid a_{ij} \mid},$$

where $E = (e_{ij})$. This implies $|e_{ij}| \leq \mathrm{RelError}\left(a_{ij}, \tilde{a}_{ij}\right) |a_{ij}|$ for all all pairs of indices (i, j).

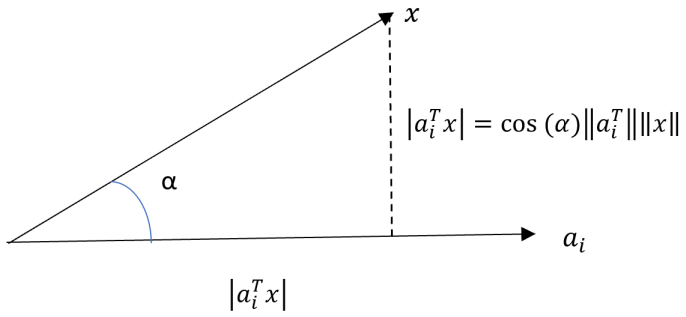


Figure 2.1: Sketch of the geometry in the proof of Proposition 2.5.

Let a_i be the rows of A and e_i are the rows of E. Taking norms of the k-th row vectors we can get the following inequality:

$$\|e_k\| \leq \|a_k\| \max_{i,j} \mathrm{RelError}\left(\tilde{a}_{ij}, \tilde{a}_{ij}\right), \tag{2.2}$$

We also have:

$$\mathrm{RelError}(b_k, \tilde{b}_k) = \frac{\mid e_k^T x \mid}{\mid a_k^T x \mid} \leq \frac{\|e_k\| \, \|x\|}{\mid a_k^T x \mid} \leq \frac{\|a_k\| \, \|x\| \max_{i,j} \mathrm{RelError}\left(a_{ij}, \tilde{a}_{ij}\right)}{\mid a_k^T x \mid};$$

the first inequality by Cauchy-Schwartz and the second by (2.2).

Combined this implies

$$\frac{\text{RelError}(b_k, \tilde{b}_k)}{\max_{i,j} \text{RelError}\left(a_{ij}, \tilde{a}_{ij}\right)} \leq \left| \frac{a_k^T x}{\|a_k\| \, \|x\|} \right|^{-1}.$$

Plugging this into (2.1) we see that

$$\text{CW}[g]\left(A\right) \leq \max_i \frac{1}{|\, \frac{a_i^T x}{\|a_i\| \|x\|} \,|} = \max_i \frac{1}{|\, \cos \sphericalangle \left(a_i, x\right) \,|}.$$

This finishes the proof. □

2.3 Ill-posedness

Let us look again at the general setting where we we have a map $f : I \to O$ between a set of input I and outputs O.

Definition 2.6. The set

$$\Sigma_\mu := \{x \in I \mid \mu[f]\left(x\right) = \infty\}$$

for either $\mu = \kappa$, $\mu = \kappa_{REL}$ or $\mu = \text{CW}$ is called the set of ill-posed inputs. If the condition number μ is clear from the context, we also omit the subscript and simply write Σ.

In the case of matrix vector multiplication the ill-posed inputs are characterized as follows:

$$\begin{aligned}
\Sigma_\kappa &= \emptyset; \\
\Sigma_{\kappa_{\text{REL}}} &= \{A \in \mathbb{R}^{n \times n} \mid Ax = 0\} \quad ; \\
\Sigma_{\text{CW}} &\subseteq \{A \in \mathbb{R}^{n \times n} \mid \exists i : a_i^T = 0\} \\
&= \{A \in \mathbb{R}^{n \times n} \mid \exists i : a_i \in x^\perp\};
\end{aligned}$$

the first and the second equation are by by Proposition 2.4 and the third is by Proposition 2.5. All of these are real algebraic varieties!

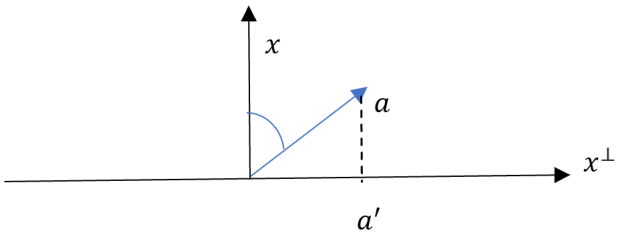
We consider in more detail the CW-condition number and Σ_{CW}. Recall:

$$\mathrm{CW}[g]\,(A) \leq \max_i \frac{1}{|\cos \sphericalangle\,(a_i, x)\,|}.$$

For fixed i, let a_i be the orthogonal projection of a_i onto $x^\perp = \{y \in \mathbb{R}^n \mid x^T y = 0\}$. Then:

$$|\cos \sphericalangle\,(a_i, x)\,| = \frac{\|a_i - a_i'\|}{\|a_i\|} = \min_{y \in x^\perp} \frac{\|a_i - y\|}{\|a_i\|}.$$

This relation is depicted in the picture just below.



This shows that

$$\mathrm{CW}[g]\,(A) \leq \max_i \frac{\|a_i\|}{\min\limits_{y \in x^\perp} \|a_i - y\|} = \max_i \frac{\|a_i\|}{\mathrm{dist}\,(a_i, x^\perp)},$$

where dist is the usual distance from a point to a set in the Euclidean metric. We have shown that the CW-condition number at A is bounded from above by the normalized distance of A to $\{A \in \mathbb{R}^{n \times n} \mid \exists i : a_i \in x^\perp\}$. The latter contains the ill-posed inputs. This is a first instance of an equation called condition number theorem. Such theorems relate condition numbers to inverse distance to ill-posedness; see [Dem87] for a detailed discussion.

2.4 Global analysis of condition numbers

The previous discussion has shown that condition numbers can be arbitrary large. Consequently, with finite precision arithmetic we can't compute for all instances $x \in I$ their output $f\,(x)$. The worst-case analysis can be $\sup_{x \in I} \mu\,(x) = \infty$ for $\mu \in \{\kappa, \kappa_{\mathrm{REL}}, CW\}$.

However, it could be the case that $\mu(x) = \infty$ is extremely rare.
This motivates the following alternative global measures of condition:

Worst-Case Analysis:	$\sup_{x \in I} \mu[f](x)$
Average Analysis:	$\mathbb{E}_{x \sim d} \mu[f](x)$, where d is some probability distribution on the space of inputs I.
Smoothed Analysis:	$\sup_{\overline{x} \in I} \mathbb{E}_{x \sim \mathrm{Unif}(B(\overline{x}, \sigma))} \mu[f]$, where $B(\overline{x}, \sigma)$ is the ball of radius σ centered at x and $\mathrm{Unif}(B(\overline{x}, \sigma))$ is the uniform distribution.
Without the Black Swans:	$\mathbb{E}_{x \sim d}[\mu[f](x) \mid x \notin W]$, where d is some probability distribution on the space of inputs I and $W \subset I$ is a set of small measure.

Usually, it is harder to compute the condition number of a given problem instance than solving the problem itself [Dem88, DDM01]. To avoid this issue Smale initiated the study of the probability distribution of condition numbers. If the condition number of a problem is large with neglectably small probability, then it is reasonable to speak of a "numerically feasible problem", because one can assume that the loss of precision is small.
The different items in the list above represent different viewpoints to this approach: average analysis assumes a probability distribution on the whole input space I; smoothed analysis was invented by Spielman and Teng [ST04] and assumes a local probability distribution around a data point $\overline{x} \in I$. The analysis without the black swans was invented by Amelunxen and Lotz [AL17] and sits between the previous two approaches. It considers a global probability distribution but removes a "difficult set" of small measure W from the analysis.
In the literature, many computational problems have been studied under one of these viewpoints: polynomial eqations [Sma81], linear algebra [BC10, CD05, ES05], systems of polynomial equations in diverse settings [EPR19, SS93], linear systems of inequalities [HM09], linear and convex programming [AB15, ST03], eigenvalue and eigenvectors in the classic and other settings [AC15], polynomial eigenvalue problems [AB19, BK20], tensor decompositions [BV19, BBV19a, BBV19b], and other computational models [CMnPSM02].

Lecture 3

3.1 Matrix norms

Matrix norms are vector norms in vector spaces whose elements are matrices. We have the following definition:

Definition 3.1. Let $A \in \mathbb{R}^{n\times m}$ be a matrix and $x \in \mathbb{R}^m$ be a vector.

1. We define the r-norm of x such that $\|x\|_r := (\sum_{i=1}^m \mid x_i \mid^r)^{\frac{1}{r}}$ and ∞-norm such that $\|x\|_\infty := \max_i \mid x_i \mid$
2. We define the $r \to s$ norm of A such that

$$\|A\|_{r\to s} := \max_{\|x\|_r=1} \|Ax\|_s .$$

3. In case $r = s$ we write $\|A\|_{r\to s} ; = \|A\|_r$.
4. In case $r = s = 2$ $\|A\|_{r\to s} := \|A\|$, the spectral norm of A.
5. The Frobenius norm $\|A\|_F := \left(\sum_{j=1}^m \sum_{i=1}^n \mid a_{ij} \mid^2\right)^{\frac{1}{2}}$.

The $r \to s$ norm defines a distance on the space of matrices:

$$\mathrm{dist}_{r\to s}(A, B) := \|A - B\|_{r\to s}, \quad \text{for } A, B \in \mathbb{R}^{n\times m}. \tag{3.1}$$

If $r = s$, we also write $\mathrm{dist}_r(A, B) := \mathrm{dist}_{r\to s}(A, B)$.

The matrix norms satisfy several properties; see, e.g., [BC13, Section 1.1].

Lemma 3.2. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix then:*

1. *The $r \to \infty$ norm is $\|A\|_{r\to\infty} = \max_{1\le i\le m} \|a_i\|_2$, where a_i is the i-th row of A.*
2. *The Frobenius norm is $\|A\|_F = \text{Trace}\left(A^T A\right)^{\frac{1}{2}}$*
3. *The $r \to s$ norm satisfies $\|AB\|_{r\to s} \le \|A\|_{r\to s'} \|B\|_{s'\to s}$ for any s'.*
4. *The $s^* \to r^*$ norm is $\left\|A^T\right\|_{s^*\to r^*} = \|A\|_{r\to s}$ where: $\frac{1}{s^*} + \frac{1}{s} = 1$ and $\frac{1}{r^*} + \frac{1}{r} = 1$*

3.2 The singular value decomposition

Let $n \le m$. For all $A \in \mathbb{R}^{m\times n}$ there exists orthogonal matrices $U, V \in \mathbb{R}^{m\times n}$ (this means that $U^T U = 1_n$ and $V^T V = 1_n$) and positive values $\sigma_1 \ge \cdots \ge \sigma_n$ such that:

$$A = U\text{diag}\left(\sigma_1, \ldots, \sigma_n\right) V^T.$$

The set $\{\sigma_1, \ldots, \sigma_n\}$ is unique and it is called the set of singular values of A.
Facts:

- The spectral norm is $\|A\| = \sigma_1$.
- The Frobenius norm is $\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$.
- For $n = m$ the inverse of A is $A^{-1} = V\text{diag}\left(\sigma_1^{-1}, \ldots, \sigma_n^{-1}\right) U^T$

Knowing the fact that the 2 norm of A is the largest singular value σ_1 of A while for the inverse it is understood that the 2 norm will be $\frac{1}{\sigma_n}$. For the case $n = m$ we will prove in Theorem 3.7 below that the distance $\text{dist}_2(A, \Sigma)$ from A to the variety of singular matrices $\Sigma := \{A \in \mathbb{R}^{n\times n} \mid \det(A) = 0\}$ is equal to σ_n.

3.3 Condition number of linear equation solving

Recall from very beginning of Lecture 1 the problem of solving linear equations $Ax = b$, where for small perturbation on the input we had large error in the output depending on the problem itself. We want to explain this.

Let D be the set of the matricies A such that $D = \{A \in \mathbb{R}^{n\times n} \mid \det(A) \neq 0\}$ and the two computational problems:

- $\varphi : D \times \mathbb{R}^n \to \mathbb{R}^n$, $(A, b) \mapsto x = A^{-1}b$.
- $\psi : D \to D$, $A \mapsto A^{-1}$.

Those two problems are related. One can compute $\varphi(A, b)$ by first computing $\psi(A) = A^{-1}$ and then multiplying $\psi(A)\, b = x$. And we can compute $\psi(A)$ by computing $\phi(A, e_i)$ for $i = 1, \ldots, n$, where $e_i = (0, \cdots, 0, 1, 0, \cdots, 0)$ are the standard basis vectors.
Let us write

$$\mathrm{C}_{rs}(A) := \|A\|_{r\to s} \left\|A^{-1}\right\|_{r\to s} \tag{3.2}$$

(for $r = s = 2 : \mathrm{C}_{rs}(A) = \|A\| \, \|A^{-1}\|$)
We have the following result connecting $\mathrm{C}_{rs}(A)$ with the condition number of ψ. We prove the theorem in a separate subsection below.

Theorem 3.3. *The relative condition number of ψ of A is*

$$\kappa_{\mathrm{REL}}[\psi](A) = \mathrm{C}_{rs}(A), \tag{3.3}$$

where the error in the input is measured with $\|\cdot\|_{r\to s}$ and in the output with $\|\cdot\|_{s\to r}$.

If we consider the spectral norm ($r = s = 2$) then the number $\mathrm{C}_2(A) := \|A\| \, \|A^{-1}\|$ is often considered *the* condition number in numerical analysis. Another common name for $\mathrm{C}_2(A)$ is "Turing's condition number" [A.M47]. Because $\mathrm{C}_2(A)$ was originally invented by Turing and by von Neumann and Goldstine. Using the SVD of A we can present $\mathrm{C}_2(A)$ in the following alternative way.

Corollary 3.4. *The relative condition number of ψ at the input $A \in \mathbb{R}^{n\times n}$ measured in spectral norm is*

$$\kappa_{\mathrm{REL}}[\psi](A) = \mathrm{C}_2(A) = \|A\| \left\|A^{-1}\right\| = \frac{\sigma_1}{\sigma_n},$$

where σ_1 is the largest singular value of A and σ_n is the smallest.

The importance of C_2 is due to the central role of linear equation solving in numerical mathematics. However, $\mathrm{C}_2(A)$ is the condition number of matrix inversion ψ, and not of linear equation solving ϕ. Of course, we have the relation $\varphi(A,b) = g \circ \varphi(A)$, where $g(M) = M \cdot b$ – inverting a matrix followed by matrix vector multiplication. Therefore, a naive bound for relative condition number of the linear equation solving $\kappa_{\mathrm{REL}}[\varphi](A,b)$ by Lemma 1.4 is $\kappa_{\mathrm{REL}}[\varphi](A) \leq \kappa_{\mathrm{REL}}[g](A^{-1}) \cdot \kappa_{\mathrm{REL}}[\psi](A)$.
But in fact we can show the following; see [BC13, Theorem 1.4].

Theorem 3.5. *The relative condition number of the linear equation solving is*

$$\kappa_{\mathrm{REL}}[\varphi](A,b) = \kappa_{\mathrm{REL}}[g]\left(A^{-1}\right) + \kappa_{\mathrm{REL}}[\psi](A)$$

As we can see we have addition in place of product which is much better since we want the condition number to be small.
Thus this theorem shows that the condition number of linear equation solving is smaller than the one we get by modelling this problem as "invert matrix A", then "multiply with the vector". This underlines that one should not compute inverses for numerically solving $Ax = b$. Instead we should choose another method for linear equation solving like using a LU- or QR-decomposition.
Let us now come back to the example from the beginning of Lecture 1. The input data consists of the two 2×2-matrices

$$A_1 = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1+10^{-8} \end{pmatrix}.$$

Their inverses are

$$A_1^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad A_2^{-1} = 10^8 \begin{pmatrix} 1+10^{-8} & -1 \\ -1 & 1+10^{-8} \end{pmatrix}.$$

Since $A_1^{-1} = \frac{1}{2}A_1^T$ we see that $A_1^T A_1 = 2 \cdot 1_2$. This implies $\|A_1 x\| = 2\|x\|$ and $\|A_1^{-1}x\| = \frac{1}{2}\|x\|$

for any $x \in \mathbb{R}^2$, which gives

$$C_2(A_1) = \|A_1\| \cdot \|A_1^{-1}\| = 2 \cdot \frac{1}{2} = 1.$$

On the other hand, by multiplying with the vector $e_1 = (1, 0)$ we see that $\|A_2\| \geq \|A_2 e_1\| = \sqrt{2}$ and $\|A_2^{-1}\| \geq \|A_2^{-1} e_1\| \geq 10^8$. This implies

$$C_2(A_2) = \|A_2\| \cdot \|A_2^{-1}\| \geq \sqrt{2} \cdot 10^8.$$

This shows that $C_2(A_2) \gg C_2(A_1)$, which in combination with Theorem 3.5 explains the different qualities in the outputs for the examples.

3.3.1 Proof of Theorem 3.3

For the proof of Theorem 3.3 we need an auxiliary lemma.

Lemma 3.6. *Let $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$ be two vectors. Then, we have $\left\|vw^T\right\|_{r\to s} = \|v\|_s \|w\|_{r^*}$ with $\frac{1}{r^*} + \frac{1}{r} = 1$.*

Proof. Let $x \in \mathbb{R}^m$ be a vector. Then: $\left\|vw^T\right\|_{r\to s} = \max_{\|x\|=1} \left\|vw^T x\right\|_s = \mid w^T x \mid \cdot \|v\|_s$. By Hölder's inequality: $\mid w^T x \mid \leq \|w\|_r \cdot \|x\|_r = \|w\|_{r^*}$ and this inequality is sharp. We therefore get $\left\|vw^T\right\|_{r\to s} = \|w\|_{r^*} \cdot \|v\|_s$. The proof is finished. □

Now we can prove Theorem 3.3.

Proof of the Theorem 3.3. Since $\psi(A) = A^{-1} = \frac{1}{\det(A)} \cdot \operatorname{adj}(A)$ (here, $\operatorname{adj}(A)$ is the adjoint of A), the map ψ is a rational polynomial function on D. This implies that ψ is differentiable on D and by Lemma 1.4 the relative condition number of φ can be computed in terms of the norm of the Jacobian J_ψ of ψ:

$$\kappa_{\mathrm{REL}}[\varphi](A) = \|J_\psi(A)\|_{(r\to s)\to(s\to r)} \frac{\|A\|_{r\to s}}{\|A^{-1}\|_{s\to r}} \tag{3.4}$$

Taking the derivative of the equation

$$AB = 1_n = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}$$

we have $\dot{A}B + A\dot{B} = 0$ and knowing that $B = A^{-1}$ we obtain $\dot{B} = -A^{-1}\dot{A}A^{-1}$. This implies

$$\dot{B} = J_\psi(A)\,\dot{A}.$$

By definition, we have $\|J_\psi(A)\|_{(r\to s)\to(s\to r)} = \max_{\|\dot{A}\|_{r\to s}=1} \left\|A^{-1}\dot{A}A^{-1}\right\|_{s\to r}$. Lemma 3.2 (3) gives $\|J_\psi(A)\|_{(r\to s)\to(s\to r)} \le \left(\|A^{-1}\|_{s\to r}\right)^2$ and hence

$$\kappa_{\mathrm{REL}}[\varphi](A) \le \left(\left\|A^{-1}\right\|_{s\to r}\right)^2 \frac{\|A\|_{r\to s}}{\|A^{-1}\|_{s\to r}} = \left\|A^{-1}\right\|_{s\to r}\|A\|_{r\to s} = \mathrm{C}_{rs}(A). \tag{3.5}$$

On the other hand, consider two vectors $v, w \in \mathbb{R}^n$ with $\|v\|_s = 1, \|w\|_{r^*} = 1$, such that

$$\left\|A^{-1}\right\|_{s\to r} = \left\|A^{-1}v\right\|_s \quad \text{and} \quad \left\|A^{-T}\right\|_{r^*\to s^*} = \left\|A^{-T}w\right\|_{r^*},$$

where $\frac{1}{r^*} + \frac{1}{r} = 1$ and $\frac{1}{s^*} + \frac{1}{s} = 1$. Then, for $\dot{A} = vw^T$ we have:

$$J_\psi\left(vw^T\right) = -A^{-1}\cdot v\cdot w^T\cdot A^{-1} = -\left(A^{-1}v\right)\cdot\left(A^{-T}w\right)^T$$

By Lemma 3.6:

$$\begin{aligned}\left\|\left(A^{-1}r\right)\cdot\left(A^{-T}w\right)\right\|_{s\to r} = \left\|A^{-1}v\right\|_s\left\|A^{-T}w\right\|_{r^*} &= \left\|A^{-1}\right\|_{s\to r}\left\|A^{-T}\right\|_{r^*\to s^*} \\ &= \left(\left\|A^{-1}\right\|_{s\to r}\right)^2, \qquad \text{by Lemma 3.2 (3),} \\ &= \|v\|_s\cdot\|w\|_{r^*} = 1\end{aligned}$$

This shows that $\|J_\psi(A)\|_{(r\to s)\to(s\to r)} \ge \left(\|A^{-1}\|_{s\to r}\right)^2$, and therefore $\kappa_{\mathrm{REL}}[\varphi](A) \ge \mathrm{C}_{rs}(A)$, by (3.4). Combining this with (3.5) finishes the proof. □

3.4 Distance to the nearest singular matrix

As in Section 3.2 we define the variety of singular $n \times n$-matrices to be

$$\Sigma := \{A \in \mathbb{R}^{n \times n} \mid \det(A) = 0\}.$$

The goal of this section is prove the following theorem.

Theorem 3.7. *The distance between $A \in \mathbb{R}^{n \times n} \setminus \Sigma$ and the variety of singular matrices is*

$$\operatorname{dist}_{r \to s}(A, \Sigma) = \frac{1}{\|A^{-1}\|_{s \to r}}.$$

In particular, as $\operatorname{dist}_{r \to s}(A, \Sigma) \to 0$ we have $\|A^{-1}\|_{s \to r} \to \infty$.

Theorem 3.3 and Theorem 3.7 together imply that Σ is the locus of ill-posed inputs for matrix inversion:

$$\Sigma = \{A \in \mathbb{R}^{n \times n} \mid \kappa_{\mathrm{REL}}[\psi](A) = \infty\}.$$

in the sense of Definition 2.6. Furthermore, Theorem 3.5 implies that $\Sigma \times \mathbb{R}^n$ is the locus of ill-posed inputs for linear equation solving. This makes the theory of condition numbers consistent with what we would expect as linear equations with data in Σ can't be solved uniquely.

Theorem 3.7 also implies the condition number theorem for matrix inversion.

Corollary 3.8 (The Condition Number Theorem for Matrix Inversion)**.** *The relative condition number of matrix inversion is*

$$\kappa_{\mathrm{REL}}[\psi](A) = \frac{\|A\|_{r \to s}}{\operatorname{dist}_{r \to s}(A, \Sigma)}.$$

This equation means that the condition number of matrix inversion is the normalized distance of A to the variety of singular matrices. The closest to the singular matrix to be the higher is the condition number; Σ is the hypersurface of problems which can't be solved.

Proof of Theorem 3.7. Let the distance in the $r \to s$ norm between out input data A and the nearest ill-posed one be $\operatorname{dist}_{r\to s}(A, \Sigma) = \|A - A'\|_{r\to s}$ with $A' \in \Sigma$ the closest singular matrix to A and let denote $E := A' - A$. Then, there exists $x \in \mathbb{R}^n \setminus \{0\}$ with $(A + E)\, x = 0$. Since $A \notin \Sigma$, it is invertible and so $x = -A^{-1}Ex$. We get

$$\|x\|_r = \left\|A^{-1}Ex\right\|_r \leq \left\|A^{-1}\right\|_{s\to r} \|E\|_{r\to s} \|x\|_r ;$$

by Lemma 3.2 (3). This implies $\|E\|_{r\to s} \geq \frac{1}{\|A^{-1}\|_{s\to r}}$ and so

$$\operatorname{dist}_{r\to s}(A, \Sigma) \geq \frac{1}{\|A^{-1}\|_{s\to r}} \tag{3.6}$$

Next, we find a singular matrix $\tilde{A} \in \Sigma$ such that $\operatorname{dist}_{r\to s}(A, \tilde{A}) \leq (\|A^{-1}\|_{s\to r})^{-1}$. This would show $\operatorname{dist}_{r\to s}(A, \Sigma) \geq (\|A^{-1}\|_{s\to r})^{-1}$ and thus together with (3.6) we would get the desired result. For this, we let $y \in \mathbb{R}^n$ with $\|A^{-1}\|_{s\to r} = \|A^{-1}y\|_r$ and $\|y\|_s = 1$ and

$$x := A^{-1}y \quad \text{and} \quad E := \frac{1}{x^T x} \cdot yx^T.$$

Setting $A' := A - E$ we have

$$A'x = (A + E)\, x = Ax + Ex = y + \frac{-1}{\|x\|_2^2} \cdot yx^T x = 0,$$

and therefore $A' \in \Sigma$. This implies $\operatorname{dist}_{r\to s}(A, \Sigma) \leq \|A - A'\|_{r\to s} = \|E\|_{r\to s}$.
By construction, we have $\|x\|_r = \|A^{-1}\|_{s\to r}$. Furthermore, by Lemma 3.6 (3) we have $\|E\|_{s\to r} = (x^T x)^{-1} \|y\|_s \|x\|_{r^*}$. Observe that $x^T x = \|x^T x\|_{r\to r}$ as a 1×1-matrix. A second application of Lemma 3.6 (3) therefore gives $x^T x = \|x\|_r \|x\|_{r^*}$. By assumption, $\|y\|_s = 1$ so

$$\|E\|_{s\to r} = \frac{1}{\|x\|_r} = \frac{1}{\|A^{-1}\|_{s\to r}}.$$

This shows $\operatorname{dist}_{r\to s}(A, \Sigma) \leq (\|A^{-1}\|_{s\to r})^{-1}$. In (3.6) we shown the other inequality. We therefore finally get $\operatorname{dist}_{r\to s}(A, \Sigma) = (\|A^{-1}\|_{s\to r})^{-1}$. The proof is finished. □

Lecture 4

In this lecture we will conduct average analyses (see Section 2.4) of the two problems matrix-vector multiplication and matrix inversion.

4.1 Average Analysis of Matrix-Vector-Multiplication

Recall the problem of matrix vector multiplication: $g : \mathbb{R}^{n\times n} \to \mathbb{R}^n$, $A \to Ab$, where A is the variable and $b \in \mathbb{R}^n$ is fixed. The componentwise condition number of the computation g is

$$\mathrm{CW}[g](A) := \max_j \mathrm{CW}_j[g](A) \tag{4.1}$$

where

$$\mathrm{CW}_j[g](A) = \lim_{\varepsilon\to 0} \sup_{\max_{i,k} \mathrm{RelError}\left(A_{i,k},\tilde{A}_{i,k}\right)\le\varepsilon} \frac{\mathrm{RelError}\left(g(A)_j, g(\tilde{A})_j\right)}{\max_{i,k} \mathrm{RelError}\left(A_{ik}, \tilde{A}_{ik}\right)}$$

As we know the Average Analysis assumes a probability distribution on the whole input space. In this section we assume the Gaussian distribution for the input A. So we have the following theorem:

Theorem 4.1. *Let $A \in \mathbb{R}^{n\times n}$ be a $n\times n$ matrix whose entries are independent and identically distributed standard Gaussian random variables. Then, for each j we have*

1. $\mathrm{Prob}\{\mathrm{CW}_j[g](A) \ge \frac{1}{\varepsilon}\} \le \sqrt{\frac{2n}{\pi}} \cdot \varepsilon$
2. $\mathbb{E} \log \mathrm{CW}_j[g](A) \le \frac{1}{2}\log n + 1$

The theorem implies that the expected loss of precision for matrix-vector multiplication (see Definition 2.1) is approximately $\frac{1}{2}\log n + 1$. For $n = 10^k$ we get an expected loss of precision in base $\beta = 10$ of about $\frac{k}{2} + 1$. This means that the loss of precision is expected to grow linearly with k. For instance, for matrix-vector multiplicaton with square matrices of side length ten thousand we expect to lose at most $\frac{4}{2} + 1 = 3$ digits in the 10-adic expansion.
For the proof of Theorem 4.1 we need three auxiliary lemmata. The first one is the following result. For a proof see [BC13, Corollary 2.3].

Lemma 4.2. *For an integrable function: $f : S^{n-1} \to \mathbb{R}$, where $S^{n-1} = \{x \in \mathbb{R}^n \mid x^T x = 1\}$ is the unit sphere, we have $\int_{S^{n-1}} f(x)\, \mathrm{d}x = \int_{(0,\pi)\times S^{n-2}} f(\lambda(u,\theta)) (\sin\theta)^{n-2}\, \mathrm{d}u\, \mathrm{d}\theta$, where*

- $\lambda(u,\theta) = (\cos\theta, \sin\theta \cdot u)$
- $\mathrm{d}x$, $\mathrm{d}u$, $\mathrm{d}\theta$ *are the differential forms with respect to Lebesgue measure on each space.*

The second lemma we need is a general lemma about expectations of nonnegative random variables. For the expectation we need another lemma:

Lemma 4.3. *Let $Z : M \to [0,\infty)$ be a random variable that only takes nonnegative values and has a density $\rho(x)$. Then, $\mathbb{E}Z = \int_0^\infty \mathrm{Prob}\{Z \geq t\}\, dt$.*

Proof. Let $N := \{(x,t) \in M \times [0,\infty) \mid Z(x) \geq t\}$ and $A := \int_N \rho(x)\, \mathrm{d}(x,t)$. We use Fubini's theorem: on the one hand, $A = \int_M \rho(x) \left(\int_0^{Z(x)} \mathrm{d}t\right) \mathrm{d}x = \int_M \rho(x) Z(x)\, \mathrm{d}x = \mathbb{E}Z$, and, on the other hand: $A = \int_0^\infty \left(\int_{Z(x)\geq t} \rho(x)\, \mathrm{d}x\right) \mathrm{d}t = \int_0^\infty \mathrm{Prob}(Z \geq t)\, \mathrm{d}t$. This shows the claimed equality. The proof is finished. (One can prove this lemma without densities, but we don't need this case here.) □

Recall from Proposition 2.5 that the componentwise condition number component satisfies:

$$\mathrm{CW}_j[g](A) \leq \frac{1}{|\cos \sphericalangle(a_j, b)|},$$

where a_j = j-th row of A. The right-hand side depends only on $\frac{a_j}{\|a_j\|}$ and $\frac{b}{\|b\|}$, because

$$\cos \sphericalangle(a_j, b) = \frac{a_j^T b}{\|a_j\|\, \|b\|}.$$

Facts:

- If A has i.i.d. standard Gaussian entries, a_j has standard Gaussian entries.
- If a_j has standard Gaussian entries, then $\frac{a_j}{\|a_j\|}$ has the uniform distribution in S^{n-1}.

We write $\mathrm{Unif}\,(S^{n-1})$ for uniform distribution on the sphere. By definition, a measurable subset $A \subset S^{n-1}$ has the probability $\mathrm{Prob}_{a\sim\mathrm{Unif}(S^{n-1})}(A) = \mathrm{vol}(A)/\mathrm{vol}(S^{n-1})$, where $\mathrm{vol}(\cdot)$ is the $n-1$-dimensional volume. We need a result on the ratio of volumes of spheres:

Lemma 4.4. *We have* $\mathrm{vol}(S^{n-2})/\mathrm{vol}(S^{n-1}) \leq \sqrt{\frac{n}{2\pi}}$.

Proof. See [BC13, Lemma 2.25]. □

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. The probability that we consider is

$$\mathrm{Prob}\{\mathrm{CW}_j[g]\,(A) \geq \varepsilon^{-1}\} = \ \mathrm{Prob}_{a\sim\mathrm{Unif}(S^{n-1})}\{\cos\sphericalangle\,(a,b) \leq \varepsilon\} = \frac{\mathrm{vol}(A)}{\mathrm{vol}(S^{n-1})}, \tag{4.2}$$

where

$$A := \{a \in S^{n-1} \mid \cos\sphericalangle\,(a,b) \leq \varepsilon.\}.$$

Since the uniform distribution is invariant under orthogonal transformations we can assume that $b = (1,0,\ldots,0)$. Let us also write $a = \lambda\,(u,\theta) = (\cos\theta, \sin\theta\cdot u)$. Then, $\sphericalangle\,(a,b) = \cos\theta$. Lemma 4.2 implies:

$$\mathrm{vol}(A) = \int_{[\arccos(\varepsilon),\arccos(-\varepsilon)]\times S^{n-2}} (\sin\theta)^{n-2}\ \mathrm{d}u\ \ d\theta = \mathrm{vol}(S^{n-2}) \int_{\arccos(\varepsilon)}^{\arccos(-\varepsilon)} (\sin\theta)^{n-2}\ \mathrm{d}\theta;$$

the second equality because the integrand does not depend on $u \in S^{n-2}$, so we may integrate out this factor. Now, we use that $0 \leq \sin\theta < 1$ for $\theta \in (0,\pi)$ to bound $(\sin\theta)^{n-2} \leq \sin\theta$. This yields

$$\mathrm{vol}(A) \leq \mathrm{vol}(S^{n-2}) \int_{\arccos(\varepsilon)}^{\arccos(-\varepsilon)} \sin\theta\,\mathrm{d}\theta = \mathrm{vol}(S^{n-2})\,(\varepsilon+\varepsilon) = 2\varepsilon\mathrm{vol}(S^{n-2}). \tag{4.3}$$

We plug this into Equation (4.2) to obtain

$$\text{Prob}\{\text{CW}_j[g]\,(A) \geq \varepsilon^{-1}\} \leq 2\varepsilon \frac{\text{vol}(S^{n-2})}{\text{vol}(S^{n-1})} \leq \sqrt{\frac{2n}{\pi}}, \tag{4.4}$$

where we have used $\text{vol}(S^{n-2})/\text{vol}(S^{n-1}) \leq \sqrt{\frac{n}{2\pi}}$, from Lemma 4.4. This proves the first part of the theorem.

Now, we prove the second part on the expectation. Using Lemma 4.3 we get:

$$\mathbb{E} \log \text{CW}_j[g]\,(A) = \int_0^\infty \text{Prob}\{\log \text{CW}_j[g]\,(A) \geq t\}\,\mathrm{d}t = \int_0^\infty \text{Prob}\{\text{CW}_j[g]\,(A) \geq e^t\}\,\mathrm{d}t$$

We now split the last integral at $t_0 \in [0, \infty)$ to obtain

$$\mathbb{E} \log \text{CW}_j[g]\,(A) = \int_0^{t_0} \text{Prob}\{\text{CW}_j[g]\,(A) \geq e^t\}\,\mathrm{d}t + \int_{t_0}^\infty \text{Prob}\{\text{CW}_j[g]\,(A) \geq e^t\}\,\mathrm{d}t.$$

Since a probability is always bounded from above by 1 we can bound the left integral by t_0. Furthermore, using (4.4) we can bound the right integral by $\int_{t_0}^\infty \sqrt{\frac{2n}{\pi}} \cdot e^{-t}\,\mathrm{d}t = \sqrt{\frac{2n}{\pi}} \cdot e^{-t_0}$. This yields

$$\mathbb{E} \log \text{CW}_j[g]\,(A) \leq t_0 + \sqrt{\frac{2n}{\pi}} \cdot e^{-t_0}.$$

If we now use $t_0 = \log \sqrt{\frac{2n}{\pi}}$ we obtain

$$\mathbb{E} \log \text{CW}_j[g]\,(A) \leq \log \sqrt{\frac{2n}{\pi}} + 1 \leq \frac{1}{2} \log n + 1.$$

This finishes the proof □

4.2 Avarage Analysis of Matrix-Inversion

Recall from Equation (3.2) Turing's condition number $\text{C}_{rs}\,(A) = \|A\|_{r\to s}\,\|A^{-1}\|_{r\to s}$. By Theorem 3.3, $\text{C}_{rs}\,(A)$ is the relative condition number of matrix inversion $\psi(A) = A^{-1}$ if we measure the input with $r \to s$ norm and the output with $s \to r$ norm. As before let $a_1, ..., a_n$ be the rows of the matrix $A \in \mathbb{R}^{n\times n}$. We have the following average anaysis for $\text{C}_{2\infty}$.

Theorem 4.5. *Let $A \in \mathbb{R}^{n\times n}$ such that the rows $a_1, ..., a_n$ are independent and identically uniform distributed on S^{n-1}. Then:*

1. $\text{Prob}\{C_{2\infty}(A) \geq \varepsilon^{-1}\} \leq \sqrt{\frac{2}{\pi}} n^{\frac{5}{2}} \varepsilon$.
2. $\mathbb{E} \log C_{2\infty}(A) \leq \frac{5}{2} \log n + 1$.

As in the previous section we conclude from this theorem that the expected loss of precision for an input of size $10^k \times 10^k$ is about $\frac{5k}{2} \log 10 + 1$. For instance, if $k = 4$ we get an expected loss of precision in base $\beta = 10$ of at about $\frac{5\cdot 4}{2} + 1 = 11$.

Before we prove the theorem, let us first discuss its assumptions. Why is the assumption that the rows of A are uniformly on the sphere justified. For this, we consider $A \in \mathbb{R}^{n\times n}$. Let $\alpha_i := \|a_i\|$ be the norm of the i-th row of A, and consider the matrix $\tilde{A}$ with rows a_i/α_i. Then, $\psi(A) = \text{diag}(\alpha_1^{-1}, \ldots, \alpha_n^{-1}) \cdot \psi(\tilde{A})$ – for computing $\psi(A)$ it suffices to compute $\psi(\tilde{A})$ plus a matrix-vector multiplicaton. From a different perspective, if we consider $\mathrm{C}_{2\infty}(A)$ in the context of linear equation solving (see Theorem 3.5), then $Ax = b$ if and only if $\tilde{A}x = \tilde{b}$, where $\tilde{b} = (b_i/\alpha_i)_{i=1}^n$. This shows that we get equivalent problems when considering $\tilde{A}$ instead of A. The process of going from A to $\tilde{A}$ is called preprocessing.

Why do we need this twist and why is this helpful? We know from Corollary 3.8 that $C_{2\infty}(A) = \min_{S\in\Sigma} \frac{\max_i \|a_i\|}{\max_j \|a_j - s_j\|}$, where $\Sigma = \{A \in \mathbb{R}^{n\times n} \mid \det(A) = 0\}$ is the determinant surface and $s_1, ..., s_n$ are the rows of S. For a fixed S we have:

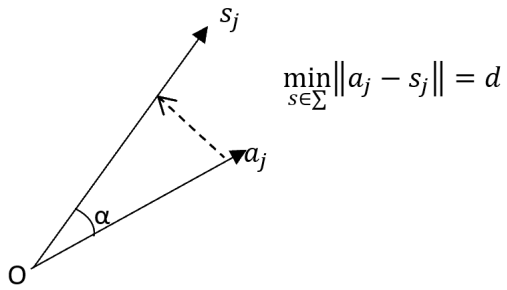
$$\frac{\max_i \|a_i\|}{\max_j \|a_j - s_j\|} = \min_j \frac{\max_i \|a_i\|}{\|a_j\| \left\|\tilde{a}_j - \frac{s_j}{\|a_j\|}\right\|} \geq \min_j \frac{1}{\left\|\tilde{a}_j - \frac{s_j}{\|a_j\|}\right\|} \geq \text{dist}_{2\infty}\left(\tilde{A}, \Sigma\right),$$

which implies

$$C_{2\infty}(A) \geq C_{2\infty}\left(\tilde{A}\right).$$

From the above evidence, we can see that preprocessing gives us an easier problem from the numerical point of view. In other words, we have done this transformation to get an equivalent but numerically better conditioned problem.

Let us now prove Theorem 4.5.



Proof of Theorem 4.5. Let $A \in (S^{n-1})^n$. Since Σ is a cone:

$$\mathrm{dist}_{2\infty}(A,\Sigma) = \min_{S\in\Sigma}\max_j \|a_j - s_j\| = \min_{S\in\Sigma}\max_j \sin\sphericalangle(a_j,s_j) = \min_{S\in\Sigma_s}\max_j \sin\sphericalangle(a_j,s_j)$$

Let $\Sigma_s := \Sigma \cap (S^{n-1})^n$ be the variety of singular matrices whose rows have unit lengths. So, we have shown that

$$\mathrm{Prob}\{C_{2\infty}(A) \geq \varepsilon^{-1}\} = \ \mathrm{Prob}\{\min_{S\in\Sigma_s}\max_j \sphericalangle(a_j,s_j) \leq \varepsilon\}.$$

For a fixed $A \in (S^{n-1})^n$, let $u, v \in \mathbb{R}^n$ be such that $\mathrm{dist}_{2\infty}(A,\Sigma) = \left\|uv^T\right\|_{2\infty}$ (we know from the proof of Theorem 3.7 that the minimizer is rank one). Note:

- $\left\|uv^T\right\|_{2\infty} = \|v\|\,\|u\|_\infty = \|v\| \cdot \max_i \mid u_i \mid$
- $\mathrm{dist}_{\infty,2}\left(A^T,\Sigma\right) = \left\|vu^T\right\|_{\infty,2} = \|v\|\,\|u\|_\infty$

Without loss of generality we can assume $\|u\|_\infty = |u_1|$. We also know: $A^T u = v$. By Cramer's rule: $u_1 \cdot \det(A) = \det([v, a_2, \ldots, a_n])$, which implies

$$\det(A - U) = 0, \text{ where } U = [\frac{1}{u_1}v \mid 0 \cdots \mid 0].$$

The norm of U is

$$\|U\|_{\infty\to 2} = \frac{1}{|u_1|}\|v\| \leq \|v\|\,\|u\|_\infty \cdot n, \tag{4.5}$$

where in the last inequality we have used that $\|u\|_\infty \geq \sqrt{n}\|u\|_2$ twice. We shown the

following

$$\text{Prob}\{\min_{S\in\Sigma_s}\max_j \sphericalangle(a_j, s_j) \leq \varepsilon\} \leq \text{Prob}(W_1 \cup \cdots \cup W_n),$$

where $W_i \subset \mathbb{R}^{n\times n}$ is defined as

$$W_i := \{A = [a_1, \ldots, a_n] \mid \exists c \in \mathbb{R}^n : \sin\sphericalangle(a_i, c) \leq n\varepsilon, [a_1, \ldots, a_{i-1}, c, a_{i+1}, \ldots a_n] \in \Sigma\}.$$

By symmetry, the W_i all have equal probability. The probability of the union $W_1 \cup \cdots \cup W_n$ is bounded from above by the sum of the probabilities for each of this events called the union bound and so we have

$$\text{Prob}\{\min_{S\in\Sigma_s}\max_j \sphericalangle(a_j, s_j) \leq \varepsilon\} \leq n\text{Prob}(W_1).$$

The volume of W_1 is

$$\text{vol}(W_1) = \int_{a_n} \cdots \int_{a_2}\int_{a_1} \text{vol}\left(n\varepsilon\text{-neighborhood of span}\{a_2, \ldots, a_n\}\right) \mathrm{d}a_1\mathrm{d}a_2 \cdots \mathrm{d}a_n.$$

By rational symmetry the inner integral does not depend on $a_2, \ldots, a_n$ and from Equation (4.3) we see that it is bounded by

$$\int_{a_1} \text{vol}\left(n\varepsilon\text{-neighborhood of span}\{a_2, \ldots, a_n\}\right) \mathrm{d}a_1 \leq 2\varepsilon n\text{vol}(S^{n-2}).$$

So we get $\text{vol}(W_1) \leq 2\varepsilon n\text{vol}(S^{n-2})\text{vol}(S^{n-1})^{n-1}$. The probability of W_1 is then

$$\text{Prob}(W_1) = \frac{\text{vol}(W_1)}{\text{vol}(S^{n-1})^n} \leq 2\varepsilon n\frac{\text{vol}(S^{n-2})}{\text{vol}(S^{n-1})} \leq n^{\frac{3}{2}}\varepsilon\sqrt{\frac{2}{\pi}},$$

the last inequality by Lemma 4.4. This implies

$$\text{Prob}\{C_{2\infty}\}(A) \leq n^{\frac{5}{2}}\varepsilon\sqrt{\frac{2}{\pi}}.$$

We have thus proved the first part of the theorem. For the second part we proceed exactly like we did for the second part of Theorem 4.1. The proof is finished. □

Bibliography

[AB15] D. Amelunxen and P. Bürgisser. Probabilistic analysis of the Grassmann condition number. *Found. Comput. Math.*, 15(1):3–51, 2015. (cited on p. 15).

[AB19] D. Armentano and C. Beltrán. The polynomial eigenvalue problem is well conditioned for random inputs. *SIAM J. Matrix Anal. Appl.*, 40(1):175–193, 2019. (cited on p. 15).

[AC15] D. Armentano and F. Cucker. A randomized homotopy for the Hermitian eigenpair problem. *Found. Comput. Math.*, 15(1):281–312, 2015. (cited on p. 15).

[AL17] D. Amelunxen and M. Lotz. Average-case complexity without the black swans. *Journal of Complexity*, 41:82–101, 2017. (cited on p. 15).

[A.M47] A.M.Turing. Rounding-off errors in matrix processes. *National Physical Laboratory, Teddington, Middlesex*, 1947. (cited on p. 18).

[BBV19a] C. Beltrán, P. Breiding, and N. Vannieuwenhoven. The average condition number of most tensor rank decomposition problems is infinite, 2019. (cited on p. 15).

[BBV19b] C. Beltrán, P. Breiding, and N. Vannieuwenhoven. Pencil-based algorithms for tensor rank decomposition are not stable. *SIAM Journal on Matrix Analysis and Applications*, 40(2):739–773, Jan 2019. (cited on p. 15).

[BC10] P. Bürgisser and F. Cucker. Smoothed analysis of Moore-Penrose inversion. *SIAM J. Matrix Anal. Appl.*, 31(5):2769–2783, 2010. (cited on p. 15).

[BC13] P. Bürgisser and F. Cucker. *Condition: The Geometry of Numerical Algorithms*, volume 349 of *Grundlehren der mathematischen Wissenschaften*. Springer, Heidelberg, 2013. (cited on p. ii, 8, 16, 19, 25, 26).

[BK20] C. Beltrán and K. Kozhasov. The real polynomial eigenvalue problem is well conditioned on the average. *Found. Comput. Math.*, 20(2):291–309, 2020. (cited on p. 15).

[BV19] P. Breiding and N. Vannieuwenhoven. On the average condition number of tensor rank decompositions. *IMA Journal of Numerical Analysis*, 40(3):1908–1936, Jun 2019. (cited on p. 15).

[CD05] Z. Chen and J. J. Dongarra. Condition numbers of Gaussian random matrices. *SIAM J. Matrix Anal. Appl.*, 27(3):603–620, 2005. (cited on p. 15).

[CMnPSM02] D. Castro, J. L. Montaña, L. M. Pardo, and J. San Martín. The distribution of condition numbers of rational data of bounded bit length. *Found. Comput. Math.*, 2:1–52, 2002. (cited on p. 15).

[DDM01] J. W. Demmel, B. Diament, and G. Malajovich. On the complexity of computing error bounds. *Found. Comput. Math.*, 1:101–125, 2001. (cited on p. 15).

[Dem87] J. W. Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Courant Institute of Mathematical Sciences, 251 Mercer Str, 10012, New York, NY, USA*, 51:251–289, 1987. (cited on p. 14).

[Dem88] J. W. Demmel. The probability that a numerical analysis problem is difficult. *Math. Comp.*, 50:449–480, 1988. (cited on p. 15).

[Dem96] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1996. (cited on p. 1).

[EPR19] A A. Ergür, G. Paouris, and J. M. Rojas. Probabilistic condition number estimates for real polynomial systems I: A broader family of distributions. *Found. Comput. Math.*, 19(1):131–157, 2019. (cited on p. 15).

[ES05] A. Edelman and B. D. Sutton. Tails of condition number distributions. *SIAM J. Matrix Anal. Appl.*, 27:547–560, 2005. (cited on p. 15).

[Hig96] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, second edition, 1996. (cited on p. 5, 8).

[HM09] R. Hauser and T. Müller. Conditioning of random conic systems under a general family of input distributions. *Found. Comput. Math.*, 9:335–358, 2009. (cited on p. 15).

[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. (cited on p. 6).

[Knu98] D. E. Knuth. *The Art of Computer Programming*, volume 2. SIAM, 3 edition, 1998. (cited on p. 5).

[Ric66] J. R. Rice. A theory of condition. *SIAM J. Numer. Anal*, 3:287–310, 1966. (cited on p. 3).

[Sma81] S. Smale. The fundamental theorem of algebra and complexity theory. *Bull. Amer. Math. Soc.*, 4:1–36, 1981. (cited on p. 15).

[SS93] M. Shub and S. Smale. Complexity of Bezout's theorem. II. Volumes and probabilities. In *Computational algebraic geometry (Nice, 1992)*, volume 109 of *Progr. Math.*, pages 267–285. Birkhäuser Boston, Boston, MA, 1993. (cited on p. 15).

[ST03] D. A. Spielman and S.-H. Teng. Smoothed analysis of termination of linear programming algorithms. volume 97, pages 375–404. 2003. ISMP, 2003 (Copenhagen). (cited on p. 15).

[ST04] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, May 2004. (cited on p. 15).

[Ste97] P. H. Sterbenz. *Floating-Point Computation*. SIAM, 1997. (cited on p. 5).

[TB97] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997. (cited on p. 3, 4).