MA4230 Matrix Computation

Chapter 5: Conditioning and Stability

5.1 Conditioning of mathematical problems

5.2 Floating point numbers and floating point arithmetic

5.3 Stability of numerical algorithms

5.4 Stability of solution algorithms for linear systems

5.1 Conditioning of mathematical problems

Well-conditioned and ill-conditioned problems

"Conditioning" $\Longleftrightarrow$ perturbation behavior of mathematical problems.

We regard a **(mathematical) problem** as a function

$$f : X \to Y$$

with normed vector spaces X (*the data space*) and Y (*the solution space*).

- A problem f, together with a particular data point $x \in X$ ((f, x) is called **problem instance** or simply **problem**), is **well-conditioned** if small changes in x only lead to small changes in $f(x)$.
- Otherwise, i.e., if a small change in x can lead to a large change in $f(x)$, we call the problem (instance) **ill-conditioned**.

$\Longrightarrow$ How can we decide whether a problem is well- or ill-conditioned?

Condition number

Condition number = measure for the perturbation behavior of a problem.

Definition (Absolute and relative condition number)

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces. For a problem $f : X \to Y$ and a given data point $x \in X$, we define

(i) the **absolute condition number** $\hat{\kappa} = \hat{\kappa}(x)$ by

$$\hat{\kappa} := \lim_{\delta \to 0} \sup_{\substack{\Delta x \in X \\ 0 < \|\Delta x\|_X \leq \delta}} \frac{\|f(x + \Delta x) - f(x)\|_Y}{\|\Delta x\|_X},$$

(ii) and, if $x \in X \backslash \{0\}$ and $f(x) \in Y \backslash \{0\}$, the **relative condition number** $\kappa = \kappa(x)$ by

$$\kappa := \lim_{\delta \to 0} \sup_{\substack{\Delta x \in X \\ 0 < \|\Delta x\|_X \leq \delta}} \frac{\frac{\|f(x+\Delta x) - f(x)\|_Y}{\|f(x)\|_Y}}{\frac{\|\Delta x\|_X}{\|x\|_X}}.$$

If κ is small (e.g., $1, 10, 100$), the problem is called well-conditioned, and if κ is large (e.g., $10^6, 10^{12}$), the problem is called ill-conditioned.

Condition numbers: a special case

Let $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ with chosen norms $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$ and $\|\cdot\|_{(m)}$ on $\mathbb{R}^m$. Consider a problem $f : X \to Y$, a given data point $x \in \mathbb{R}^n$, and assume that f is differentiable at x.

Then, we have

$$\hat{\kappa} = \|J_f(x)\|_{(m,n)}, \qquad \kappa = \frac{\|J_f(x)\|_{(m,n)}\|x\|_{(n)}}{\|f(x)\|_{(m)}}$$

where $J_f(x) \in \mathbb{R}^{m\times n}$ denotes the Jacobian of f at x whose entries are given by $(J_f(x))_{ij} = \partial_j f_i$, and $\|\cdot\|_{(m,n)}$ denotes the matrix norm on $\mathbb{R}^{m\times n}$ induced by the norms $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$ and $\|\cdot\|_{(m)}$ on $\mathbb{R}^m$.

Example 1: constant multiple of a real number

For $X = Y = \mathbb{R}$ with norm $\|\cdot\|_{(1)} := |\cdot|$ on $\mathbb{R}$, consider the problem

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto 7x,$$

i.e., the problem of obtaining $7x$ from $x \in \mathbb{R}$.

Note that f is differentiable on $\mathbb{R}$ and $J_f(x) = \big(f'(x)\big) = \big(7\big) \in \mathbb{R}^{1\times 1}$ for all $x \in \mathbb{R}$. Hence,

$$\kappa = \frac{\|J_f(x)\|_{(1,1)}\|x\|_{(1)}}{\|f(x)\|_{(1)}} = \frac{|7||x|}{|7x|} = 1.$$

$\Longrightarrow$ The problem is well-conditioned.

Example 2: addition of two real numbers

For $X = \mathbb{R}^2$ with norm $\|\cdot\|_{(2)} := \|\cdot\|_2$ on $\mathbb{R}^2$, and $Y = \mathbb{R}$ with norm $\|\cdot\|_{(1)} := |\cdot|$ on $\mathbb{R}$, consider the problem

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad (x_1, x_2) \mapsto x_1 + x_2,$$

i.e., the problem of finding the sum of two real numbers.

f is differentiable on $\mathbb{R}^2$ and $J_f(x) = \begin{pmatrix} \partial_1 f(x) & \partial_2 f(x) \end{pmatrix} = \begin{pmatrix} 1 & 1 \end{pmatrix} \in \mathbb{R}^{1\times 2}$ for all $x \in \mathbb{R}^2$. Hence,

$$\kappa = \frac{\|x\|_{(2)}}{\|f(x)\|_{(1)}} \|J_f(x)\|_{(1,2)} = \frac{\sqrt{x_1^2 + x_2^2}}{|x_1 + x_2|} \sup_{\substack{z\in\mathbb{R}^2 \\ \|z\|_2=1}} |\begin{pmatrix} 1 & 1 \end{pmatrix} z| = \sqrt{2}\frac{\sqrt{x_1^2 + x_2^2}}{|x_1 + x_2|}.$$

Note that when $x_2 \approx -x_1$ and $x_1 \neq 0$ we have that κ is large and the problem is ill-conditioned. This phenomenon is called **cancellation error**.

Example 3: polynomial root-finding

Consider the polynomials

$$p_1(t) := t^2 - 2t + 1, \qquad [\text{double root } t = 1],$$

$$p_x(t) := t^2 - 2t + x \quad \text{for} \quad x \leq 1, \qquad [\text{roots } t = 1 \pm \sqrt{1-x}].$$

Set $X = Y = \mathbb{R}$ with norm $|\cdot|$ on $\mathbb{R}$ and define the problem $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto f(x)$ by setting $f(x)$ to be the largest root of p_x if $x \leq 1$, and set $f(x) := f(1) = 1$ for all $x > 1$ (note this doesn't introduce perturbation errors to the right of $x = 1$ as $f(1 + \Delta x) - f(1) = 0$ for $\Delta x > 0$).

Let us show that $\kappa(1) = \infty$, i.e., the problem is **severely ill-conditioned**. Observe that $f(1) = 1$. If we perturb $x = 1$ by some $\Delta x < 0$, we find a change in $f(x)$ of size $|f(1 + \Delta x) - f(1)| = \sqrt{-\Delta x}$. (If we perturb $x = 1$ by some $\Delta x > 0$, we find no change in $f(x)$). Hence, for any $\delta > 0$ have

$$\sup_{\Delta x \in [-\delta, \delta] \setminus \{0\}} \frac{|f(1 + \Delta x) - f(1)|}{|\Delta x|} \frac{|1|}{|f(1)|} = \sup_{\Delta x \in [-\delta, 0)} \frac{\sqrt{-\Delta x}}{-\Delta x} = \infty.$$

$$\implies \kappa(1) = \infty.$$

Central conditioning problems of numerical linear algebra

- Conditioning of matrix-vector multiplication
- Conditioning of linear systems
- Conditioning of least squares problems

Conditioning of matrix-vector multiplication

Let $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ with chosen norms $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$ and $\|\cdot\|_{(m)}$ on $\mathbb{R}^m$, and let $A \in \mathbb{R}^{m\times n}$. Look at the problem

$$f : \mathbb{R}^n \to \mathbb{R}^m, \quad x \mapsto Ax,$$

i.e., the problem of computing $Ax \in \mathbb{R}^m$ from $x \in \mathbb{R}^n$.

Note f is differentiable and $J_f(x) = A$ for all $x \in \mathbb{R}^n$. Hence,

$$\kappa = \frac{\|J_f(x)\|_{(m,n)}\|x\|_{(n)}}{\|f(x)\|_{(m)}} = \frac{\|A\|_{(m,n)}\|x\|_{(n)}}{\|Ax\|_{(m)}},$$

where $\|\cdot\|_{(m,n)}$ matrix norm on $\mathbb{R}^{m\times n}$ induced by the norms $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$ and $\|\cdot\|_{(m)}$ on $\mathbb{R}^m$. If $m = n$, $\|\cdot\|_{(m)} = \|\cdot\|_{(n)}$, and A is invertible, then

$$\kappa = \|A\|_{(n,n)}\frac{\|A^{-1}Ax\|_{(n)}}{\|Ax\|_{(n)}} \le \|A\|_{(n,n)}\|A^{-1}\|_{(n,n)}.$$

This upper bound is attained for certain choices of x.

Condition number of a matrix

Definition (Condition number of a matrix)

Let $A \in \mathbb{R}^{n \times n}$ be invertible and let $\|\cdot\|$ be a norm on $\mathbb{R}^{n \times n}$. Then, we define the **condition number of** A with respect to the norm $\|\cdot\|$ to be

$$\kappa_{\|\cdot\|}(A) := \|A\| \, \|A^{-1}\|.$$

If this quantity is small, we call A well-conditioned. Otherwise, we call A ill-conditioned.

The condition number of a singular square matrix is typically set to ∞.

Theorem (Conditioning of matrix-vector multiplication)

Let $A \in \mathbb{R}^{n\times n}$ be invertible. Consider the vector space $\mathbb{R}^n$ with a chosen norm $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$, and let $\|\cdot\|_{(n,n)}$ denote the matrix norm on $\mathbb{R}^{n\times n}$ induced by the vector norm $\|\cdot\|_{(n)}$. Then, we have the following:

(i) *For the problem $f : \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto Ax$, i.e., the problem of finding $b = Ax$ from $x \in \mathbb{R}^n$, the condition number $\kappa = \kappa(x)$ is given by*

$$\kappa = \|A\|_{(n,n)} \frac{\|x\|_{(n)}}{\|b\|_{(n)}} \leq \kappa_{\|\cdot\|_{(n,n)}}(A). \tag{1}$$

If $\|\cdot\|_{(n)} = \|\cdot\|_2$ is the vector 2-norm (and hence, $\|\cdot\|_{(n,n)} = \|\cdot\|_2$ the spectral norm), we have equality in (1) if x is a multiple of a right singular vector of A corresponding to the smallest singular value σ_n.

(ii) *For the problem $f : \mathbb{R}^n \to \mathbb{R}^n$, $b \mapsto A^{-1}b$, i.e., the problem of finding the solution $x \in \mathbb{R}^n$ to $Ax = b$ from the right-hand side $b \in \mathbb{R}^n$, the condition number $\kappa = \kappa(b)$ is given by*

$$\kappa = \|A^{-1}\|_{(n,n)} \frac{\|b\|_{(n)}}{\|x\|_{(n)}} \leq \kappa_{\|\cdot\|_{(n,n)}}(A). \tag{2}$$

If $\|\cdot\|_{(n)} = \|\cdot\|_2$ is the vector 2-norm (and hence, $\|\cdot\|_{(n,n)} = \|\cdot\|_2$ the spectral norm), we have equality in (2) if b is a multiple of a left singular vector of A corresponding to the largest singular value σ_1.

Let us revisit the problem for $A \in \mathbb{R}^{m\times n}$ being a rectangular matrix with $m \geq n$ and $\mathrm{rk}(A) = n$.

Then, observing that $A^\dagger A = I_n$, i.e., the Moore-Penrose inverse $A^\dagger \in \mathbb{R}^{n\times m}$ is a left-inverse, we find that

$$\kappa = \|A\|_{(m,n)} \frac{\|A^\dagger A x\|_{(n)}}{\|Ax\|_{(m)}} \leq \|A\|_{(m,n)} \|A^\dagger\|_{(n,m)},$$

where $\|\cdot\|_{(m,n)}$ is the induced matrix norm on $\mathbb{R}^{m\times n}$, and $\|\cdot\|_{(n,m)}$ is the induced matrix norm on $\mathbb{R}^{n\times m}$ (induced by the vector norms $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$, $\|\cdot\|_{(m)}$ on $\mathbb{R}^m$). We define the **condition number of** A to be

$$\kappa_{\|\cdot\|_{(m,n)},\|\cdot\|_{(n,m)}}(A) := \|A\|_{(m,n)} \|A^\dagger\|_{(n,m)}.$$

Conditioning of linear systems

Let $X = \mathbb{R}^{n\times n}$ and $Y = \mathbb{R}^n$ with a chosen norm $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$ and induced matrix norm $\|\cdot\|_{(n,n)}$ on $\mathbb{R}^{n\times n}$. Let $b \in \mathbb{R}^n$ be fixed. Consider the problem

$$f : A \mapsto A^{-1}b \in \mathbb{R}^n \quad \text{for } A \in \mathbb{R}^{n\times n} \text{ invertible,}$$

i.e., the problem of finding the solution $x \in \mathbb{R}^n$ to $Ax = b$.

Rk: Although the space of invertible $n \times n$ matrices is not a vector space, we can still study the perturbation behavior of f since a perturbed invertible matrix is still invertible if the perturbation is sufficiently small:

Lemma (Perturbation lemma)

Let $A \in \mathbb{R}^{n\times n}$ be invertible, and let $\|\cdot\|$ be a submultiplicative norm on $\mathbb{R}^{n\times n}$ (i.e., $\|M_1M_2\| \le \|M_1\|\,\|M_2\|$ for any $M_1, M_2 \in \mathbb{R}^{n\times n}$).
Then, for any $\Delta A \in \mathbb{R}^{n\times n}$ with $\|\Delta A\| < \|A^{-1}\|^{-1}$, the perturbed matrix $A + \Delta A \in \mathbb{R}^{n\times n}$ is invertible and there holds

$$\|(A + \Delta A)^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|\Delta A\|\,\|A^{-1}\|}.$$

Lemma (Perturbation lemma)

Let $A \in \mathbb{R}^{n\times n}$ be invertible, and let $\|\cdot\|$ be a submultiplicative norm on $\mathbb{R}^{n\times n}$ (i.e., $\|M_1 M_2\| \le \|M_1\|\,\|M_2\|$ for any $M_1, M_2 \in \mathbb{R}^{n\times n}$).
Then, for any $\Delta A \in \mathbb{R}^{n\times n}$ with $\|\Delta A\| < \|A^{-1}\|^{-1}$, the perturbed matrix $A + \Delta A \in \mathbb{R}^{n\times n}$ is invertible and there holds

$$\|(A+\Delta A)^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|\Delta A\|\,\|A^{-1}\|}.$$

Proof: Use the following fact without pf: For any $X \in \mathbb{R}^{n\times n}$ with $\|X\| < 1$, we have that $I_n - X$ is invertible and there holds $(I_n - X)^{-1} = \sum_{i=0}^{\infty} X^i$ (*Neumann series*) and $\|(I_n - X)^{-1}\| \le \frac{1}{1-\|X\|}$.

Let $A \in \mathbb{R}^{n\times n}$ invertible and $\Delta A \in \mathbb{R}^{n\times n}$ with $\|\Delta A\| < \|A^{-1}\|^{-1}$. Write

$$A + \Delta A = (I_n - X)A \quad \text{with} \quad X := -(\Delta A)A^{-1} \in \mathbb{R}^{n\times n}.$$

Then, $\|X\| = \|(\Delta A)A^{-1}\| \le \|\Delta A\|\,\|A^{-1}\| < 1$. Hence, we find that $A + \Delta A$ is invertible as a product of invertible matrices, and

$$\|(A+\Delta A)^{-1}\| \le \|A^{-1}\|\|(I_n - X)^{-1}\| \le \frac{\|A^{-1}\|}{1-\|X\|} \le \frac{\|A^{-1}\|}{1 - \|\Delta A\|\,\|A^{-1}\|}. \qquad \square$$

Conditioning of linear systems

Recall: $X = \mathbb{R}^{n\times n}$, $Y = \mathbb{R}^n$ with norm $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$ and induced norm $\|\cdot\|_{(n,n)}$ on $\mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$. Let $f : A \mapsto A^{-1}b$ for $A \in \mathbb{R}^{n\times n}$ invertible.

Since induced norm $\|\cdot\|_{(n,n)}$ is submultiplicative, the perturbation lemma can be applied. Let $A \in \mathbb{R}^{n\times n}$ be invertible and let $\Delta A \in \mathbb{R}^{n\times n}$ be such that $\|\Delta A\|_{(n,n)} < \|A^{-1}\|^{-1}_{(n,n)}$. We are interested in the quantity

$$q(\Delta A) := \frac{\|f(A+\Delta A) - f(A)\|_{(n)}}{\|\Delta A\|_{(n,n)}} \frac{\|A\|_{(n,n)}}{\|f(A)\|_{(n)}} = \frac{\|(A+\Delta A)^{-1}b - A^{-1}b\|_{(n)}}{\|\Delta A\|_{(n,n)}} \frac{\|A\|_{(n,n)}}{\|A^{-1}b\|_{(n)}}.$$

Write $(A+\Delta A)^{-1}b = x + \Delta x$ for some $\Delta x \in \mathbb{R}^n$ where $x := A^{-1}b$:

$$Ax = b, \qquad (A+\Delta A)(x+\Delta x) = b.$$

$\Longrightarrow (\Delta A)x + (A+\Delta A)\Delta x = 0$, i.e., $\Delta x = -(A+\Delta A)^{-1}(\Delta A)x$, and thus,

$$(A+\Delta A)^{-1}b - A^{-1}b = \Delta x = -(A+\Delta A)^{-1}(\Delta A)A^{-1}b.$$

We find that

$$
\begin{aligned}
q(\Delta A) &= \frac{\|(A+\Delta A)^{-1}b - A^{-1}b\|_{(n)}}{\|\Delta A\|_{(n,n)}} \frac{\|A\|_{(n,n)}}{\|A^{-1}b\|_{(n)}} \\
&= \frac{\|-(A+\Delta A)^{-1}(\Delta A)A^{-1}b\|_{(n)}}{\|\Delta A\|_{(n,n)}} \frac{\|A\|_{(n,n)}}{\|A^{-1}b\|_{(n)}} \\
&\leq \|(A+\Delta A)^{-1}\|_{(n,n)}\|A\|_{(n,n)} \\
&\leq \frac{\|A\|_{(n,n)}\|A^{-1}\|_{(n,n)}}{1-\|\Delta A\|_{(n,n)}\|A^{-1}\|_{(n,n)}} = \frac{\kappa_{\|\cdot\|_{(n,n)}}(A)}{1-\frac{\|\Delta A\|_{(n,n)}}{\|A\|_{(n,n)}}\kappa_{\|\cdot\|_{(n,n)}}(A)},
\end{aligned}
$$

and it follows that the condition number for the problem f at the matrix A is bounded by the condition number of the matrix A:

$$
\kappa = \lim_{\delta\to 0} \sup_{\substack{\Delta A\in\mathbb{R}^{n\times n} \\ 0<\|\Delta A\|_{(n,n)}\leq\delta}} q(\Delta A) \leq \kappa_{\|\cdot\|_{(n,n)}}(A).
$$

It can actually be shown that there holds equality in the above estimate (we omit the proof) and we arrive at the following theorem:

Conditioning of linear systems

Theorem (Conditioning of linear systems)

Consider the vector space $\mathbb{R}^n$ *with a chosen norm* $\|\cdot\|_{(n)}$ *on* $\mathbb{R}^n$*, and let* $\|\cdot\|_{(n,n)}$ *denote the matrix norm on* $\mathbb{R}^{n\times n}$ *induced by* $\|\cdot\|_{(n)}$*. Then, for a fixed* $b \in \mathbb{R}^n$*, the condition number for the problem of finding the solution* $x \in \mathbb{R}^n$ *of* $Ax = b$ *from* $A \in \{M \in \mathbb{R}^{n\times n} : M \text{ invertible}\}$ *is given by*

$$\kappa = \kappa_{\|\cdot\|_{(n,n)}}(A).$$

Conditioning of least squares problems

Given $A \in \mathbb{R}^{m\times n}$, $m \geq n$, $\mathrm{rk}(A) = n$, and $b \in \mathbb{R}^m$, consider LS problem

$$\text{Minimize } \|Av - b\|_2 \text{ over } v \in \mathbb{R}^n.$$

Recall that in this situation we have

- $x = A^\dagger b$ is the unique solution to the least squares problem, i.e., the unique vector $x \in \mathbb{R}^n$ satisfying $\|Ax - b\|_2 = \inf_{v\in\mathbb{R}^n} \|Av - b\|_2$,
- $y = Ax = AA^\dagger b$ is the unique vector $y \in \mathscr{R}(A)$ satisfying $\|y - b\|_2 = \inf_{w\in\mathscr{R}(A)} \|w - b\|_2$.

We consider the following mathematical problems:

(i) obtain y from b for fixed A, i.e., $f_{b\mapsto y} : \mathbb{R}^m \to \mathbb{R}^m$, $b \mapsto AA^\dagger b$,

(ii) obtain x from b for fixed A, i.e., $f_{b\mapsto x} : \mathbb{R}^m \to \mathbb{R}^n$, $b \mapsto A^\dagger b$,

(iii) obtain y from A for fixed b, i.e., $f_{A\mapsto y} : A \mapsto AA^\dagger b \in \mathbb{R}^m$ for $A \in \mathbb{R}^{m\times n}$, $\mathrm{rk}(A) = n$,

(iv) obtain x from A for fixed b, i.e., $f_{A\mapsto x} : A \mapsto A^\dagger b \in \mathbb{R}^n$ for $A \in \mathbb{R}^{m\times n}$, $\mathrm{rk}(A) = n$,

and we consider the 2-norm on $\mathbb{R}^m$ and $\mathbb{R}^n$, and the spectral norm on $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{n\times m}$.

Theorem (Conditioning of least squares problems)

In this situation, there holds

$$\kappa_{b\mapsto y} = \frac{1}{\cos(\theta)},\ \kappa_{b\mapsto x} = \frac{\kappa(A)}{\eta\cos(\theta)},\ \kappa_{A\mapsto y} \leq \frac{\kappa(A)}{\cos(\theta)},\ \kappa_{A\mapsto x} \leq \kappa(A) + \frac{(\kappa(A))^2\tan(\theta)}{\eta},$$

where $\kappa_{i\mapsto j}$ $(i \in \{b, A\},\ j \in \{x, y\})$ *condition number for* $f_{i\mapsto j}$, *and*

$$\kappa(A) := \|A\|_2\|A^\dagger\|_2 \geq 1, \theta := \cos^{-1}\left(\frac{\|AA^\dagger b\|_2}{\|b\|_2}\right) \in \left[0, \frac{\pi}{2}\right], \eta := \frac{\|A\|_2\|A^\dagger b\|_2}{\|AA^\dagger b\|_2} \in [1, \kappa(A)]$$

Observations:

- For $A \in \mathbb{R}^{m\times n}$, $m \geq n$, $\mathrm{rk}(A) = n$, the condition number in the spectral norm is given by $\kappa(A) = \|A\|_2\|A^\dagger\|_2 = \frac{\sigma_1}{\sigma_n} \in [1, \infty)$.
- θ is a measure for the closeness of the projection $AA^\dagger b$ to b.
- If $m = n$, we have $A^\dagger = A^{-1}$ and hence $\theta = 0$. In particular, we find $\kappa_{b\mapsto x} = \frac{\kappa(A)}{\eta} = \frac{\|A^{-1}\|_2\|b\|_2}{\|A^{-1}b\|_2}$ and $\kappa_{A\mapsto x} \leq \kappa(A) = \|A\|_2\|A^{-1}\|_2$, i.e., we recover the previous results on the conditioning of linear systems.

Proof of (i)

Let $A \in \mathbb{R}^{m\times n}$, $m \geq n$, $\mathrm{rk}(A) = n$ be fixed, and consider

$$f_{b\mapsto y} : \mathbb{R}^m \to \mathbb{R}^m, \qquad b \mapsto AA^\dagger b.$$

Recall from PS3: $P := AA^\dagger$ is the orthogonal projector onto $\mathscr{R}(A)$.
Note $P \neq 0_{m\times m}$ as $A \neq 0_{m\times n}$.

$\Longrightarrow \|AA^\dagger\|_2 = 1$.

We find that the condition number $\kappa_{b\mapsto y} = \kappa_{b\mapsto y}(b)$ of $f_{b\mapsto y}$ is given by

$$\kappa_{b\mapsto y} = \frac{\|J_{f_{b\mapsto y}}(b)\|_2\|b\|_2}{\|f_{b\mapsto y}(b)\|_2} = \frac{\|AA^\dagger\|_2\|b\|_2}{\|AA^\dagger b\|_2} = \frac{\|b\|_2}{\|AA^\dagger b\|_2} = \frac{1}{\cos(\theta)}.$$

Proof of (ii)

Let $A \in \mathbb{R}^{m\times n}$, $m \geq n$, with $\mathrm{rk}(A) = n$ be fixed, and consider the problem

$$f_{b\mapsto x} : \mathbb{R}^m \to \mathbb{R}^n, \qquad b \mapsto A^\dagger b.$$

Then, the condition number $\kappa_{b\mapsto x} = \kappa_{b\mapsto x}(b)$ of $f_{b\mapsto x}$ is given by

$$\begin{aligned}
\kappa_{b\mapsto x} = \frac{\|J_{f_{b\mapsto x}}(b)\|_2\|b\|_2}{\|f_{b\mapsto x}(b)\|_2} &= \frac{\|A^\dagger\|_2\|b\|_2}{\|A^\dagger b\|_2} \\
&= \|A\|_2\|A^\dagger\|_2 \frac{\|AA^\dagger b\|_2}{\|A\|_2\|A^\dagger b\|_2}\frac{\|b\|_2}{\|AA^\dagger b\|_2} \\
&= \frac{\kappa(A)}{\eta\cos(\theta)}.
\end{aligned}$$

Proof of (iii),(iv) omitted. □

5.2 Floating point numbers and floating point arithmetic

How are real numbers represented on a computer?

Note: Computers use a finite number of bits to represent a number.

$\Longrightarrow$ there is

- a largest represented number (rn): $x^+_{\max} > 0$,
- a smallest rn: $x^-_{\min} < 0$,
- a smallest positive rn: $x^+_{\min} > 0$,
- a largest negative rn: $x^-_{\max} < 0$,

i.e., the set of all rn's is a finite subset of $[x^-_{\min}, x^-_{\max}] \cup \{0\} \cup [x^+_{\min}, x^+_{\max}]$.

$\Longrightarrow$ there must be gaps between represented numbers.

Floating Point System (FPS)

Definition (Floating point system)

Given

- $\beta \in \mathbb{N}$ with $\beta \geq 2$ (**base**, usually taken to be 2),
- $t \in \mathbb{N}$ (**precision**),
- $e_{\min}, e_{\max} \in \mathbb{Z}$ (**minimal/maximal exponent**),

we define the **floating point system** $F = F(\beta, t, e_{\min}, e_{\max}) \subseteq \mathbb{R}$ to be the set of real numbers that can be written as

$$x = (-1)^s \cdot (m_1 \beta^{-1} + \cdots + m_t \beta^{-t}) \cdot \beta^e =: (-1)^s \cdot [0.m_1 \ldots m_t]_\beta \cdot \beta^e$$

for some $m_1, \ldots, m_t \in \{0, 1, \ldots, \beta - 1\}$, $e \in \mathbb{Z} \cap [e_{\min}, e_{\max}]$, $s \in \{0, 1\}$.

We call the number $[0.m_1 \ldots m_t]_\beta \in [0, 1)$ the **mantissa** of x, and the number $e \in \mathbb{Z}$ the **exponent** of x.

Requiring $m_1 \neq 0$ if $x \neq 0$ and $m_1 = 0$ if $x = 0$, representation is unique.

Largest/smallest rn's

Recall: $x = (-1)^s \cdot (m_1\beta^{-1} + \cdots + m_t\beta^{-t}) \cdot \beta^e$ for some $m_1, \ldots, m_t \in \{0, 1, \ldots, \beta - 1\}$, $e \in \mathbb{Z} \cap [e_{\min}, e_{\max}]$, $s \in \{0, 1\}$.

In a FPS $F = F(\beta, t, e_{\min}, e_{\max})$, the largest rn is

$$x^+_{\max} = (\beta - 1)\left(\sum_{i=1}^{t} \beta^{-i}\right) \cdot \beta^{e_{\max}} = (1 - \beta^{-t})\beta^{e_{\max}},$$

the smallest rn is $x^-_{\min} = -(1 - \beta^{-t})\beta^{e_{\max}}$, the smallest positive rn is

$$x^+_{\min} = \beta^{-1} \cdot \beta^{e_{\min}} = \beta^{e_{\min}-1},$$

and the largest negative rn is $x^-_{\max} = -\beta^{e_{\min}-1}$. Therefore, we have

$$F \subseteq [-(1 - \beta^{-t})\beta^{e_{\max}}, -\beta^{e_{\min}-1}] \cup \{0\} \cup [\beta^{e_{\min}-1}, (1 - \beta^{-t})\beta^{e_{\max}}].$$

IEEE double precision arithmetic

One uses $\beta = 2$, $t = 53$, and the rn's are of the form

$$\begin{aligned} x &= (-1)^s \cdot (m_1 2^{-1} + \cdots + m_{53} 2^{-53}) \cdot 2^{(c_{10}2^{10} + \cdots + c_0 2^0) - 1022} \\ &= (-1)^s \cdot [0.m_1 \ldots m_{53}]_2 \cdot 2^{[c_{10} \ldots c_0]_2 - 1022} \end{aligned} \tag{3}$$

with $s, c_i, m_i \in \{0, 1\}$, **biased exponent** $[c_{10} \ldots c_0]_2 \in \{1, \ldots, 2046\}$, and $m_1 = 1$. The excluded numbers $[c_{10} \ldots c_0]_2 \in \{0, 2047\}$ are used for representing 0 and "NaN". The number x from (3) is equivalent to

$$\begin{aligned} x &= (-1)^s \cdot (m_1 + m_2 2^{-1} \cdots + m_{53} 2^{-52}) \cdot 2^{(c_{10}2^{10} + \cdots + c_0 2^0) - 1023} \\ &= (-1)^s \cdot (1 + [0.m_2 \ldots m_{53}]_2) \cdot 2^{[c_{10} \ldots c_0]_2 - 1023} \\ &= (-1)^s \cdot [1.m_2 \ldots m_{53}]_2 \cdot 2^{[c_{10} \ldots c_0]_2 - 1023} \end{aligned}$$

and is stored as the binary number

$$| \underbrace{s}_{1 \text{ bit}} | \underbrace{c_{10}|c_9|c_8|\ldots|c_2|c_1|c_0}_{11 \text{ bits}} | \underbrace{m_2|m_3|m_4|\ldots|m_{51}|m_{52}|m_{53}}_{52 \text{ bits}} |.$$

Note $x^+_{\max} = (1 - 2^{-53})2^{1024} \approx 1.8 \cdot 10^{308}$, $x^+_{\min} = 2^{-1022} \approx 2.2 \cdot 10^{-308}$, $x^-_{\min} = -(1 - 2^{-53})2^{1024} \approx -1.8 \cdot 10^{308}$, $x^-_{\max} = -2^{-1022} \approx -2.2 \cdot 10^{-308}$

Observe that, in IEEE double precision arithmetic, the rn's

- in the interval $[1, 2]$ are $\{1 + j \cdot 2^{-52} \,|\, j \in \{0, 1, \ldots, 2^{52}\}\}$,
- in the interval $[2, 4]$ are $\{2 + j \cdot 2^{-51} \,|\, j \in \{0, 1, \ldots, 2^{52}\}\}$,
- in the interval $[2^k, 2^{k+1}]$ are $\{2^k + j \cdot 2^{k-52} \,|\, j \in \{0, 1, \ldots, 2^{52}\}\}$.

$\Longrightarrow$ distance between adjacent numbers in relative sense at most $2^{-52} \approx 2.2 \cdot 10^{-16}$.

(Note that the rn's in $[2^{52}, 2^{53}]$ are precisely the integers $\mathbb{N} \cap [2^{52}, 2^{53}]$).

A measure for the resolution of F: the number $\varepsilon_{\text{machine}}$.

Definition (machine epsilon)

To a FPS $F = F(\beta, t, e_{\min}, e_{\max})$, define the **machine epsilon**

$$\varepsilon_{\text{machine}} = \frac{1}{2}\beta^{1-t}.$$

The machine epsilon in IEEE double precision arithmetic is given by

$$\varepsilon_{\text{machine}} = \frac{2^{1-53}}{2} = 2^{-53} \approx 1.1 \cdot 10^{-16}.$$

(half the distance between 1 and next larger rn)

Rounding

For any $x \in [x^-_{\min}, x^-_{\max}] \cup [x^+_{\min}, x^+_{\max}]$ there exists a rn $x' \in F$ satisfying

$$\frac{|x - x'|}{|x|} \leq \varepsilon_{\text{machine}}, \tag{4}$$

i.e., the distance between x and x' in a relative sense is at most $\varepsilon_{\text{machine}}$.

Define a **rounding operator** $\text{fl} : [x^-_{\min}, x^-_{\max}] \cup \{0\} \cup [x^+_{\min}, x^+_{\max}] \to F$ with the property

$$|x - \text{fl}(x)| = \inf_{y \in F} |x - y|$$

for all $x \in [x^-_{\min}, x^-_{\max}] \cup \{0\} \cup [x^+_{\min}, x^+_{\max}]$. Then, $x' = \text{fl}(x)$ satisfies (4).
$\Longrightarrow \forall x \in [x^-_{\min}, x^-_{\max}] \cup \{0\} \cup [x^+_{\min}, x^+_{\max}]\ \exists\, \varepsilon \in [-\varepsilon_{\text{machine}}, \varepsilon_{\text{machine}}]$:

$$\text{fl}(x) = x(1 + \varepsilon).$$

Floating Point Arithmetic

Floating point operations: analogue of elementary operations $(+, -, \times, /)$ for two numbers of a FPS.

Definition (Floating point operations)

Let F be a FPS. Define the **floating point operations** $\oplus, \ominus, \otimes, \oslash$ on F by

$$x \circledast y := \mathrm{fl}(x * y), \qquad (x, y \in F)$$

for $\circledast \in \{\oplus, \ominus, \otimes, \oslash\}$.

In view of $\mathrm{fl}(x) = x(1+\varepsilon)$ for some ε with $|\varepsilon| \leq \varepsilon_{\text{machine}}$:

Theorem (Fundamental axiom of floating point arithmetic)

Let F be a FPS and $\circledast \in \{\oplus, \ominus, \otimes, \oslash\}$. Then, for all $x, y \in F$ ($y \neq 0$ if $\circledast = \oslash$) there exists $\varepsilon \in [-\varepsilon_{\text{machine}}, \varepsilon_{\text{machine}}]$ such that

$$x \circledast y = (x * y)(1+\varepsilon).$$

In particular, $|x \circledast y - x * y| \leq \varepsilon_{\text{machine}} |x * y|$ for all $x, y \in F$.

5.3 Stability of numerical algorithms

What is an algorithm?

Simplification: From now on, we consider an idealized FPS $F = F(\beta, t)$ ignoring overflow and underflow (all integer exponents $e \in \mathbb{Z}$ allowed).

Question: What is an algorithm for "solving" a mathematical problem $f : X \to Y$ (with X, Y normed vector spaces)?

Suppose we have a computer with FPS satisfying the fundamental axiom. We regard an **algorithm** for the problem as a map

$$\tilde{f} : X \to Y,$$

where for $x \in X$, $\tilde{f}(x)$ is defined as follows:

1. Round x to a floating point number $\mathrm{fl}(x)$.
2. Run the (fixed) implementation of the algorithm with input $\mathrm{fl}(x)$.
3. Output is defined as $\tilde{f}(x)$ (collection of floating point numbers in Y).

Landau symbol $\mathcal{O}$

Definition (Landau symbol $\mathcal{O}$)

For real-valued functions $u = u(t)$ and $v = v(t)$ of a variable $t \in \mathbb{R}_{>0}$:

$$u(t) = \mathcal{O}(v(t)) \text{ as } t \searrow 0 \iff \exists t_0, C > 0 : |u(t)| \leq Cv(t) \quad \forall t \in (0, t_0),$$

$$u(t) = \mathcal{O}(v(t)) \text{ as } t \to \infty \iff \exists t_0, C > 0 : |u(t)| \leq Cv(t) \quad \forall t \in (t_0, \infty).$$

Examples:

- $u(t) := 2t^2 + 9t^3 = \mathcal{O}(t^2)$ as $t \searrow 0$. Pf:

$$\frac{|2t^2 + 9t^3|}{t^2} = 2 + 9t \leq 11 \quad \forall t \in (0, 1).$$

- $u(t) := 3\log(t) + 4t - t^3 = \mathcal{O}(t^3)$ as $t \to \infty$. Pf:

$$\frac{|3\log(t) + 4t - t^3|}{t^3} = \left|3\frac{\log(t)}{t^3} + 4\frac{1}{t^2} - 1\right| \longrightarrow 1 \quad \text{as} \quad t \to \infty.$$

So, $\exists t_0 > 0$ s.t. $\frac{|u(t)|}{t^3} \leq 2$ for all $t > t_0$.

Accuracy and Stability

Definition (Accuracy, stability)

Let X and Y be normed vector spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. Let $f : X \to Y$ be a problem and $\tilde{f} : X \to Y$ be an algorithm for f.

(i) $\tilde{f}$ is called **accurate** iff for each $x \in X$ there holds

$$\frac{\|\tilde{f}(x) - f(x)\|_Y}{\|f(x)\|_Y} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

(ii) $\tilde{f}$ is called **stable** iff for each $x \in X$ there holds

$$\frac{\|\tilde{f}(x) - f(\tilde{x})\|_Y}{\|f(\tilde{x})\|_Y} = \mathcal{O}(\varepsilon_{\text{machine}}) \text{ for some } \tilde{x} \in X \text{ with } \frac{\|\tilde{x} - x\|_X}{\|x\|_X} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

Statements of the form $\frac{\|p(x,\varepsilon_{\text{machine}})\|}{\|q(x,\varepsilon_{\text{machine}})\|} = \mathcal{O}(\varepsilon_{\text{machine}})$ are meant in the sense

$$\frac{\|p(x,\varepsilon_{\text{machine}})\|}{\|q(x,\varepsilon_{\text{machine}})\|} = \mathcal{O}(\varepsilon_{\text{machine}}) \quad \text{as} \quad \varepsilon_{\text{machine}} \searrow 0, \quad \text{uniformly in } x, \text{ i.e.,}$$

$$\exists\, \varepsilon_0, C > 0 : \ \|p(x,\varepsilon_{\text{machine}})\| \le C\varepsilon_{\text{machine}}\|q(x,\varepsilon_{\text{machine}})\| \ \forall \varepsilon_{\text{machine}} \in (0,\varepsilon_0),\ x \in X.$$

We dream of accuracy, but ...

Accuracy: $\frac{\|\tilde{f}(x)-f(x)\|_Y}{\|f(x)\|_Y} = \mathcal{O}(\varepsilon_{\text{machine}})$.

Stability:
$\frac{\|\tilde{f}(x)-f(\tilde{x})\|_Y}{\|f(\tilde{x})\|_Y} = \mathcal{O}(\varepsilon_{\text{machine}})$ for some $\tilde{x} \in X$ with $\frac{\|\tilde{x}-x\|_X}{\|x\|_X} = \mathcal{O}(\varepsilon_{\text{machine}})$.

If f ill-conditioned, there is little hope to construct an accurate $\tilde{f}$:
Even if the only error would stem from rounding the input data (and say everything else is performed exactly), this small perturbation can already lead to large changes in the result.

$\Longrightarrow$ Appropriate goal in constructing algorithms is stability:

A stable algorithm gives the almost right answer to an almost right question.

Backward stability: A stronger condition than stability

Definition (Backward stability)

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces. Let $f : X \to Y$ be a problem and $\tilde{f} : X \to Y$ be an algorithm for f.
Then, $\tilde{f}$ is called **backward stable** iff for each $x \in X$ there holds

$$\tilde{f}(x) = f(\tilde{x}) \quad \text{for some } \tilde{x} \in X \text{ with } \frac{\|\tilde{x} - x\|_X}{\|x\|_X} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

Any backward stable algorithm is stable.

A backward stable algorithm gives the exact answer to an almost right question.

Illustration

Some results

Theorem (Independence of norm)

If X, Y are finite-dimensional, the definitions of accuracy, stability, and backward stability are independent of the choice of norms in X and Y in the sense that the corresponding conditions either all hold or fail independently of the choice of norms.

Theorem (Accuracy of backward stable algorithms)

Let X and Y be normed vector spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. Consider a problem $f : X \to Y$ with condition number κ, and a backward stable algorithm $\tilde{f} : X \to Y$ for f. Then, there holds

$$\frac{\|\tilde{f}(x) - f(x)\|_Y}{\|f(x)\|_Y} = \mathcal{O}(\kappa(x)\,\varepsilon_{\text{machine}}).$$

In particular, if $\kappa(x) = \mathcal{O}(1)$, then $\tilde{f}$ is accurate.

Example 1: Stability of floating point operation $\oplus$

The floating point operations $\oplus, \ominus, \otimes, \oslash$ are all backward stable. We prove this for $\oplus$ and leave the remaining operations as an exercise.

Let us consider the problem

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x_1, x_2) := x_1 + x_2,$$

and the algorithm

$$\tilde{f} : \mathbb{R}^2 \to \mathbb{R}, \quad \tilde{f}(x_1, x_2) := \mathrm{fl}(x_1) \oplus \mathrm{fl}(x_2).$$

Choose $\|\cdot\|_1$ as norm on $\mathbb{R}^2$ and $|\cdot|$ as norm on $\mathbb{R}$. Let $x = (x_1, x_2)^{\mathrm{T}} \in \mathbb{R}^2$.

Then, $\mathrm{fl}(x_1) = x_1(1+\varepsilon_1)$ and $\mathrm{fl}(x_2) = x_2(1+\varepsilon_2)$ with $|\varepsilon_1|, |\varepsilon_2| \leq \varepsilon_{\text{machine}}$, and we have $\mathrm{fl}(x_1) \oplus \mathrm{fl}(x_2) = (\mathrm{fl}(x_1) + \mathrm{fl}(x_2))(1+\varepsilon_3)$ with $|\varepsilon_3| \leq \varepsilon_{\text{machine}}$.

Recall: $\mathrm{fl}(x_1) = x_1(1+\varepsilon_1)$ and $\mathrm{fl}(x_2) = x_2(1+\varepsilon_2)$ with $|\varepsilon_1|, |\varepsilon_2| \leq \varepsilon_{\text{machine}}$, and we have $\mathrm{fl}(x_1) \oplus \mathrm{fl}(x_2) = (\mathrm{fl}(x_1) + \mathrm{fl}(x_2))(1+\varepsilon_3)$ with $|\varepsilon_3| \leq \varepsilon_{\text{machine}}$.

Therefore, we find

$$\begin{aligned}\tilde{f}(x) = \mathrm{fl}(x_1) \oplus \mathrm{fl}(x_2) &= (\mathrm{fl}(x_1) + \mathrm{fl}(x_2))(1+\varepsilon_3) \\ &= (x_1(1+\varepsilon_1) + x_2(1+\varepsilon_2))(1+\varepsilon_3) \\ &= x_1(1+\varepsilon_1)(1+\varepsilon_3) + x_2(1+\varepsilon_2)(1+\varepsilon_3) \\ &= \tilde{x}_1 + \tilde{x}_2 = f(\tilde{x})\end{aligned}$$

with $\tilde{x}_1 = x_1(1+\varepsilon_1)(1+\varepsilon_3)$, $\tilde{x}_2 = x_2(1+\varepsilon_2)(1+\varepsilon_3)$ and $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)^{\mathrm{T}}$. We have

$$|\tilde{x}_1 - x_1| = |\varepsilon_1 + \varepsilon_3 + \varepsilon_1\varepsilon_3|\,|x_1| \leq (|\varepsilon_1| + |\varepsilon_3| + |\varepsilon_1|\,|\varepsilon_3|)|x_1| \leq C(\varepsilon_{\text{machine}})|x_1|,$$
$$|\tilde{x}_2 - x_2| = |\varepsilon_2 + \varepsilon_3 + \varepsilon_2\varepsilon_3|\,|x_2| \leq (|\varepsilon_2| + |\varepsilon_3| + |\varepsilon_2|\,|\varepsilon_3|)|x_2| \leq C(\varepsilon_{\text{machine}})|x_2|,$$

with $C(\varepsilon_{\text{machine}}) := 2\varepsilon_{\text{machine}} + \varepsilon_{\text{machine}}^2$, and hence,

$$\|\tilde{x} - x\|_1 = |\tilde{x}_1 - x_1| + |\tilde{x}_2 - x_2| \leq C(\varepsilon_{\text{machine}})(|x_1| + |x_2|) = C(\varepsilon_{\text{machine}})\|x\|_1.$$

Since $C(\varepsilon_{\text{machine}}) = 2\varepsilon_{\text{machine}} + \varepsilon_{\text{machine}}^2 = \mathcal{O}(\varepsilon_{\text{machine}})$, it follows that $\tilde{f}$ is backward stable.

Example 2: Stability of adding 1

Let us consider the problem

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) := x + 1,$$

and the algorithm

$$\tilde{f} : \mathbb{R} \to \mathbb{R}, \qquad \tilde{f}(x) := \mathrm{fl}(x) \oplus 1.$$

Then, $\tilde{f}$ is stable. Proof: We choose the absolute value $|\cdot|$ as norm on $\mathbb{R}$. For $x \in \mathbb{R}$ set $\tilde{x} = \mathrm{fl}(x)$ so that we have $|\tilde{x} - x| \leq \varepsilon_{\text{machine}}|x|$ and

$$\begin{aligned} |\tilde{f}(x) - f(\tilde{x})| = |(\mathrm{fl}(x) \oplus 1) - (\tilde{x} + 1)| &= |(\tilde{x} \oplus 1) - (\tilde{x} + 1)| \\ &\leq \varepsilon_{\text{machine}}|\tilde{x} + 1| = \varepsilon_{\text{machine}}|f(\tilde{x})|. \end{aligned}$$

It follows that $\tilde{f}$ is stable. Exercise: $\tilde{f}$ is not backward stable.

Example 3: Stability of computing inner/outer products

(i) Inner product: Consider $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $f(x, y) := x^{\mathrm{T}} y$. Then, the algorithm $\tilde{f} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$\tilde{f}(x, y) := [[[(\mathrm{fl}(x_1) \otimes \mathrm{fl}(y_1)) \oplus (\mathrm{fl}(x_2) \otimes \mathrm{fl}(y_2))] \oplus (\mathrm{fl}(x_3) \otimes \mathrm{fl}(y_3))] \oplus \dots] \oplus (\mathrm{fl}(x_n) \otimes \mathrm{fl}(y_n))$$

is backward stable.

(ii) Outer product: Consider $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^{m \times n}$, $f(x, y) := x y^{\mathrm{T}}$. Then, the algorithm $\tilde{f} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^{m \times n}$ given by

$$\tilde{f}(x, y) := \begin{pmatrix} \mathrm{fl}(x_1) \otimes \mathrm{fl}(y_1) & \cdots & \mathrm{fl}(x_1) \otimes \mathrm{fl}(y_n) \\ \vdots & & \vdots \\ \mathrm{fl}(x_m) \otimes \mathrm{fl}(y_1) & \cdots & \mathrm{fl}(x_m) \otimes \mathrm{fl}(y_n) \end{pmatrix}$$

is stable, but not backward stable.

Example 4: (In)stability of computing eigenvalues

Consider the following algorithm for computing eigenvalues of $A \in \mathbb{R}^{n \times n}$:

1. First, find the coefficients of the characteristic polynomial (i.e., $\lambda \mapsto \det(\lambda I_n - A)$).
2. Find the roots of the obtained polynomial.

This algorithm is unstable (hence, this is not used in practice).

Note that for e.g. $A = I_2 \in \mathbb{R}^{2 \times 2}$ we have the characteristic polynomial $t \mapsto t^2 - 2t + 1$.

Computing the coefficients of the characteristic polynomial, we have errors of order $\mathcal{O}(\varepsilon_{\text{machine}})$, leading to errors in the roots of order $\mathcal{O}(\sqrt{\varepsilon_{\text{machine}}})$.
In IEEE double precision arithmetic, this means loss of 8 digits of accuracy.

5.4 Stability of solution algorithms for linear systems

Stability of solving $Ax = b$ via QR (using Householder)

Given: $A \in \mathbb{R}^{n\times n}$ invertible, $b \in \mathbb{R}^n$. Find $x \in \mathbb{R}^n$ s.t. $Ax = b$.

1) Use Householder to obtain factor $R \in \mathbb{R}^{n\times n}$ of a QR factn $A = QR$, and reflection vectors $v_1, \ldots, v_n \in \mathbb{R}^n$ (Q is not explicitly formed).
2) Compute $y := Q^{\mathrm{T}} b \in \mathbb{R}^n$ from the vectors $v_1, \ldots, v_n$ and b.
3) Solve the upper-triangular system $Rx = y$ by backward substitution.

Theorem

The above algorithm is backward stable in the sense that

$$(A + \Delta A)\tilde{x} = b \quad \text{for some } \Delta A \in \mathbb{R}^{n\times n} \text{ with } \frac{\|\Delta A\|}{\|A\|} = \mathcal{O}(\varepsilon_{\text{machine}})$$

for all norms $\|\cdot\|$ on $\mathbb{R}^{n\times n}$, where $\tilde{x} \in \mathbb{R}^n$ is the computed soln. Further,

$$\frac{\|\tilde{x} - A^{-1}b\|_{(n)}}{\|A^{-1}b\|_{(n)}} = \mathcal{O}(\kappa_{\|\cdot\|_{(n,n)}}(A)\, \varepsilon_{\text{machine}})$$

for any norm $\|\cdot\|_{(n)}$ on $\mathbb{R}^n$ with corresponding induced norm $\|\cdot\|_{(n,n)}$.

Backward stability of QR via Householder

Theorem (Backward stability of QR via Householder)

Suppose we apply Householder to an invertible matrix $A \in \mathbb{R}^{n\times n}$, leading to outputs $\tilde{R} \in \mathbb{R}^{n\times n}$ and $\tilde{v}_1, \ldots, \tilde{v}_n \in \mathbb{R}^n$ (the computed factor R and reflection vectors v_i in floating point computation). Writing $\tilde{Q} := \tilde{Q}_1\tilde{Q}_2 \ldots \tilde{Q}_n$ with $\tilde{Q}_i$ denoting the orthogonal matrix from Section (Householder) corresponding to the reflection vector $\tilde{v}_i$, there holds

$$\tilde{Q}\tilde{R} = A + \Delta A \quad \text{for some } \Delta A \in \mathbb{R}^{n\times n} \text{ with } \frac{\|\Delta A\|}{\|A\|} = \mathcal{O}(\varepsilon_{\text{machine}})$$

for all matrix norms $\|\cdot\|$ on $\mathbb{R}^{n\times n}$.

Stability of Gaussian elimination

Theorem

(i) *Gauß without pivoting: Suppose a LU factorization $A = LU$ of an invertible matrix $A \in \mathbb{R}^{n\times n}$, for which there exists a LU factorization, is computed by Gauß. Then,*

$$\tilde{L}\tilde{U} = A + \Delta A \quad \text{for some } \Delta A \in \mathbb{R}^{n\times n} \text{ with } \frac{\|\Delta A\|}{\|L\|\|U\|} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

(ii) *Gauß with partial pivoting: Suppose a PA=LU factorization of an invertible matrix $A \in \mathbb{R}^{n\times n}$ is computed by Gauß with partial pivoting. Then,*

$$\tilde{L}\tilde{U} = \tilde{P}A + \Delta A \quad \text{for some } \Delta A \in \mathbb{R}^{n\times n} \text{ with } \frac{\|\Delta A\|}{\|A\|} = \mathcal{O}(\rho\,\varepsilon_{\text{machine}})$$

where ρ denotes the **growth factor** *of A defined by*

$$\rho := \frac{\max_{i,j\in\{1,\dots,n\}}|u_{ij}|}{\max_{i,j\in\{1,\dots,n\}}|a_{ij}|}.$$

If $|l_{ij}| < 1$ for all $i > j$, then $\tilde{P} = P$ for $\varepsilon_{\text{machine}}$ sufficiently small.

Stability of Gauß without pivoting

Recall $\tilde{L}\tilde{U} = A + \Delta A$ for some $\Delta A \in \mathbb{R}^{n\times n}$ with $\frac{\|\Delta A\|}{\|L\|\|U\|} = \mathcal{O}(\varepsilon_{\text{machine}})$.

$\Longrightarrow$ backward stability if $\|L\|\|U\| = \mathcal{O}(\|A\|)$. Otherwise, backward instability is to be expected.

It is known that both L and U can be unboundedly large and that Gaussian elimination without pivoting is unstable, and hence, should not be used in general. We give a simple example illustrating the problem:

Consider $A := \begin{pmatrix} 10^{-20} & 1 \\ 1 & 1 \end{pmatrix}$. Gauß performed exactly gives

$$A = LU, \qquad L := \begin{pmatrix} 1 & 0 \\ 10^{20} & 1 \end{pmatrix}, \quad U := \begin{pmatrix} 10^{-20} & 1 \\ 0 & 1 - 10^{20} \end{pmatrix}.$$

In IEEE double precision arithmetic, the computed result would be

$$\tilde{L} := \begin{pmatrix} 1 & 0 \\ 10^{20} & 1 \end{pmatrix}, \quad \tilde{U} := \begin{pmatrix} 10^{-20} & 1 \\ 0 & -10^{20} \end{pmatrix} \qquad \Longrightarrow \tilde{L}\tilde{U} = \begin{pmatrix} 10^{-20} & 1 \\ 1 & 0 \end{pmatrix}.$$

Considering $Ax = b := (1, 0)^{\mathrm{T}}$ with exact solution $x \approx (-1, 1)^{\mathrm{T}}$, we find from $\tilde{L}\tilde{U}\tilde{x} = b$ that $\tilde{x} = (0, 1)^{\mathrm{T}}$, terrible!

Stability of Gauß with partial pivoting

Recall

$$\tilde{L}\tilde{U} = \tilde{P}A + \Delta A \quad \text{for some } \Delta A \in \mathbb{R}^{n\times n} \text{ with } \frac{\|\Delta A\|}{\|A\|} = \mathcal{O}(\rho\, \varepsilon_{\text{machine}})$$

where

$$\rho := \frac{\max_{i,j\in\{1,\dots,n\}} |u_{ij}|}{\max_{i,j\in\{1,\dots,n\}} |a_{ij}|}.$$

Problem sheets: $\rho \leq 2^{n-1}$ and this is sharp.

A growth factor of 2^n means a loss of around n bits of precision, which is a huge problem for high-dimensional problems (as they arise in practice).

Still, according to our definition, Gaussian elimination with partial pivoting is backward stable (as dependence of the constant on the dimension is allowed). However, we should rather think of it as stable for most problems, but very unstable for certain matrices.

In practice, for problems with real applications, Gaussian elimination with partial pivoting performs in a stable way.

Stability of solving $Ax = b$ via Cholesky

Given a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a vector $b \in \mathbb{R}^n$, do the following to obtain the solution $x \in \mathbb{R}^n$ to $Ax = b$.

1) Find the factor $R \in \mathbb{R}^{n\times n}$ of the Cholesky factorization $A = R^{\mathrm{T}}R$.
2) Solve $R^{\mathrm{T}}y = b$ for $y \in \mathbb{R}^n$ by forward substitution.
3) Solve $Rx = y$ for $x \in \mathbb{R}^n$ by backward substitution.

We have:

(i) Backward stability of Cholesky factorization: Suppose we apply Cholesky to a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$. Then,

$$\tilde{R}^{\mathrm{T}}\tilde{R} = A + \Delta A \quad \text{for some } \Delta A \in \mathbb{R}^{n\times n} \text{ with } \frac{\|\Delta A\|}{\|A\|} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

(ii) Solving Ax=b via Cholesky is backward stable in the sense that

$$(A + \Delta A)\tilde{x} = b \quad \text{for some } \Delta A \in \mathbb{R}^{n\times n} \text{ with } \frac{\|\Delta A\|}{\|A\|} = \mathcal{O}(\varepsilon_{\text{machine}}).$$

An intuitive reason for the stability of Cholesky factorization, compared to LU factn, is that the factor R in the Cholesky factorization $A = R^{\mathrm{T}}R$ cannot become very large compared to A (e.g., $\|R\|_2^2 = \|A\|_2$).

End of “Chapter 5: Conditioning and Stability”.