

Tutorial 3: Error Analysis and Conditioning

Floating-point arithmetic and error analysis (AFAE)

Sorbonne Université

Contents

1	Exercise 1: Summation	2
1.1	Question 1: Condition Number of Summation	2
1.2	Question 2: Backward Stability of Recursive Summation	3
1.3	Question 3: Relative Error Bound for Summation	3
1.4	Question 4: Dot Product Analysis	4
1.4.1	(a) Condition Number of Dot Product	4
1.4.2	(b) Backward Stability of Dot Product	4
1.4.3	(c) Relative Error Bound	5
2	Exercise 2: Polynomial Evaluation	5
2.1	Question 1: Condition Number Formula	5
2.2	Question 2: Backward Stability of Horner Scheme	6
2.3	Question 3: Relative Error Bound	6
2.4	Question 4: Distance to Nearest Root	7
3	Exercise 3: Roots of Polynomials	7
3.1	Question 1: Condition Number of a Simple Root	7
3.2	Question 2: When is a Simple Root Ill-Conditioned?	8
4	Exercise 4: Conditioning of Matrix Inverse	9
4.1	Question 1: Show that $\kappa(A) = \ A\ \ A^{-1}\ $	9
4.2	Question 2: Distance to Singularity	10
4.3	Question 3: Express $\kappa(A)$ in Terms of Singular Values	10
5	Summary	11
5.1	Key Concepts	11
5.2	Condition Numbers Derived	12
5.3	Backward Stable Algorithms	12
5.4	When Problems Are Ill-Conditioned	12

1 Exercise 1: Summation

1.1 Question 1: Condition Number of Summation

Goal: Show that

$$\text{cond} \left(\sum_{i=1}^n p_i \right) = \frac{\sum_{i=1}^n |p_i|}{\left| \sum_{i=1}^n p_i \right|} \quad (1)$$

Given Definition:

$$\text{cond} \left(\sum_{i=1}^n p_i \right) := \limsup_{\varepsilon \rightarrow 0} \left\{ \frac{\left| \sum_{i=1}^n \tilde{p}_i - \sum_{i=1}^n p_i \right|}{\varepsilon \left| \sum_{i=1}^n p_i \right|} : |\tilde{p}_i - p_i| \leq \varepsilon |p_i| \text{ for } i = 1, \dots, n \right\} \quad (2)$$

Proof:

Let $S = \sum_{i=1}^n p_i$ and $\tilde{S} = \sum_{i=1}^n \tilde{p}_i$.

The perturbation is:

$$\tilde{S} - S = \sum_{i=1}^n \tilde{p}_i - \sum_{i=1}^n p_i = \sum_{i=1}^n (\tilde{p}_i - p_i) \quad (3)$$

Given the constraint $|\tilde{p}_i - p_i| \leq \varepsilon |p_i|$, we can write:

$$\tilde{p}_i - p_i = \delta_i |p_i| \quad (4)$$

where $|\delta_i| \leq \varepsilon$.

Therefore:

$$\tilde{S} - S = \sum_{i=1}^n \delta_i |p_i| \quad (5)$$

Taking absolute values:

$$|\tilde{S} - S| = \left| \sum_{i=1}^n \delta_i |p_i| \right| \leq \sum_{i=1}^n |\delta_i| |p_i| \leq \varepsilon \sum_{i=1}^n |p_i| \quad (6)$$

The supremum is achieved when all δ_i have the same sign as $|p_i|$:

$$\sup |\tilde{S} - S| = \varepsilon \sum_{i=1}^n |p_i| \quad (7)$$

Dividing by $\varepsilon |S|$:

$$\text{cond}(S) = \lim_{\varepsilon \rightarrow 0} \frac{\varepsilon \sum_{i=1}^n |p_i|}{\varepsilon |S|} = \frac{\sum_{i=1}^n |p_i|}{\left| \sum_{i=1}^n p_i \right|} \quad (8)$$

$$\boxed{\text{cond} \left(\sum_{i=1}^n p_i \right) = \frac{\sum_{i=1}^n |p_i|}{\left| \sum_{i=1}^n p_i \right|}} \quad (9)$$

Interpretation

- If all p_i have the same sign, then $\text{cond} = 1$ (well-conditioned)
- If there is massive cancellation, $|\sum p_i| \ll \sum |p_i|$, the condition number is large (ill-conditioned)

1.2 Question 2: Backward Stability of Recursive Summation

Recursive Summation Algorithm:

s	=	p	
s	=	s	p
s	=	s	p
...			
s	=	s	p

Where \oplus denotes floating-point addition: $a \oplus b = (a + b)(1 + \delta)$ with $|\delta| \leq u$ (machine precision).

Backward Stability: An algorithm is backward stable if the computed result $\tilde{f}(x)$ satisfies:

$$\tilde{f}(x) = f(\tilde{x}) \quad (10)$$

where \tilde{x} is a slightly perturbed input: $|\tilde{x} - x| = O(u)|x|$.

Proof:

For the recursive summation:

$$\tilde{s}_k = ((\tilde{s}_{k-1} + p_k)(1 + \delta_k)) \quad (11)$$

Expanding recursively:

$$\tilde{s}_n = ((p_1(1 + \delta_1) + p_2)(1 + \delta_2) + p_3)(1 + \delta_3) \cdots + p_n)(1 + \delta_n) \quad (12)$$

We can rewrite this as:

$$\tilde{s}_n = p_1 \prod_{j=1}^n (1 + \delta_j) + p_2 \prod_{j=2}^n (1 + \delta_j) + \cdots + p_n (1 + \delta_n) \quad (13)$$

Let $\theta_i = \prod_{j=i}^n (1 + \delta_j) - 1$. Using the fact that $\prod (1 + \delta_j) \approx 1 + \sum \delta_j$ for small δ_j :

$$|\theta_i| \leq (n - i + 1)u + O(u^2) \approx (n - i + 1)u \quad (14)$$

Therefore:

$$\tilde{s}_n = \sum_{i=1}^n p_i (1 + \theta_i) = \sum_{i=1}^n \tilde{p}_i \quad (15)$$

where $\tilde{p}_i = p_i(1 + \theta_i)$ with $|\theta_i| \leq nu$.

This shows that the computed sum is the exact sum of slightly perturbed values \tilde{p}_i .

$$\boxed{\text{Recursive summation is backward stable}} \quad (16)$$

Backward Stability

The computed result equals the exact sum of the inputs perturbed by at most $O(nu)$.

1.3 Question 3: Relative Error Bound for Summation

Combining backward stability with conditioning:

Backward Stability gives us:

$$\tilde{s}_n = \sum_{i=1}^n p_i (1 + \theta_i) \quad (17)$$

with $|\theta_i| \leq nu$.

Forward Error:

$$\left| \frac{\tilde{s}_n - s_n}{s_n} \right| = \left| \frac{\sum_{i=1}^n p_i \theta_i}{\sum_{i=1}^n p_i} \right| \quad (18)$$

Using $|\theta_i| \leq nu$:

$$\left| \frac{\tilde{s}_n - s_n}{s_n} \right| \leq \frac{\sum_{i=1}^n |p_i| |\theta_i|}{\left| \sum_{i=1}^n p_i \right|} \leq nu \cdot \frac{\sum_{i=1}^n |p_i|}{\left| \sum_{i=1}^n p_i \right|} \quad (19)$$

$$\boxed{\left| \frac{\tilde{s}_n - s_n}{s_n} \right| \leq nu \cdot \text{cond} \left(\sum_{i=1}^n p_i \right)} \quad (20)$$

Interpretation:

$$\text{Relative Error} \leq \text{Machine Precision} \times \text{Number of Operations} \times \text{Condition Number} \quad (21)$$

1.4 Question 4: Dot Product Analysis

Dot Product: $d = \sum_{i=1}^n x_i y_i$

1.4.1 (a) Condition Number of Dot Product

Let $d = \sum_{i=1}^n x_i y_i$ with perturbations \tilde{x}_i, \tilde{y}_i .

$$\tilde{d} = \sum_{i=1}^n \tilde{x}_i \tilde{y}_i = \sum_{i=1}^n (x_i + \Delta x_i)(y_i + \Delta y_i) \quad (22)$$

$$= \sum_{i=1}^n (x_i y_i + x_i \Delta y_i + y_i \Delta x_i + \Delta x_i \Delta y_i) \quad (23)$$

Neglecting second-order terms:

$$\tilde{d} - d \approx \sum_{i=1}^n (x_i \Delta y_i + y_i \Delta x_i) \quad (24)$$

With $|\Delta x_i| \leq \varepsilon |x_i|$ and $|\Delta y_i| \leq \varepsilon |y_i|$:

$$|\tilde{d} - d| \leq \varepsilon \sum_{i=1}^n (|x_i| |y_i| + |y_i| |x_i|) = 2\varepsilon \sum_{i=1}^n |x_i y_i| \quad (25)$$

Therefore:

$$\boxed{\text{cond}(x \cdot y) = \frac{\sum_{i=1}^n |x_i y_i|}{\left| \sum_{i=1}^n x_i y_i \right|}} \quad (26)$$

This has the same form as the summation condition number!

1.4.2 (b) Backward Stability of Dot Product

The dot product computation involves both multiplication and addition:

\mathbf{t}	$=$	\mathbf{x}	\mathbf{y}
\mathbf{t}	$=$	\mathbf{t}	$(\mathbf{x} \quad \mathbf{y})$
\mathbf{t}	$=$	\mathbf{t}	$(\mathbf{x} \quad \mathbf{y})$
\dots			

Each multiplication: $x_i \otimes y_i = x_i y_i (1 + \delta_i^{\text{mult}})$ with $|\delta_i^{\text{mult}}| \leq u$

Each addition has error as before.

Following similar analysis to summation:

$$\tilde{d} = \sum_{i=1}^n \tilde{x}_i \tilde{y}_i \quad (27)$$

where $\tilde{x}_i \tilde{y}_i = x_i y_i (1 + \theta_i)$ with $|\theta_i| \leq 2nu$.

Dot product computation is backward stable

(28)

1.4.3 (c) Relative Error Bound

$$\left| \frac{\tilde{d} - d}{d} \right| \leq 2nu \cdot \text{cond}(x \cdot y) = 2nu \cdot \frac{\sum_{i=1}^n |x_i y_i|}{|\sum_{i=1}^n x_i y_i|}$$

(29)

Orthogonal Vectors

When $x \perp y$ (nearly orthogonal), $\sum x_i y_i \approx 0$ while $\sum |x_i y_i|$ is not small. This makes the dot product ill-conditioned!

2 Exercise 2: Polynomial Evaluation

2.1 Question 1: Condition Number Formula

For $p(x) = \sum_{i=0}^n a_i x^i$, the condition number of evaluating p at x is:

$$\text{cond}(p, x) = \frac{\sum_{i=0}^n |a_i| |x|^i}{|p(x)|} = \frac{\tilde{p}(|x|)}{|p(x)|}$$

(30)

where $\tilde{p}(x) = \sum_{i=0}^n |a_i| x^i$ is the polynomial with absolute value coefficients.

Derivation:

Consider perturbations $\tilde{a}_i = a_i (1 + \delta_i)$ with $|\delta_i| \leq \varepsilon$:

$$\tilde{p}(x) = \sum_{i=0}^n \tilde{a}_i x^i = \sum_{i=0}^n a_i (1 + \delta_i) x^i \quad (31)$$

$$\tilde{p}(x) - p(x) = \sum_{i=0}^n a_i \delta_i x^i \quad (32)$$

$$|\tilde{p}(x) - p(x)| \leq \varepsilon \sum_{i=0}^n |a_i| |x|^i \quad (33)$$

$$\text{cond}(p, x) = \frac{\sum_{i=0}^n |a_i| |x|^i}{|p(x)|} \quad (34)$$

2.2 Question 2: Backward Stability of Horner Scheme

Horner Scheme:

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + x \cdot a_n) \cdots)) \quad (35)$$

Algorithm:

$$\begin{array}{lcl} b & = & a \\ b & = & a \quad (x \quad b) \\ b & = & a \quad (x \quad b) \\ \dots & & \\ b & = & a \quad (x \quad b) \end{array}$$

Proof of Backward Stability:

At each step, we have:

$$\tilde{b}_i = (a_i + x\tilde{b}_{i+1}(1 + \delta_i^{\text{mult}}))(1 + \delta_i^{\text{add}}) \quad (36)$$

Working backwards from $\tilde{b}_n = a_n$:

$$\tilde{b}_{n-1} = (a_{n-1} + xa_n(1 + \delta_n^{\text{mult}}))(1 + \delta_{n-1}^{\text{add}}) \quad (37)$$

After complete expansion, we can show:

$$\tilde{b}_0 = \sum_{i=0}^n \tilde{a}_i x^i \quad (38)$$

where $\tilde{a}_i = a_i(1 + \theta_i)$ with $|\theta_i| \leq 2nu$.

This means the computed value is the exact evaluation of a slightly perturbed polynomial.

Horner scheme is backward stable

(39)

Efficiency

Horner scheme uses only n multiplications and n additions (optimal!)

2.3 Question 3: Relative Error Bound

Combining backward stability with the condition number:

$$\left| \frac{\tilde{p}(x) - p(x)}{p(x)} \right| \leq 2nu \cdot \text{cond}(p, x) = 2nu \cdot \frac{\tilde{p}(|x|)}{|p(x)|} \quad (40)$$

When is polynomial evaluation ill-conditioned?

- When $|p(x)| \ll \tilde{p}(|x|)$, i.e., when there is significant cancellation in the sum
- Near roots of the polynomial: as $x \rightarrow \alpha$ where $p(\alpha) = 0$, we have $|p(x)| \rightarrow 0$

2.4 Question 4: Distance to Nearest Root

Given: Distance $d(p, q) = \max_i \{|a_i - b_i|/|a_i|\}$

To Show:

$$\min\{d(p, q) : q(z) = 0\} = \frac{1}{\text{cond}(p, z)} \quad (41)$$

Proof:

We want to find the smallest relative perturbation $d(p, q)$ such that $q(z) = 0$.

Let $q(x) = \sum_{i=0}^n b_i x^i$ where $b_i = a_i(1 + \delta_i)$.

For $q(z) = 0$:

$$\sum_{i=0}^n a_i(1 + \delta_i)z^i = 0 \quad (42)$$

$$p(z) + \sum_{i=0}^n a_i \delta_i z^i = 0 \quad (43)$$

The minimum distance occurs when all δ_i are chosen to optimally cancel $p(z)$:

$$\sum_{i=0}^n a_i \delta_i z^i = -p(z) \quad (44)$$

The minimum value of $\max_i |\delta_i|$ needed to achieve this is:

$$d_{\min} = \frac{|p(z)|}{\sum_{i=0}^n |a_i| |z|^i} = \frac{|p(z)|}{\tilde{p}(|z|)} = \frac{1}{\text{cond}(p, z)} \quad (45)$$

$$\boxed{\min\{d(p, q) : q(z) = 0\} = \frac{1}{\text{cond}(p, z)}} \quad (46)$$

Interpretation:

- If $\text{cond}(p, z)$ is large (ill-conditioned), then a tiny perturbation in coefficients can create a root at z
- This explains why roots of polynomials are sensitive to coefficient errors

3 Exercise 3: Roots of Polynomials

3.1 Question 1: Condition Number of a Simple Root

Given: $p(\alpha) = 0$ and $p'(\alpha) \neq 0$ (simple root)

Definition:

$$K(p, \alpha) := \lim_{\varepsilon \rightarrow 0} \sup_{|\Delta a_i| \leq \varepsilon |a_i|} \left\{ \frac{|\Delta \alpha|}{\varepsilon |\alpha|} \right\} \quad (47)$$

To Show:

$$K(p, \alpha) = \frac{\tilde{p}(|\alpha|)}{|\alpha| |p'(\alpha)|} \quad (48)$$

Proof:

Consider the perturbed polynomial:

$$\tilde{p}(x) = \sum_{i=0}^n (a_i + \Delta a_i) x^i \quad (49)$$

Let $\tilde{\alpha} = \alpha + \Delta \alpha$ be the perturbed root: $\tilde{p}(\tilde{\alpha}) = 0$.

Taylor expansion of \tilde{p} around α :

$$\tilde{p}(\tilde{\alpha}) = \tilde{p}(\alpha) + \tilde{p}'(\alpha)\Delta\alpha + O((\Delta\alpha)^2) = 0 \quad (50)$$

Since $\tilde{p}(\alpha) = p(\alpha) + \sum_{i=0}^n \Delta a_i \alpha^i = \sum_{i=0}^n \Delta a_i \alpha^i$ (as $p(\alpha) = 0$):

$$\sum_{i=0}^n \Delta a_i \alpha^i + p'(\alpha)\Delta\alpha + O(\varepsilon^2) = 0 \quad (51)$$

To first order:

$$\Delta\alpha \approx -\frac{\sum_{i=0}^n \Delta a_i \alpha^i}{p'(\alpha)} \quad (52)$$

Taking absolute values:

$$|\Delta\alpha| \leq \frac{\sum_{i=0}^n |\Delta a_i| |\alpha|^i}{|p'(\alpha)|} \leq \frac{\varepsilon \sum_{i=0}^n |a_i| |\alpha|^i}{|p'(\alpha)|} \quad (53)$$

Therefore:

$$\frac{|\Delta\alpha|}{\varepsilon |\alpha|} \leq \frac{\sum_{i=0}^n |a_i| |\alpha|^i}{|\alpha| |p'(\alpha)|} = \frac{\tilde{p}(|\alpha|)}{|\alpha| |p'(\alpha)|} \quad (54)$$

$$\boxed{K(p, \alpha) = \frac{\tilde{p}(|\alpha|)}{|\alpha| |p'(\alpha)|}} \quad (55)$$

3.2 Question 2: When is a Simple Root Ill-Conditioned?

A simple root α is **ill-conditioned** when $K(p, \alpha)$ is large.

From the formula: $K(p, \alpha) = \frac{\tilde{p}(|\alpha|)}{|\alpha| |p'(\alpha)|}$

Ill-conditioning occurs when:

1. **Small derivative:** $|p'(\alpha)| \ll \tilde{p}(|\alpha|)/|\alpha|$

- This happens when the root is nearly a multiple root
- The polynomial is nearly flat at the root
- Example: $p(x) = (x - \alpha)^2 + \varepsilon$ has $p'(\alpha) = O(\sqrt{\varepsilon})$

2. **Large $|\alpha|$ combined with small $|p'(\alpha)|$:**

- High-degree polynomials evaluated at large $|\alpha|$
- The numerator $\tilde{p}(|\alpha|) = \sum |a_i| |\alpha|^i$ grows rapidly with $|\alpha|$

3. **Coefficient growth:**

- When $|a_i|$ are large, especially for high powers
- Wilkinson's polynomial: $(x - 1)(x - 2) \cdots (x - 20)$ has enormous coefficients

Wilkinson's Polynomial

The polynomial $W(x) = \prod_{i=1}^{20} (x - i)$ is famously ill-conditioned. A tiny perturbation to one coefficient can move roots dramatically. This is because $|W'(i)|$ is small relative to the coefficient magnitudes.

Summary: A root is ill-conditioned when:

- **Near multiple root:** $p'(\alpha) \approx 0$

- **Large root magnitude:** $|\alpha|$ is large relative to coefficient scale
- **Coefficient imbalance:** Large variation in coefficient magnitudes

$$\boxed{\text{Ill-conditioned when: } |p'(\alpha)| \ll \frac{\tilde{p}(|\alpha|)}{|\alpha|}} \quad (56)$$

4 Exercise 4: Conditioning of Matrix Inverse

4.1 Question 1: Show that $\kappa(A) = |A||A^{-1}|$

Given Definition:

$$\kappa(A) := \lim_{\varepsilon \rightarrow 0} \sup_{|\Delta A| \leq \varepsilon |A|} \left(\frac{|(A + \Delta A)^{-1} - A^{-1}|}{\varepsilon |A^{-1}|} \right) \quad (57)$$

Proof:

We need a perturbation formula for $(A + \Delta A)^{-1}$.

Lemma (Matrix Inversion Perturbation Formula): If $|\Delta A| \cdot |A^{-1}| < 1$, then $A + \Delta A$ is invertible and:

$$(A + \Delta A)^{-1} = A^{-1} - A^{-1} \Delta A A^{-1} + A^{-1} \Delta A A^{-1} \Delta A A^{-1} - \dots \quad (58)$$

Derivation:

$$(A + \Delta A)^{-1} = (A(I + A^{-1} \Delta A))^{-1} = (I + A^{-1} \Delta A)^{-1} A^{-1} \quad (59)$$

Using the Neumann series $(I + E)^{-1} = I - E + E^2 - E^3 + \dots$ for $|E| < 1$:

$$(I + A^{-1} \Delta A)^{-1} = I - A^{-1} \Delta A + O(|\Delta A|^2) \quad (60)$$

Therefore:

$$(A + \Delta A)^{-1} - A^{-1} = -A^{-1} \Delta A A^{-1} + O(|\Delta A|^2) \quad (61)$$

Taking norms:

$$|(A + \Delta A)^{-1} - A^{-1}| = |A^{-1} \Delta A A^{-1}| + O(|\Delta A|^2) \quad (62)$$

$$\leq |A^{-1}| |\Delta A| |A^{-1}| + O(|\Delta A|^2) \quad (63)$$

For the supremum over $|\Delta A| \leq \varepsilon |A|$, the worst case is when ΔA is aligned with the singular vectors to maximize the norm:

$$\sup_{|\Delta A| \leq \varepsilon |A|} |A^{-1} \Delta A A^{-1}| = \varepsilon |A| |A^{-1}|^2 \quad (64)$$

Therefore:

$$\kappa(A) = \lim_{\varepsilon \rightarrow 0} \frac{\varepsilon |A| |A^{-1}|^2 + O(\varepsilon^2)}{\varepsilon |A^{-1}|} = |A| |A^{-1}| \quad (65)$$

$$\boxed{\kappa(A) = |A| |A^{-1}|} \quad (66)$$

Standard Condition Number

This is the standard definition of the condition number of a matrix!

- $\kappa(A) \geq 1$ (equality when A is orthogonal/unitary)

- Large $\kappa(A)$ means A is close to singular
- $\kappa(A) = \infty$ when A is singular

4.2 Question 2: Distance to Singularity

Distance to Singularity:

$$\text{dist}(A) := \min \left\{ \frac{|\Delta A|}{|A|} : A + \Delta A \text{ is singular} \right\} \quad (67)$$

To Show: $\text{dist}(A) = \kappa(A)^{-1}$

Proof:

$A + \Delta A$ is singular if and only if there exists a unit vector v ($|v| = 1$) such that:

$$(A + \Delta A)v = 0 \quad (68)$$

$$Av = -\Delta Av \quad (69)$$

Taking norms:

$$|Av| = |\Delta Av| \leq |\Delta A||v| = |\Delta A| \quad (70)$$

Since $|v| = 1$:

$$|Av| \leq |\Delta A| \quad (71)$$

The minimum $|\Delta A|$ occurs when we choose v to minimize $|Av|$:

$$\min_{|v|=1} |Av| = \sigma_{\min}(A) = \frac{1}{|A^{-1}|} \quad (72)$$

where $\sigma_{\min}(A)$ is the smallest singular value of A .

Therefore:

$$\min |\Delta A| = \sigma_{\min}(A) = \frac{1}{|A^{-1}|} \quad (73)$$

And:

$$\text{dist}(A) = \frac{\sigma_{\min}(A)}{|A|} = \frac{1}{|A||A^{-1}|} = \frac{1}{\kappa(A)} \quad (74)$$

$$\boxed{\text{dist}(A) = \kappa(A)^{-1}} \quad (75)$$

Interpretation:

- If $\kappa(A)$ is large, A is close to a singular matrix
- A relative perturbation of size $1/\kappa(A)$ can make A singular
- Well-conditioned matrices ($\kappa(A) \approx 1$) are far from singular

4.3 Question 3: Express $\kappa(A)$ in Terms of Singular Values

Singular Value Decomposition (SVD):

$$A = U\Sigma V^T \quad (76)$$

where:

- U, V are orthogonal matrices

- $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$

Properties:

- $|A|_2 = \sigma_1 = \sigma_{\max}(A)$ (largest singular value)
- $|A^{-1}|_2 = 1/\sigma_n = 1/\sigma_{\min}(A)$ (reciprocal of smallest singular value)

Therefore:

$$\kappa(A) = |A|_2 |A^{-1}|_2 = \sigma_{\max}(A) \cdot \frac{1}{\sigma_{\min}(A)} \quad (77)$$

$$\boxed{\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \frac{\sigma_1}{\sigma_n}} \quad (78)$$

Special Cases:

1. **Orthogonal/Unitary matrices:** $\sigma_1 = \sigma_n = 1$

$$\kappa(Q) = 1 \quad (\text{perfectly conditioned}) \quad (79)$$

2. **Diagonal matrices:** $\Sigma = \text{diag}(d_1, \dots, d_n)$

$$\kappa(\Sigma) = \frac{\max_i |d_i|}{\min_i |d_i|} \quad (80)$$

3. **Near-singular matrices:** $\sigma_n \approx 0$

$$\kappa(A) \rightarrow \infty \quad (\text{ill-conditioned}) \quad (81)$$

Practical Interpretation

The condition number is the ratio of the largest to smallest “stretching factors” of the matrix.

- $\kappa(A)$ measures how much the matrix amplifies relative errors
- In floating-point arithmetic with precision u , expect errors of order $\kappa(A) \cdot u$

5 Summary

5.1 Key Concepts

Concept	Formula	Meaning
Condition Number	$\text{cond}(f, x) = \frac{\ f'(x)\ \ x\ }{\ f(x)\ }$	Amplification of relative errors
Backward Stability	$\tilde{f}(x) = f(\tilde{x})$ where $\tilde{x} \approx x$	Computed result = exact result of perturbed input
Forward Bound	$\frac{\ \tilde{y} - y\ }{\ y\ } \leq \text{accuracy} \times \text{cond}$	Relative error bounded by algorithm accuracy \times conditioning

5.2 Condition Numbers Derived

1. **Summation:**

$$\text{cond}\left(\sum p_i\right) = \frac{\sum |p_i|}{|\sum p_i|} \quad (82)$$

2. **Dot Product:**

$$\text{cond}(x \cdot y) = \frac{\sum |x_i y_i|}{|\sum x_i y_i|} \quad (83)$$

3. **Polynomial Evaluation:**

$$\text{cond}(p, x) = \frac{\tilde{p}(|x|)}{|p(x)|} \quad (84)$$

4. **Polynomial Root:**

$$K(p, \alpha) = \frac{\tilde{p}(|\alpha|)}{|\alpha| |p'(\alpha)|} \quad (85)$$

5. **Matrix Inverse:**

$$\kappa(A) = |A| |A^{-1}| = \frac{\sigma_{\max}}{\sigma_{\min}} \quad (86)$$

5.3 Backward Stable Algorithms

- ✓ Recursive summation: $O(nu)$ error
- ✓ Dot product: $O(nu)$ error
- ✓ Horner scheme: $O(nu)$ error per evaluation

5.4 When Problems Are Ill-Conditioned

- **Summation:** Massive cancellation ($\sum p_i \approx 0$ but $\sum |p_i|$ large)
- **Dot Product:** Nearly orthogonal vectors
- **Polynomial Evaluation:** Near roots or significant cancellation
- **Polynomial Roots:** Nearly multiple roots ($p'(\alpha) \approx 0$)
- **Matrix Inverse:** Nearly singular ($\sigma_{\min} \approx 0$)