AMS526: Numerical Analysis I (Numerical Linear Algebra for Computational and Data Sciences)

Lecture 5: Conditioning and Condition Number; Floating Point Arithmetic

Xiangmin Jiao

SUNY Stony Brook

Outline

Overview of Error Analysis

- Error analysis is important subject of numerical analysis
- Given a problem f and an algorithm $\tilde{f}$ with an input x, the *absolute error* is $\|\tilde{f}(x) - f(x)\|$ and relative error is $\|\tilde{f}(x) - f(x)\| / \|f(x)\|$
- What are possible sources of errors?

Overview of Error Analysis

- Error analysis is important subject of numerical analysis
- Given a problem f and an algorithm $\tilde{f}$ with an input x, the *absolute error* is $\|\tilde{f}(x) - f(x)\|$ and relative error is $\|\tilde{f}(x) - f(x)\|/\|f(x)\|$
- What are possible sources of errors?
  - Round-off error (input, computation) – main concern of NLA
  - truncation (approximation) error – main concern for AMS 527
- We would like the solution to be *accurate*, i.e., with small errors
- Error depends on property (*conditioning*) of the problem, property (*stability*) of the algorithm

Example Using SVD Analysis

- Suppose $A \in \mathbb{R}^{n\times n}$ is nonsingular, and let $A = U\Sigma V^T$ be its SVD
- Then

$$Ax = U\Sigma V^T x = \sum_{i=1}^{n} \sigma_i v_i^T x u_i$$

and

$$x = A^{-1}b = \left(U\Sigma V^T\right)^{-1} b = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i$$

- Whether matrix multiplication and linear system are sensitive to small changes in A or b depends on distribution of singular values; nearly zero σ_n can amplify errors in v_i
- How do we formalize this analysis?

Absolute Condition Number

- Condition number is a measure of *sensitivity* of a problem
- *Absolute condition number* of a problem f at x is

$$\hat{\kappa} = \lim_{\varepsilon \to 0} \sup_{\|\delta x\| \leq \varepsilon} \frac{\|\delta f\|}{\|\delta x\|}$$

  where $\delta f = f(x + \delta x) - f(x)$
- Less formally, $\hat{\kappa} = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}$ for infinitesimally small δx
- If f is differentiable, then

$$\hat{\kappa} = \|J(x)\|$$

  where J is the Jacobian of f at x, with $J_{ij} = \partial f_i / \partial x_j$, and matrix norm is induced by vector norms on ∂f and ∂x
- Question: What is absolute condition number of $f(x) = \alpha x$?
- Question: Is absolute condition number scale invariant?

Relative Condition Number

- *Relative condition number* of f at x is

$$\kappa = \lim_{\varepsilon \to 0} \sup_{\|\delta x\| \leq \varepsilon} \frac{\|\delta f\| / \|f(x)\|}{\|\delta x\| / \|x\|}$$

- Less formally, $\kappa = \sup_{\delta x} \frac{\|\delta f\| / \|\delta x\|}{\|f(x)\| / \|x\|}$ for infinitesimally small δx
- Note: we can use different types of norms to get different condition numbers
- If f is differentiable, then

$$\kappa = \frac{\|J(x)\|}{\|f(x)\| / \|x\|}$$

- Question: What is relative condition number of $f(x) = \alpha x$?
- Question: Is relative condition number scale invariant?
- In numerical analysis, we in general use relative condition number
- A problem is *well-conditioned* if κ is small and is *ill-conditioned* if κ is large

Examples

- Example: Function $f(x) = \sqrt{x}$

Examples

- Example: Function $f(x) = \sqrt{x}$
  - Absolute condition number of f at x is $\hat{\kappa} = \|J\| = 1/(2\sqrt{x})$
    - Note: We are talking about the condition number of the problem for a given x
  - Relative condition number $\kappa = \frac{\|J\|}{\|f(x)\|/\|x\|} = \frac{1/(2\sqrt{x})}{\sqrt{x}/x} = 1/2$
- Example: Function $f(x) = x_1 - x_2$, where $x = (x_1, x_2)^T$

Examples

- Example: Function $f(x) = \sqrt{x}$
  - Absolute condition number of f at x is $\hat{\kappa} = \|J\| = 1/(2\sqrt{x})$
    - Note: We are talking about the condition number of the problem for a given x
  - Relative condition number $\kappa = \frac{\|J\|}{\|f(x)\|/\|x\|} = \frac{1/(2\sqrt{x})}{\sqrt{x}/x} = 1/2$
- Example: Function $f(x) = x_1 - x_2$, where $x = (x_1, x_2)^T$
  - Absolute condition number of f at x in ∞-norm is $\hat{\kappa} = \|J\|_\infty = \|(1, -1)\|_\infty = 2$
  - Relative condition number $\kappa = \frac{\|J\|_\infty}{\|f(x)\|_\infty/\|x\|_\infty} = \frac{2}{|x_1 - x_2|/\max\{|x_1|,|x_2|\}}$
  - κ is arbitrarily large (f is ill-conditioned) if $x_1 \approx x_2$ (hazard of cancellation error)
- Note: From now on, we will talk about only relative condition number

Condition Number of Matrix

- Consider $f(x) = Ax$, with $A \in \mathbb{R}^{m \times n}$

$$\kappa = \frac{\|J\|}{\|f(x)\|/\|x\|} = \frac{\|A\|\|x\|}{\|Ax\|}$$

- If A is square and nonsingular, since $\|x\|/\|Ax\| \leq \|A^{-1}\|$

$$\kappa \leq \|A\|\|A^{-1}\|$$

- Note that for $f(b) = A^{-1}b$, its condition number $\kappa \leq \|A\|\|A^{-1}\|$
- We define *condition number of matrix A* as

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

- It is the upper bound of the condition number of $f(x) = Ax$ for any x
- For any induced matrix norm, $\kappa(I) = 1$ and $\kappa(A) \geq 1$
- Note about the distinction between the condition number of a *problem* (the map $f(x)$) and the condition number of a *problem instance* (the evaluation of $f(x)$ for specific x)

Geometric Interpretation of Condition Number

- Another way to interpret at $\kappa(A)$ is

$$\kappa(A) = \sup_{\delta x, x} \frac{\|\delta f\| / \|\delta x\|}{\|f(x)\| / \|x\|} = \frac{\sup_{\delta x} \|A\delta x\| / \|\delta x\|}{\inf_x \|Ax\| / \|x\|}$$

- Question: For what x and δx is the equality achieved?

Geometric Interpretation of Condition Number

- Another way to interpret at $\kappa(A)$ is

$$\kappa(A) = \sup_{\delta x, x} \frac{\|\delta f\|/\|\delta x\|}{\|f(x)\|/\|x\|} = \frac{\sup_{\delta x} \|A\delta x\|/\|\delta x\|}{\inf_x \|Ax\|/\|x\|}$$

- Question: For what x and δx is the equality achieved?
  - Answer: When x is in direction of minimum magnification, and δx is in direction of maximum magnification
- Define *maximum magnification* of A as

$$\text{maxmag}(A) = \max_{\|x\|=1} \|Ax\|$$

  and *minimum magnification* of A as

$$\text{minmag}(A) = \min_{\|x\|=1} \|Ax\|$$

- Then condition number of matrix is $\kappa(A) = \text{maxmag}(A)/\text{minmag}(A)$
- For 2-norm, $\kappa(A) = \sigma_1/\sigma_n$, ratio of largest and smallest singular values

Example of Ill-Conditioned Matrix

Example

Let $A = \begin{bmatrix} 1000 & 999 \\ 999 & 998 \end{bmatrix}$. It is easy to verify that $A^{-1} = \begin{bmatrix} -998 & 999 \\ 999 & -1000 \end{bmatrix}$. So

$$\kappa_\infty(A) = \kappa_1(A) = 1999^2 = 3.996 \times 10^6.$$

Example of Ill-Conditioned Matrix

Example

A famous example is Hilbert matrix, defined by $h_{ij} = 1/(i + j - 1)$, $1 \leq i, j \leq n$. The matrix is ill-conditioned for even quite small n.

For $n \leq 4$, we have

$$H_4 = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{bmatrix},$$

with condition number $\kappa_2(H_4) \approx 1.6 \times 10^4$, and $\kappa_2(H_8) \approx 1.5 \times 10^{10}$.

Outline

Floating Point Representations

- Computers use finite number of bits to represent real numbers
  - Numbers cannot be arbitrarily large or small (associated risks of *overflow* and *underflow*)
  - There must be gaps between representable numbers (potential round-off errors)
- Commonly used computer-representations are floating point representations, which resemble scientific notation

$$\pm(d_0 + d_1\beta^{-1} + \cdots + d_{p-1}\beta^{-p+1})\beta^e,\ 0 \le d_i < \beta$$

  where β is base, p is digits of precision, and e is exponent between e_{min} and e_{max}
- Normalize if $d_0 \neq 0$ (except for 0)
- Gaps between adjacent numbers scale with size of numbers
- Relative resolution given by *machine epsilon* $\epsilon_{\text{machine}} = 0.5\beta^{1-p}$
- For all x, there exists a floating point x' such that $|x - x'| \le \epsilon_{\text{machine}}|x|$

IEEE Floating Point Representations

- Single precision: 32 bits
  - 1 sign bit (S), 8 exponent bits (E), 23 significant bits (M), $(-1)^S \times 1.M \times 2^{E-127}$
  - $\epsilon_{\text{machine}}$ is $2^{-24} \approx 6 \times 10^{-8}$
- Double precision: 64 bits
  - 1 sign bit (S), 11 exponent bits (E), 52 significant bits (M), $(-1)^S \times 1.M \times 2^{E-1023}$
  - $\epsilon_{\text{machine}}$ is $2^{-53} \approx 1.11 \times 10^{-16}$
- Special quantities
  - $+\infty$ and $-\infty$ when operation overflows; e.g., $x/0$ for nonzero x
  - NaN (Not a Number) is returned when an operation has no well-defined result; e.g., $0/0$, $\sqrt{-1}$, $\arcsin(2)$, NaN

Floating Point Arithmetic

- Define $\mathrm{fl}(x)$ as closest floating point approximation to x
- By definition of $\epsilon_{\text{machine}}$, we have:

$$\text{For all } x \in \mathbb{R}, \text{ there exists } \epsilon \text{ with } |\epsilon| \leq \epsilon_{\text{machine}}$$
$$\text{such that } \mathrm{fl}(x) = x(1+\epsilon)$$

- Given operation $+$, $-$, $\times$, and $/$ (denoted by $*$), floating point numbers x and y, and corresponding floating point arithmetic (denoted by $\circledast$), we require that $x \circledast y = \mathrm{fl}(x * y)$
- This is guaranteed by IEEE floating point arithmetic
- Fundamental axiom of floating point arithmetic:

$$\text{For all } x, y \in \mathbb{F}, \text{ there exists } \epsilon \text{ with } |\epsilon| \leq \epsilon_{\text{machine}}$$
$$\text{such that } x \circledast y = (x * y)(1+\epsilon)$$

- These properties will be the basis of error analysis with rounding errors
- Note that floating point arithmetic is **not associative**