Lecture 08
Floating Point Arithmetic; Condition Numbers

Songting Luo

Department of Mathematics
Iowa State University

MATH 562 Numerical Analysis II

Outline

Outline

1. Floating Point Arithmetic

2. Conditioning and Condition Numbers

Floating Point Representations

- Computers can only use finite number of bits to represent a real number
  - 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

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

  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 $\epsilon_{machine} = 0.5\beta^{1-p}$
- For all x, there exists a floating point x' such that $|x - x'| \leqslant \epsilon_{machine}|x|$

IEEE Floating Point Representations

- Single precision: 32 bit
  - 1 sign bit (S), 8 exponent bits (E), 23 significant bits (M)
    $(-1)^S \times 1.M \times 2^{E-127}$
  - $\epsilon_{machine}$ is $2^{-24} \approx 6 \times 10^{-8}$
- Double precision: 64 bits
  - 1 sign bit (S), 11 exponent bits (E), 52 significant bits (M)
    $(-1)^S \times 1.M \times 2^{E-1023}$
  - $\epsilon_{machine}$ is $2^{-53} \approx 1 \times 10^{-16}$
- 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.

Machine Epsilon

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

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

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

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

- These properties will be the basis of error analysis with rounding errors

Outline

Overview of Error Analysis

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

Overview of Error Analysis

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

Overview of Error Analysis

- Error analysis is important subject of numerical analysis
- Given a problem $\mathbf{f}$ and an algorithm $\tilde{\mathbf{f}}$ with an input $\mathbf{x}$, the absolute error is $\|\tilde{\mathbf{f}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|$ and relative error is $\|\tilde{\mathbf{f}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| / \|\mathbf{f}(\mathbf{x})\|$
- What are possible sources of errors?
    - Round-off error (input, computation), truncation (approximation) error
- We would like the solution to be accurate, i.e., with small errors

Overview of Error Analysis

- Error analysis is important subject of numerical analysis
- Given a problem $\mathbf{f}$ and an algorithm $\tilde{\mathbf{f}}$ with an input $\mathbf{x}$, the absolute error is $\|\tilde{\mathbf{f}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|$ and relative error is $\|\tilde{\mathbf{f}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|/\|\mathbf{f}(\mathbf{x})\|$
- What are possible sources of errors?
  - Round-off error (input, computation), truncation (approximation) error
- We would like the solution to be accurate, i.e., with small errors
- The error depends on property (conditioning) of the problem, property (stability) of the algorithm.
  - A well-conditioned problem: small perturbations of $\mathbf{x}$ lead to small changes in $\mathbf{f}(\mathbf{x})$;
  - An ill-conditioned problem: small perturbations of $\mathbf{x}$ lead to large changes in $\mathbf{f}(\mathbf{x})$.

Absolute Condition Number

- Condition number is a measure of sensitivity of a problem
- Absolute condition number of a problem $\mathbf{f}$ at $\mathbf{x}$ is

$$\hat{\kappa} = \lim_{\epsilon \to 0} \sup_{\|\delta \mathbf{x}\| \leqslant \epsilon} \frac{\|\delta \mathbf{f}\|}{\|\delta \mathbf{x}\|}$$

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

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

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

Relative Condition Number

- Relative condition number of a problem **f** at **x** is

$$\kappa = \lim_{\epsilon \to 0} \sup_{\|\delta \mathbf{x}\| \leqslant \epsilon} \frac{\|\delta \mathbf{f}\| / \|\mathbf{f}(\mathbf{x})\|}{\|\delta \mathbf{x}\| / \|\mathbf{x}\|}$$

- Less formally, $\kappa = \sup_{\delta \mathbf{x}} \frac{\|\delta \mathbf{f}\| / \|\delta \mathbf{x}\|}{\|\mathbf{f}(\mathbf{x})\| / \|\mathbf{x}\|}$ for infinitesimally small $\delta \mathbf{x}$
- If **f** is differentiable, then

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

- Question: What is relative condition number of $f(x) = \alpha x$?
- Answer: ?
- 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 $f(x) = \sqrt{x}$
  - Absolute condition number of f at x is $\hat{\kappa} = \|\mathbf{J}\| = 1/(2\sqrt{x})$
    - Note: We are talking about the condition number of the problem for a given x
  - Relative condition number $\kappa = \frac{\|\mathbf{J}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|/\|\mathbf{x}\|} = \frac{1/(2\sqrt{x})}{\sqrt{x}/x} = 1/2$
- Example: function $f(\mathbf{x}) = f(x_1, x_2) = x_1 - x_2$
  - Absolute condition number of f at x in ∞-norm is

$$\hat{\kappa} = \|\mathbf{J}\|_\infty = 2$$

  - Relative condition number $\kappa = \frac{\|\mathbf{J}\|_\infty}{\|\mathbf{f}(\mathbf{x})\|_\infty/\|\mathbf{x}\|_\infty} = \frac{2}{|x_1 - x_2|/\max\{|x_1|,|x_2|\}}$
  - 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