AMS526: Numerical Analysis I (Numerical Linear Algebra)

Lecture 08: Floating Point Arithmetic; Condition Numbers

Xiangmin Jiao

SUNY Stony Brook

Outline

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

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

  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 $|x - x'| \leq \epsilon_{\text{machine}}|x|$

IEEE Floating Point Representations

- Single precision: 32 bit
  - 1 sign bit (S), 8 exponent bits (E), 23 significant bits (M), $(-1)^S \times 1.M \times 2^{E-127}$
  - $\epsilon_{\text{machine}}$ is $2^{-24} \approx 6e-8$
- 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 $2^{-53} \approx e-16$
- 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 $\mathrm{fl}(x)$ 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

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), 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

Absolute Condition Number

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

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

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

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

  where $\boldsymbol{J}$ is the Jacobian of $\boldsymbol{f}$ at $\boldsymbol{x}$, with $J_{ij} = \partial f_i / \partial x_j$, and the matrix norm is induced by vector norms on $\partial \boldsymbol{f}$ and $\partial \boldsymbol{x}$
- 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 $\boldsymbol{f}$ at $\boldsymbol{x}$ is

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

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

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

- 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

Condition Numbers

- *Absolute condition number* of a problem $\boldsymbol{f}$ at $\boldsymbol{x}$ is

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

  where $\delta \boldsymbol{f} = \boldsymbol{f}(\boldsymbol{x} + \delta \boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})$

- Less formally, $\hat{\kappa} = \sup_{\delta \boldsymbol{x}} \frac{\|\delta \boldsymbol{f}\|}{\|\delta \boldsymbol{x}\|}$ for infinitesimally small $\delta \boldsymbol{x}$
- *Relative condition number* of $\boldsymbol{f}$ at $\boldsymbol{x}$ is

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

- Less formally, $\kappa = \sup_{\delta \boldsymbol{x}} \frac{\|\delta \boldsymbol{f}\| / \|\delta \boldsymbol{x}\|}{\|\boldsymbol{f}(\boldsymbol{x})\| / \|\boldsymbol{x}\|}$ for infinitesimally small $\delta \boldsymbol{x}$

Examples

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

Examples

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

Examples

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