CS 450: Numerical Anlaysis[1]

Introduction to Scientific Computing

University of Illinois at Urbana-Champaign

[1] *These slides have been drafted by Edgar Solomonik as lecture templates and supplementary material for the book "Scientific Computing: An Introductory Survey" by Michael T. Heath (slides).*

Scientific Computing Applications and Context

- **Mathematical modelling for computational science** *Typical scientific computing problems are numerical solutions to PDEs*
  - *Newtonian dynamics: simulating particle systems in time*
  - *Fluid and air flow models for engineering*
  - *PDE-constrained numerical optimization: finding optimal configurations (used in engineering of control systems)*
  - *Quantum chemistry (electronic structure calculations): many-electron Schrödinger equation*
- **Linear algebra and computation**
  - *Linear algebra and numerical optimization are building blocks for machine learning methods and data analysis*
  - *Computer architecture, compilers, and parallel computing use numerical algorithms (matrix multiplication, Gaussian elimination) as benchmarks*

Example: Mechanics[2]

- Newton's laws provide incomplete particle-centric picture
- Physical systems can be described in terms of *degrees of freedom* (DoFs)
  - A piston moving up and down requires ____________ DoFs
  - 1-particle system requires ____________ DoFs
  - 2-particle system requires ____________ DoFs
  - 2-particles at a fixed distance require ____________ DoFs
- N-particle system *configuration* described by $3N$ DoFs

[2] *Variational Principles of Mechanics*, Cornelius Lanczos, Dover Books on Physics, 1949.

Course Structure

- **Complex numerical problems are generally reduced to simpler problems**
- **The course topics will follow this hierarchical structure**

Numerical Analysis

- **Numerical Problems involving Continuous Phenomena**:

- **Error Analysis**:

Sources of Error

Demo: Floating Point vs Program Logic

- **Representation of Numbers**:
- **Propagated Data Error**:
- **Computational Error** = $\hat{f}(x) - f(x)$ = **Truncation Error + Rounding Error**

Error Analysis

- **Forward Error**:

- **Backward Error**:

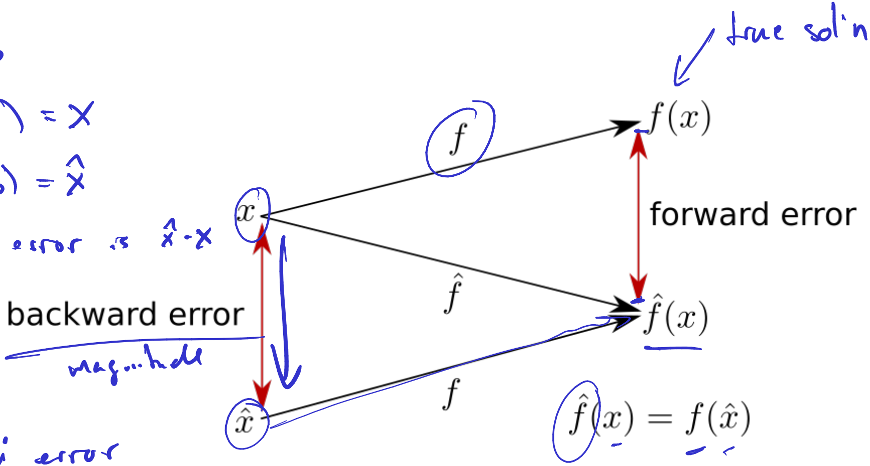
Visualization of Forward and Backward Error

$A \overset{?}{x} = b$

$f(A, b) = x$

$\hat{f}(A, b) = \hat{x}$

forward error is $\hat{x} - x$



backward error

$A\hat{x} - b = A\hat{x} - Ax = \Delta b$

$f(A, b + \Delta b) = \hat{x}$

Demo: *Conditioning of Evaluating tan*

Conditioning

- **Absolute Condition Number:** ≥ 0

evaluate $\tilde{f}$, at x,

evaluate f at $\hat{x}$

for some $\hat{x} \in D$

$$\kappa(f,x) = |f'(x)|$$

$$\kappa(f,D) = \max_{x \in D} \overbrace{\kappa(f,x)}^{f'(x)}$$

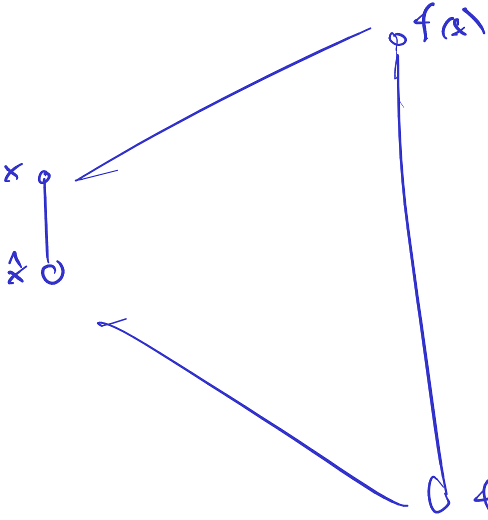
$$\lim_{\Delta x \to 0} \frac{|f(\hat{x}) - f(x)|}{\Delta x} \to f'(x)$$

- **(Relative) Condition Number:**

relative distance $\left\{ \begin{matrix} \hat{x} \\ \uparrow \\ x \end{matrix} \right. \longrightarrow \begin{matrix} f(\hat{x}) \\ \uparrow \\ f(x) \end{matrix} \Big\}$ relative distance

$\hat{x} - x = \Delta$

$$\lim_{\Delta x \to 0} \frac{|f(\hat{x}) - f(x)| / |f(x)|}{|\Delta x| / |x|}$$



amplification factor is bounded by $\kappa(f,x)$

$|\text{forward error}| \leq \kappa(f,x)\,|\text{backward error}|$

Posedness and Conditioning

- **What is the condition number of an ill-posed problem?**

$\kappa(f, x) = \infty$

$\nexists$

soln exists is unique and

- f changes continuously with x

Stability and Accuracy

- **Accuracy**: forward error is small

$\hat{f}(x) \approx f(x)$ for all $x \in D$

algorithm in exact arithmetic (no round-off error)

input domain

- **Stability**:

sensitivity of the algorithm to round-off error

$\tilde{f}(x) \approx \hat{f}(x)$

$\neq \hat{f}(fl(x))$

$\tilde{f}$ is the algorithm in finite precision

Error and Conditioning

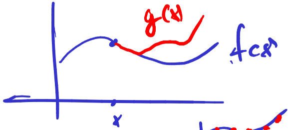
true

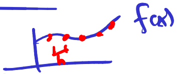
$\tilde{f} \approx \hat{f}$ $\quad$ $\hat{f} \approx f$

- Two major sources of error: *roundoff* and *truncation* error.
  - roundoff error concerns floating point error due to finite precision
  - truncation error concerns error incurred due to algorithmic approximation, e.g. the representation of a function by a finite Taylor series

$$f(x+h) \approx g(h) = \sum_{i=0}^{k} \frac{f^{(i)}(x)}{i!} h^i$$

$$f(x+h) - g(h) = \sum_{i=k+1}^{\infty} \frac{f^{(i)}(x)}{i!} h^i = O(h^{k+1}) \text{ as } h \to 0$$





- To study the propagation of roundoff error in arithmetic we can use the notion of conditioning.

$f(fl(x))$

$\uparrow$

x rounded to
the nearest floating-point number

Floating Point Numbers

Demo: Picking apart a floating point number
Demo: Density of Floating Point Numbers

- **Scientific Notation**

$$\underbrace{2.103}_{\text{significand}} \times 10^{7} \quad \text{exponent}$$

have uniformly low relative error in representation (rounding error)

- **Significand (Mantissa) and Exponent** Given x with s leading bits $x_0, \ldots, x_{s-1}$

bits in the significand and the range of exponents

$[-L, L]$

2^{-L} 2^{L}

significand

$$1.\underbrace{1100101}\times 2^{17}$$

normalized floating point number

store only bits right of the decimal place f-p system

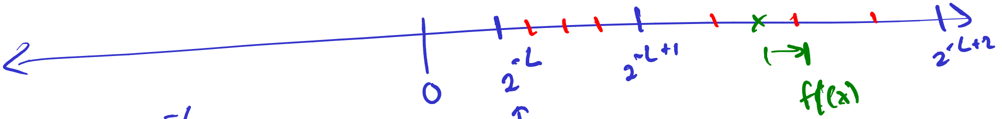
1-bit for sign

b-bits for the significand (b+1 bits of accuracy)

$\log_2(L)$-bits of exponent, to represent number in $[2^{-L}, 2^{L}]$

$\epsilon_{mach} = \overset{2^{-b}}{.01} \times 2^0 = .01$

for $x \in [2^{-L}, 2^{L}]$, $\left|\frac{fl(x)-x}{x}\right| \le \epsilon_{mach}$



underflow limit (UFL)

$$\left.\begin{array}{l} 1.00 \times 2^{-L} \\ 1.01 \times 2^{-L} \end{array}\right\} .01 \times 2^{-L}$$

$\vdots$

$$\left.\begin{array}{l} 1.00 \times 2^{-L+1} \\ 1.01 \times 2^{-L+1} \end{array}\right\} .01 \times 2^{-L+1} = .10 \times 2^{-L}$$

Rounding Error

- **Maximum Relative Representation Error (Machine Epsilon)**

$$\epsilon_{mach} = \underset{\epsilon}{\operatorname{argmin}}\ fl(1+\epsilon) \neq 1$$

$$1.000$$

$$+\ .00011$$

$$1.00011$$

$$\Downarrow$$

$$1.001$$

$$2^{-b}$$

Rounding Error in Operations (I)

Demo: *Catastrophic Cancellation*

Activity: *Cancellation in Standard Deviation Computation*

- **Addition and Subtraction**

subtraction → addition with a negation of an operand

$x - y = x + (-y)$

$3.124 + (-3.102) = .022 \Rightarrow 2.2 \times 10^{-2}$

2.200×10^{-2}

4-digits of accuracy

2 digits of accuracy

catastrophic cancellation

$$\frac{|(x+y) - (fl(x) + fl(y))|}{|x+y|} \leq \frac{\epsilon(|x| + |y|)}{|x+y|}$$

addition for arbitrary x, y is ill-posed

Rounding Error in Operations (II)

- **Multiplication and Division**

multiplication by the reciprocal

$x/y = x \cdot (1/y)$

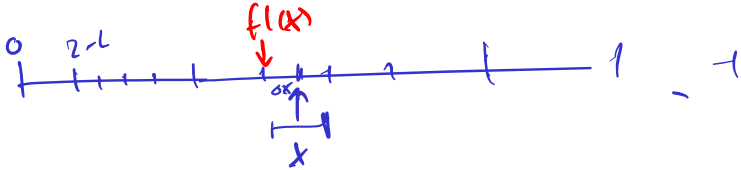
3.120×10^{2}

1.230×10^{-3}

$= 3.89 \cdot 10^{-1}$

$$\frac{|xy - fl(fl(x)fl(y))|}{|xy|}$$

$$\frac{|xy - (1+\varepsilon)^3 xy|}{|xy|} \approx 3\varepsilon$$



$$\frac{|f(x)-x|}{|x|} \le \epsilon_{mech}$$

$$|\Delta x| \le \epsilon |x|$$

$$\epsilon|x| + \epsilon|y|$$

$$\frac{|(x+y)-(\overset{x+\Delta x}{f(x)}+\overset{y+\Delta y}{f(y)})|}{|x+y|} \le \frac{|\Delta x|+|\Delta y|}{|x+y|} \le \frac{\epsilon|x|+\epsilon|y|}{|x+y|}$$

Exceptional and Subnormal Numbers

- **Exceptional Numbers**

- **Subnormal (Denormal) Number Range**

- **Gradual Underflow: Avoiding underflow in addition**

Floating Point Number Line