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

- Newton’s laws provide incomplete particle-centric picture
- Physical systems can be described in terms of \textit{degrees of freedom} (DoFs)
  - A piston moving up and down requires \underline{\text{?}} DoFs
  - 1-particle system requires \underline{\text{?}} DoFs
  - 2-particle system requires \underline{\text{?}} DoFs
  - 2-particles at a fixed distance require \underline{\text{?}} DoFs
- \textit{N}-particle system \textit{configuration} described by $3N$ DoFs

\[2\text{ Variational Principles of Mechanics, Cornelius Lanczos, Dover Books on Physics, 1949.}\]
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

- **Representation of Numbers:**

- **Propagated Data Error:**

- **Computational Error** = \( \hat{f}(x) - f(x) = \text{Truncation Error} + \text{Rounding Error} \)
Error Analysis

- **Forward Error:**

- **Backward Error:**
Visualization of Forward and Backward Error

\[ A\hat{x} = b \]
\[ f(A, b) = \chi \]
\[ \hat{f}(A, b) = \hat{\chi} \]

Forward error is \( \hat{x} - \chi \)

Backward error

backward error

\[ A\hat{x} - b = A\hat{x} - Ax = \Delta b \]
\[ \hat{f}(A, b + \Delta b) = \hat{\chi} \]

\[ f(x) \]

\[ \hat{f}(x) = f(\hat{x}) \]
Conditioning

- Absolute Condition Number:
  
  \[ k(f, x) = \frac{|f'(x)|}{1 + |f(x)|} \]

  evaluate \( f \) at \( x \), for some \( x \in D \)

- (Relative) Condition Number:
  
  \[ k(f, D) = \max_{x \in D} k(f, x) \]

  \[ k(f, D) = \lim_{\Delta x \to 0} \left| \frac{f(x + \Delta x) - f(x)}{f(x)} \right| \]

  \[ \frac{1}{|x|} \]

  relative change \( \frac{f(x + \Delta x) - f(x)}{f(x)} \)
amplification factor $\tilde{\kappa}$ bounded by

$$\|C(f, x)\| \leq \kappa(f, x) \cdot \text{backward error}$$
Posedness and Conditioning

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

\[ \delta(f) = \infty \]

So if a solution exists, it is unique and changes continuously with \( x \).
Stability and Accuracy

- **Accuracy:** Forward error is small
  \[ \hat{f}(x) \approx f(x) \text{ for all } x \in \mathbb{D} \]
  algorithm in exact arithmetic (no round-off error)

- **Stability:** Sensitivity of the algorithm to round-off error
  \[ \tilde{f}(x) \approx \hat{f}(x) \]
  \[ \uparrow \approx \hat{f}(f(x)) \]
  \( \hat{f} \) is the algorithm in finite precision
Error and Conditioning

- Two major sources of error: \textit{roundoff} and \textit{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

\[
f(x+h) \approx g(h) = \sum_{i=0}^{k} \frac{f^{(i)}(x)}{i!} h^i \quad \text{for } f(x) = x^3 + x^2 + x + 1,
\]

\[
f(x+h) - g(h) = \sum_{i=k+1}^{\infty} \frac{f^{(i)}(x)}{i!} h^i = O(h^{k+1})
\]

To study the propagation of roundoff error in arithmetic we can use the notion of conditioning.
Floating Point Numbers

- **Scientific Notation**

\[ 2.103 \times 10^2 \]

have uniformly low relative errors in representation

- **Significand (Mantissa) and Exponent**

Given \( x \) with \( s \) leading bits \( x_0, \ldots, x_{s-1} \)

Normalized floating point number

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

1-bit for sign

\[ e_{\text{r.e.}} = 0.01 \times 2^0 = 0.01 \]

for \( x \in [2^{-L}, 2^{-L+1}] \), \( \frac{E(x) - x}{x} \leq e_{\text{r.e.}} \)

5-bit for the significand (65 bits of accuracy)

\( \log_2(L) \)-bit of exponent, to represent underflow in \([2^{-L}, 2^{1-L}]\)

-0.00 \times 2^{-L}

-1.00 \times 2^{-L+1}

-1.01 \times 2^{-L+1}

\[ f(x) \]

underflow (UFL)

-0.01 \times 2^{-L+1} \leq 0.10 \times 2^{-L+1}
Rounding Error

- Maximum Relative Representation Error (Machine Epsilon)

\[ \varepsilon_{\text{mach}} = \min \{ \varepsilon : f(1 + \varepsilon) \neq 1 \} \]
Rounding Error in Operations (I)

Demo: Catastrophic Cancellation

Activity: Cancellation in Standard Deviation Computation

Addition and Subtraction

Subtraction is addition with a negation of an operand:

\[ x - y = x + (-y) \]

\[
3.1297 + (-3.102) = 0.02667
\]

9 digits of accuracy

catastrophic cancellation

\[
| (x+y) - (f1(x) + f1(y)) | \leq \varepsilon \frac{1 | x + y |}{1 | x | + | y |}
\]

addition for arbitrary \( x, y \) is ill-posed
Rounding Error in Operations (II)

- **Multiplication and Division**

  \[
  \frac{x}{y} = x \cdot \left(\frac{1}{y}\right)
  \]

  \[
  \begin{align*}
  3.120 \times 10^2 \\
  \times \frac{1.230 \times 10^{-3}}{}
  \end{align*}
  \]

  \[= 3.29 \times 10^{-1}\]

  \[
  1 \times y - f_1(f(x) + f(y))
  \]

  \[
  1 \times y - (1 + e)^3 \times y
  \]

  \[\approx 3e\]
Exceptional and Subnormal Numbers

- Exceptional Numbers

- Subnormal (Denormal) Number Range

- Gradual Underflow: Avoiding underflow in addition
Floating Point Number Line

\[ \epsilon \cdot 2^{-L} \quad \text{smaller than or equal to the gap between} \]
\[ \text{any two floating point numbers, so} \]
\[ fl(a - b) = 0 \quad \text{iff} \quad fl(a) = fl(b) \]

- \[ 0 \]
- \[ 2^{-L} \]
- \[ UFL \]
- \[ \text{subnormal} \]
- \[ 0.xy \cdot 2^{-L} \]
- \[ \text{normalized} \]
- \[ 1.xy \cdot 2^{E}, \quad E \in [-L, L] \]