I. Introduction to Numerical Analysis

2008. 9

담당교수: 주 한 규
joohan@snu.ac.kr, x9241, Rm 32-205
원자핵공학과

Introduction to Numerical Analysis

- Need for Numerical Analysis
- Scope and Elements of Numerical Analysis
- Basic Concepts in Numerical Analysis
  - **Discretization**
    - Truncation Error
  - **Binary Representation of Numbers**
    - Round-off Error
  - **Significant Digits**
    - Loss of Significant Digits
    - Problems of Finite Digit Arithmetic
    - Significant Digits in Single and Double Precisions

Needs for Numerical Analysis

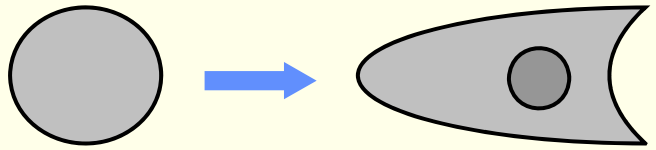
☐ Limitations of Analytic Solutions in Engineering Design

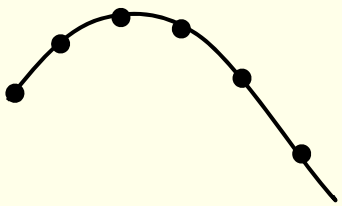
- **Automobile, Airplane, Nuclear Reactor, etc**
- **Almost all practical engineering problems <u>can not be solved analytically</u> because of**
  - Nonlinearity of governing equation: e.g. spring-mass system with non-constant spring constant and second order drag (in shock absorber)

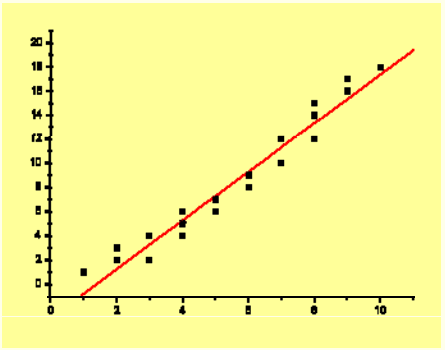
$$m\frac{d^2y}{dt^2} = F(t) - k_0 y - \beta\frac{dy}{dt} \quad \rightarrow \quad m\frac{d^2y}{dt^2} = F(t) - k_0(1-\alpha y)y - \beta\left(\frac{dy}{dt}\right)^2$$

  - Complexity of geometry and composition: e.g. heat conduction problem satisfying boundary condition

$$-\nabla \cdot k\nabla T(\vec{r}) = q'''(\vec{r})$$







☐ Data Analysis and Trend Estimate

- **Data from Measurement or Survey**
- **Draw Trend or Behavior**

☐ Simulation or Prediction

- **Weather Forecast**
- **Flight Simulator**

Scope of Numerical Analysis

- **Interpolation and Approximation**
  - Fitting of given data points for use in interpolation, differentiation or integration - Lagrange Interpolation
  - Approximate representation of data behavior or function - Least Square Method
- **Integration or Differentiation**
  - Trapezoidal formula, Gaussian Quadrature
- **Root of Nonlinear Equation**
  - Higher order polynomial or transcendental equations – Newton-Raphson Method
- **System of Linear Equations**
  - Direct Elimination Methods
  - Iterative Methods
- **Ordinary Differential Equations with Initial Conditions**
  - Runge-Kutta Method
- **Partial Differential Equations with Boundary Conditions**
  - Finite Difference Methods
  - Finite Element Methods

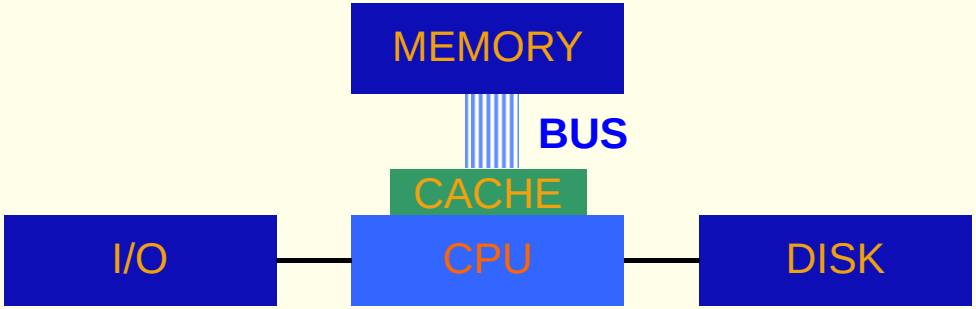
Elements of Numerical Analysis

Algorithm

- Calculation Logic
- Contents and Sequence of Calculation

Computer

- Performance Dictated by CPU Speed and Memory Capacity



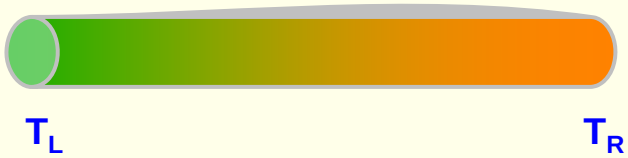
Language (Programming)

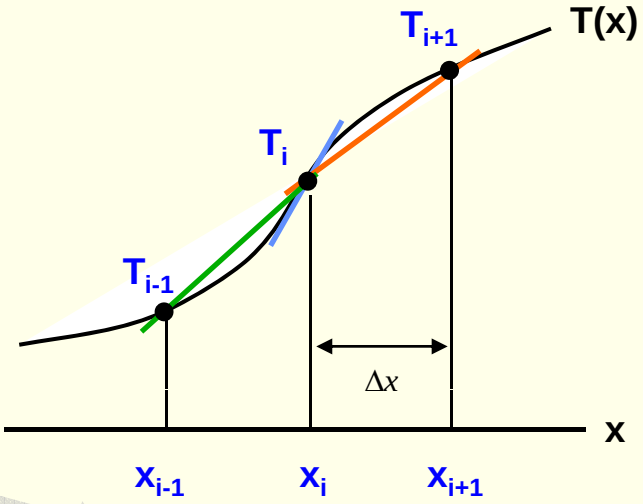
- FORTRAN90
- C, C++
- MATLAB

Basic Concepts in Numerical Analysis

☐Discretization (차분화)

- Continuous variation → Represent with discrete points





- Approximation of Derivative

$$T_i' = \left.\frac{dT}{dx}\right|_{x_i} \qquad T_i' \approx \frac{T_{i+1} - T_i}{\Delta x} \qquad T_i' \approx \frac{T_i - T_{i-1}}{\Delta x}$$

- Leads to Truncation Error (절단 오차)

$$T_{i+1} = T_i + \left.\frac{dT}{dx}\right|_{x_i} \Delta x + \frac{1}{2}\left.\frac{d^2T}{dx^2}\right|_{x_i} (\Delta x)^2 + \frac{1}{3!}\left.\frac{d^3T}{dx^3}\right|_{x_i} (\Delta x)^3 + \ldots$$

$$\Rightarrow \quad \left.\frac{dT}{dx}\right|_{x_i} = \frac{T_{i+1} - T_i}{\Delta x} + O(\Delta x^2)$$

☐Binary Numbers and Significant Digits (유효 숫자)

- Representation of real numbers with finite numbers of bits
- Leads to round-off error (끝처리 오차)

Binary Representation of Numbers

☐Integer Representation

- **4 Bytes (32 bits) in most computers**

  1101101011111010110100110100011

- **1 bit for sign**
- **31 bits for numbers**
  - Largest Positive Integer: $2^{31}-1 = 2{,}147{,}483{,}647 = 2\,\text{Giga}$

☐Real Number with Floating Point Representation (浮動소수점 표기)

$$X = \pm(.d_1 d_2 d_3 \cdots d_n)_\beta \cdot \beta^E$$

Exponent (指數)

Matissa (假數)

Base (밑)

$$\beta = \begin{cases} 10 & Demimal \\ 2 & Binary \\ 16 & Hexadecimal \end{cases} \qquad 0 \le d_i < \beta \qquad d_1 \neq 0 \qquad E = \text{integer}$$

- **n-digit Matissa**

$$f = .d_1 d_2 d_3 \cdots d_n = d_1 \times \frac{1}{\beta} + d_2 \times \frac{1}{\beta^2} + d_3 \times \frac{1}{\beta^3} \cdots + d_n \times \frac{1}{\beta^n}$$

Binary Representation of Numbers

☐ Single Precision Floating Point Number (4 Bytes)

$$\underbrace{1}_{1}\underbrace{10101010}_{8}\underbrace{11011110111111101111011}_{23}$$

Exponent Mantissa

- **Range of Exponent (1 Byte)**

$$-128 \le E \le 127 \qquad \Rightarrow 2^{127} = 1.7 \times 10^{38} \qquad \Rightarrow 2^{-128} = 2.9 \times 10^{-39}$$

  - Overflow or Underflow if a number is greater or smaller than the above bounds

- **Range of Mantissa**

$$0.5 \le f < 1.0 \qquad 0.5 \le f \le 1 - \left(\frac{1}{2}\right)^{23} = 0.99999988$$

☐ Double Precision Floating Point Number (8 Bytes)

$$\underbrace{1}_{1}\underbrace{10101010101}_{11}\underbrace{11011110111\cdots11101111011}_{52}$$

- **Range of Exponent : 8x38=304→ 10^{304}**
- **Significantly (29) More Digits for Mantissa → Increase in Significant Digits**

Significant Digits

Definition of Number of Significant Digits

- **Let $\tilde{x}$ be an approximate number representing the true number x**
  - E.g.

$$\tilde{x} = \begin{cases} 30.0 \\ 30.5 \end{cases} \quad \text{for} \quad x = 30.48$$

  - Relative Error

$$e = \left| \frac{\tilde{x} - x}{x} \right| = \begin{cases} 0.48/30.48 \approx 0.015748 \\ 0.02/30.48 \approx 0.000656 \end{cases} \quad \text{for} \quad x = 30.48$$

- **Then $\tilde{x}$ has n-significant digits if n is the largest integer bounding the relative error with the following:**

$$e = \left| \frac{\tilde{x} - x}{x} \right| < 5 \times 10^{-n} \quad = \frac{1}{2} \times \beta^{-n+1}$$

- **Simply, the number of digits in the mantissa which is correct for sure except the last digit which could be d-1 in case of round-off.**

$$\tilde{x} = \begin{cases} 0.30 \times 10^2 & \text{with 2 significant digits} \\ 0.305 \times 10^2 & \text{with 3 significant digits} \end{cases}$$

More on Significant Digits

- **Express the following numbers with 3 significant digits, and then 4 s.d.**

• 9.8664,	9.87 or 0.987×10^1	9.866 or 0.9866×10^1
• 0.12546,	0.125	0.1255
• 99.997,	100. or 0.100×10^2	100.0 or 0.1000×10^2
• 999.97,	0.100×10^3	1000. or 0.1000×10^3

- **Round-off Error**
  - **Arises Because of Using Finite Number of Significant Digits**
  - **E.g.** $\frac{1}{3} = 0.33333, \quad \frac{2}{3} = 0.66667, \quad \frac{1}{9} = 0.11111, \quad \pi = 3.1416$

Loss of Significant Digits

Addition of a Large Number and a Small Number

$$\frac{1}{3}+\frac{1}{9}\times 10^4 = 0.33333 + 0.11111\times 10^4 = (0.000033333 + 0.11111)\times 10^4$$

$$= 0.111133333\times 10^4 \longleftarrow \text{True} \qquad = 0.11113\times 10^4 \longleftarrow \text{5 S. D.}$$

- **Loses 4 Significant Digits of the Small Number**

Subtraction of Similar Numbers

$$0.12537 - 0.12533 = 0.4\times 10^{-4}$$

- **Number of Significant Digits Decreased to 1 From 5**

Problem of Finite Digits Arithmetic

Solve

$$\begin{aligned} 10^{-6}x + y &= 1 \\ x + y &= 2 \end{aligned}$$

Analytic Solution

$$x = 1.000001,\ y = 0.999999$$

- **Multiply 10^6 to Eq. 1 and then subtract from Eq.2**

$$999999\,y = 999998$$

True

$$\rightarrow 1.0000 \times 10^6\, y = 1.0000 \times 10^6$$

5 Digit Arithmetic

$$y = 1.0000$$

- **Insert y=1.0 into Eq. 1**

$$x = \frac{1 - y}{0.000001} \rightarrow x = 0.0000$$

- **Leads to a Totally Wrong Solution**

Problem of Finite Digits Arithmetic - Alternative

Solve

$$\begin{aligned} 10^{-6}x + y &= 1 \\ x + y &= 2 \end{aligned}$$

- **Eliminate y first by subtracting Eq. 1 from Eq.2**

$$0.999999x = 1.0 \qquad \rightarrow 1.0000x = 1.0000$$

True **5 Digit Arithmetic**

$$x = 1.0000$$

- **Insert x=1.0 into Eq. 1 and Solve for y**

$$y = 1 - 10^{-6}x \rightarrow y = 1.0000$$

- **Leads to a Reasonable Wrong Solution → Importance of Pivotting**

Points

- **Numerical solution always involves round-off errors**
- **Elaborated Algorithms Needed**

Significant Digits in Single and Double Precisions

Single Precision with 23 Digit Mantissa

$$.\underbrace{100\cdots00}_{23}100\cdots001 = \frac{1}{2} + \frac{1}{2^{24}} + \varepsilon \approx .\underbrace{100\cdots01}_{23} = \frac{1}{2} + \frac{1}{2^{23}}$$

$$\frac{\left|\frac{1}{2^{24}} - \frac{1}{2^{23}}\right|}{\frac{1}{2} + \frac{1}{2^{24}} + \varepsilon} \cong \frac{\frac{1}{2^{24}}}{\frac{1}{2}} = \frac{1}{2^{23}} = \frac{1}{8{,}388{,}608} = 1.2\times10^{-7} < 5\times10^{-7}$$

- **Maximum 7 Significant Digits**

Double Precision with 52 Digit Mantissa

$$\frac{1}{2^{52}} = 2.2\times10^{-16} < 5\times10^{-16}$$

- **Maximum 16 Significant Digits**
- **Requires more memory and computation time**