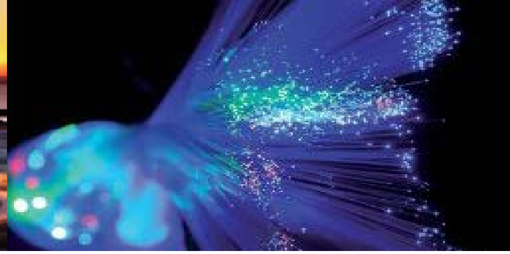
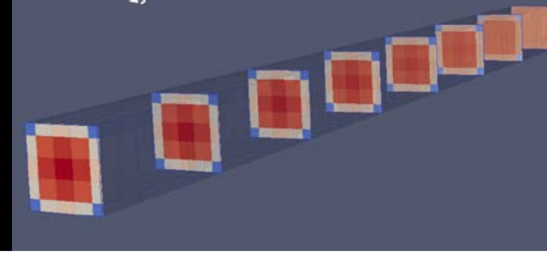
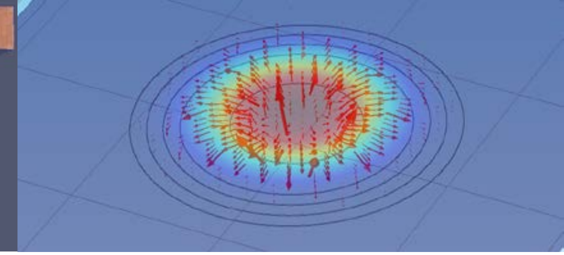


INTRODUCTION TO NUMERICAL ANALYSIS

Cho, Hyoung Kyu

Department of Nuclear Engineering

Seoul National University

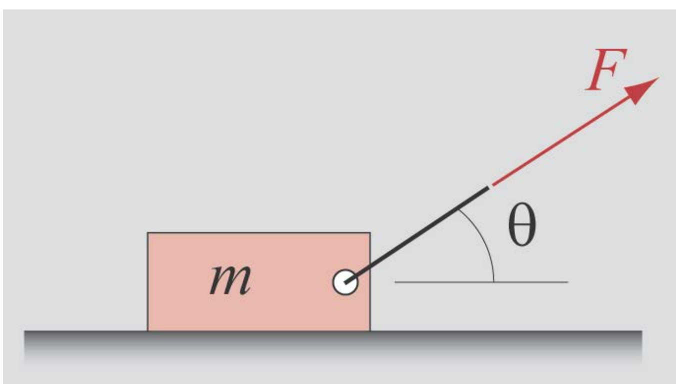
NuTHEL
Nuclear Thermal Hydraulic Engineering Lab.

1. INTRODUCTION

1.1 Introduction

Numerical methods

- Mathematical techniques used for solving mathematical problems that cannot be solved or are difficult to solve analytically
- An analytical solution
  - Exact answer in the form of a mathematical expression
- Numerical solution
  - Approximate numerical value (a number) for the solution
  - Although numerical solutions are an approximation, they can be very accurate.
  - In many numerical methods, the calculations are executed in an iterative manner until a desired accuracy is achieved.

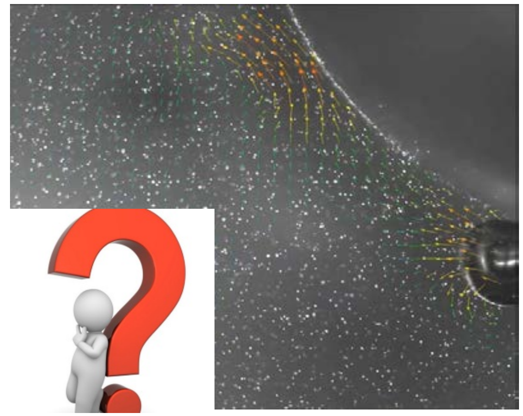


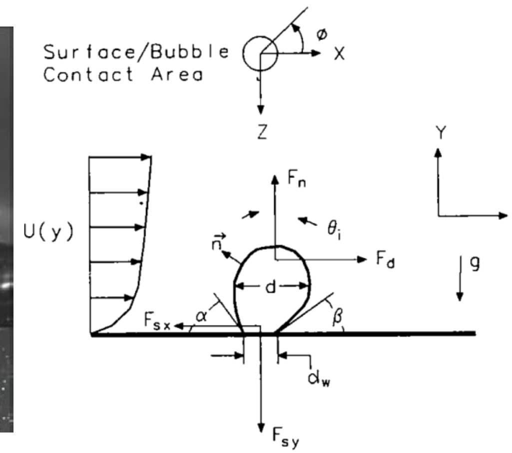
$$F = \frac{\mu m g}{\cos\theta + \mu \sin\theta}$$

1.1 Introduction

❖ Solving a problem in science and engineering

- Problem statement
  - Variables
  - Boundary/initial conditions
- Formulation of the solution
  - Model (physical laws)
  - Governing equations
- Programming (of numerical solution)
  - Selection of numerical method
    - Differ in accuracy, length of calculations, and difficulty in programming
  - Implementation
    - Algorithm + computer program
- Interpretation of the solution
  - Verification and validation





$$\Sigma F_x = F_{sx} + F_{qs} + F_{dux}$$

$$\Sigma F_y = F_{sy} + F_{duy} + F_{sL} + F_b + F_h + F_{cp}$$

$$F_{sx} = -\int_0^{\pi} d_w \sigma \cos\gamma \cos\phi \, d\phi$$

$$F_{sy} = -\int_0^{\pi} d_w \sigma \sin\gamma \, d\phi.$$

1.2 Representation of numbers on a computer

Decimal and binary representation

- Decimal system

$$\begin{array}{ccccccccc} 10^4 & 10^3 & 10^2 & 10^1 & 10^0 & 10^{-1} & 10^{-2} & 10^{-3} & 10^{-4} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 6 & 0 & 7 & 2 & 4\,. & 3 & 1 & 2 & 5 \end{array}$$

$$6\times10^4+0\times10^3+7\times10^2+2\times10^1+4\times10^0+3\times10^{-1}+1\times10^{-2}+2\times10^{-3}+5\times10^{-4}=60{,}724.3125$$

- Binary system

$$\begin{array}{cccccccc} 2^4 & 2^3 & 2^2 & 2^1 & 2^0 & 2^{-1} & 2^{-2} & 2^{-3} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 0 & 0 & 1 & 1\,. & 1 & 0 & 1 \end{array}$$

$$1\times2^4+0\times2^3+0\times2^2+1\times2^1+1\times2^0+1\times2^{-1}+0\times2^{-2}+1\times2^{-3}$$

$$1\times16+0\times8+0\times4+1\times2+1\times1+1\times0.5+0\times0.25+1\times0.125 = 19.625$$

❖ Decimal and binary representation

- Decimal to binary

$$
\begin{array}{ll}
0.188 \times 2 = 0.376 & \text{carry} = 0 \\
0.376 \times 2 = 0.752 & \text{carry} = 0 \\
0.752 \times 2 = 1.504 & \text{carry} = 1 \\
0.504 \times 2 = 1.008 & \text{carry} = 1 \\
0.008 \times 2 = 0.016 & \text{carry} = 0
\end{array}
\quad \Big\downarrow \text{MSB}
$$

$$\text{Answer} = .00110 \text{ (for five significant digits)}$$

  - $0.625 \Rightarrow ?$

1.2 Representation of numbers on a computer

❖ Floating point representation

- To accommodate large and small numbers
  - Real numbers are written in floating point representation.

$$d.\underline{dddddd} \times 10^{p}$$

  - Order of magnitude
    - Ex) 3.91×10^{-6} 6.51923×10^{3}

- Binary floating point

$$1.\underline{bbbbbb} \times 2^{bbb}$$

  - Normalizing the number with respect to the largest power of 2 that is smaller than the number itself.

$$50 = \frac{50}{2^{5}} \times 2^{5} = 1.5625 \times 2^{5}$$

$$1.1001 \times 2^{101}$$

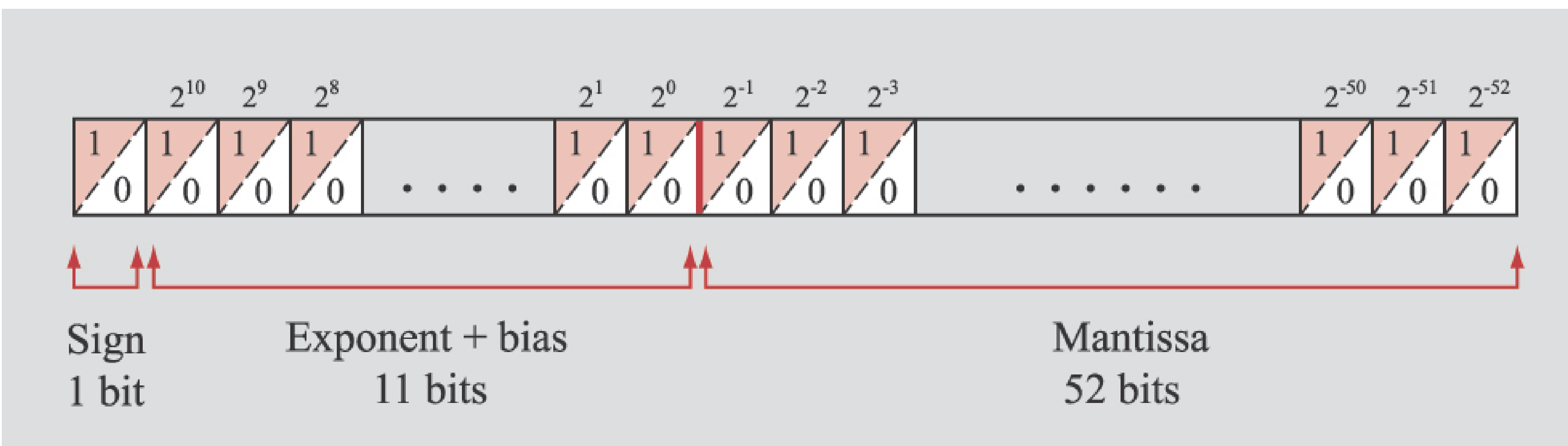
$$1344 = \frac{1344}{2^{10}} \times 2^{10} = 1.3125 \times 2^{10}$$

$$0.3125 = \frac{0.3125}{2^{-2}} \times 2^{-2} = 1.25 \times 2^{-2}$$

1.2 Representation of numbers on a computer

Storing a number in computer memory

- The computer stores
- Single precision vs. double precision
  - 4 bytes (32 bits) vs. 8 bytes (64 bits)
- First bit $\Rightarrow$ sign (0 $\Rightarrow$ + , 1 $\Rightarrow$ -)



1.2 Representation of numbers on a computer

Storing a number in computer memory

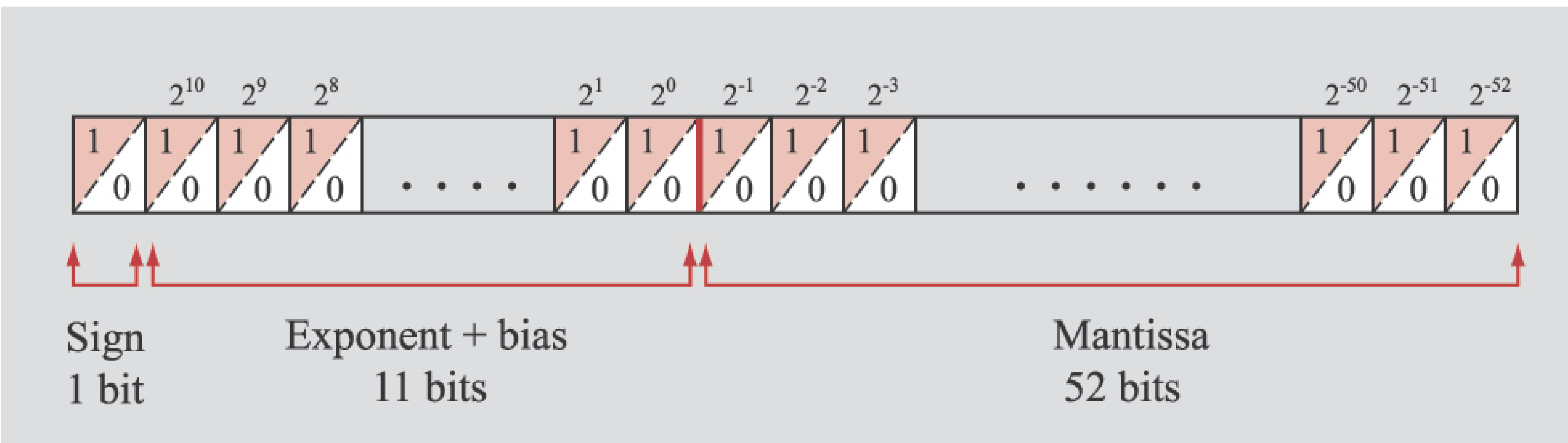
- Exponent + bias
- Mantissa

Bias?

- The bias is introduced in order to avoid using one of the bits for the sign of the exponent (since the exponent can be positive or negative)

Exponent 4 $\Rightarrow$ stored value 4+1023 = 1027

range of exponent: -1023~1024



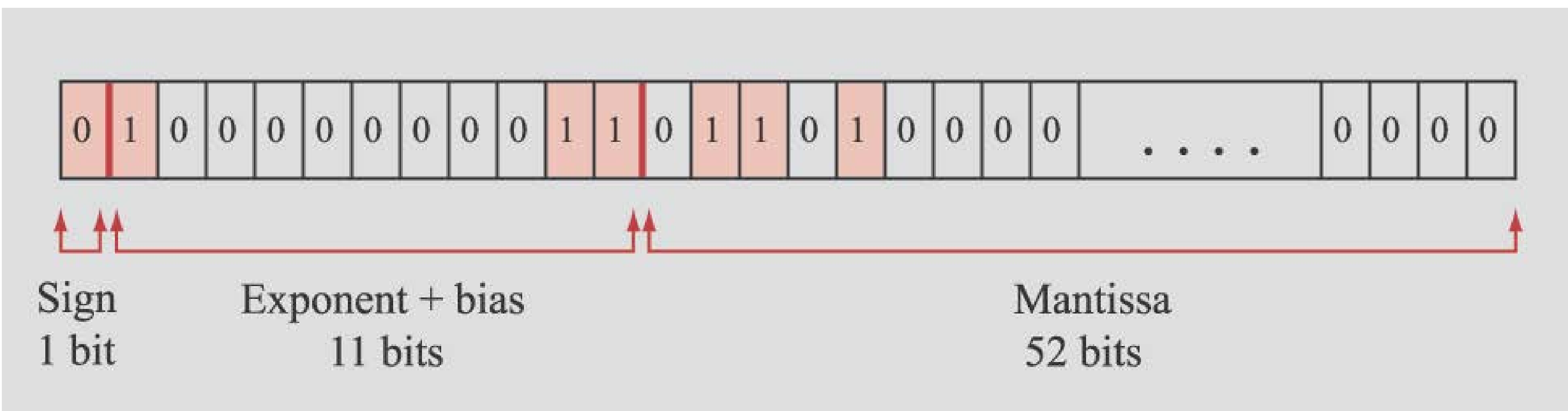
Storing a number in computer memory

- Ex)
  - 22.5 in double precision

$$\frac{22.5}{2^4}2^4 = 1.40625 \times 2^4 \xrightarrow{\text{exponent}} 4 + 1023 = 1027$$

mantissa

0.40625



Additional notes

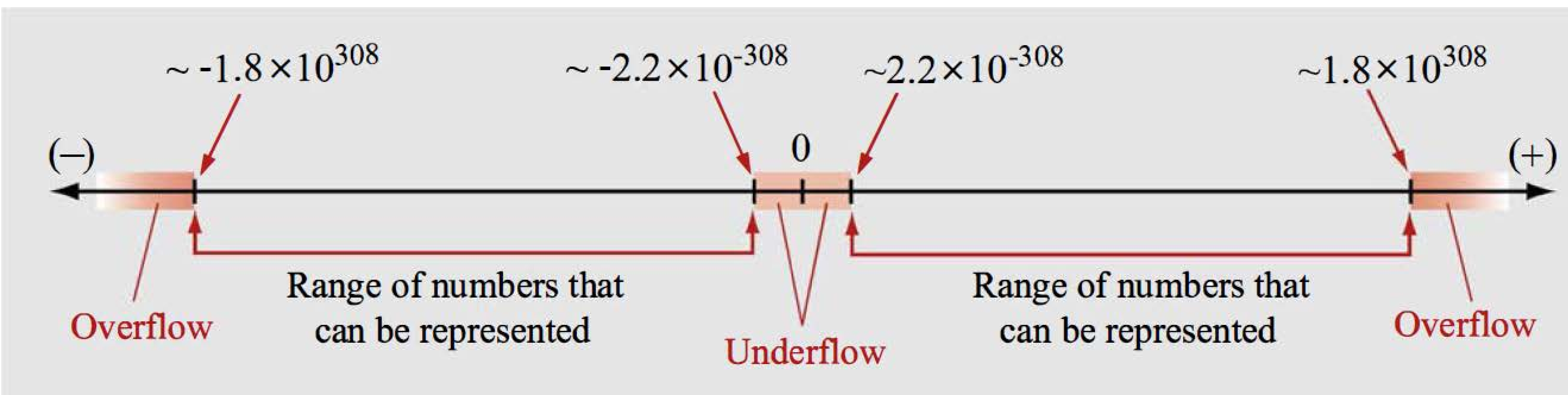
- Smallest number in double precision

  1.0×2^{-1022} $\qquad$ $2^{-1022} \approx 2.2 \times 10^{-308}$

- Largest number in double precision

  1.0×2^{1024} $\qquad$ $2^{1024} \approx 1.8 \times 10^{308}$

- Underflow and overflow



1.2 Representation of numbers on a computer

Additional notes

- Ex) Single precision, bias=127

E	Real Exponent	F	Value
0000 0000	Reserved	000…0	0_{10}
		xxx…x	Unnormalized $(-1)^S \times 2^{-126} \times (0.F)$
0000 0001	-126_{10}		Normalized $(-1)^S \times 2^{e-127} \times (1.F)$
0000 0010	-125_{10}		
…	…		
0111 1111	0_{10}		
…	…		
1111 1110	127_{10}		
1111 1111	Reserved	000…0	Infinity
		xxx…x	NaN

❖ Additional notes

- Errors
  - Ex) 0.1 $\Rightarrow 1.6 \times 2^{-4}$: 0.6 cannot be written exactly
  - The errors that are introduced are small in one step.
  - But when many operations are executed, the errors can grow to such an extent that the final answer is affected.

- Interval between numbers
  - Smallest value of the mantissa: $2^{-52} \approx 2.22 \times 10^{-16}$
  - Smallest possible difference in the mantissa between two numbers
  - The interval depends on the exponent.
  -

1.3 Errors in numerical solutions

Types of error

- Round-off errors (반올림오차)
  - Occurs because of the way that computers store numbers and execute numerical operations
- Truncation errors (절단오차)
  - Introduced by the numerical method that is used for the solution.
- Total error of the numerical solution
  - Difference between the exact solution and the numerical solution

Round-off errors

- Chopping off (discarding)
- Rounding

Example 1-2: Round-off errors

Consider the two nearly equal numbers $p = 9890.9$ and $q = 9887.1$. Use decimal floating point representation (scientific notation) with three significant digits in the mantissa to calculate the difference between the two numbers, $(p - q)$. Do the calculation first by using chopping and then by using rounding.

SOLUTION

In decimal floating point representation, the two numbers are:

$p = 9.8909 \times 10^3$ and $q = 9.8871 \times 10^3$

If only three significant digits are allowed in the mantissa, the numbers have to be shortened. If chopping is used, the numbers become:

$p = 9.890 \times 10^3$ and $q = 9.887 \times 10^3$

Using these values in the subtraction gives:

$-q = 9.890 \times 10^3 - 9.887 \times 10^3 = 0.003 \times 10^3 = 3$

If rounding is used, the numbers become:

$p = 9.891 \times 10^3$ and $q = 9.887 \times 10^3$ (q is the same as before)

Using these values in the subtraction gives:

$-q = 9.891 \times 10^3 - 9.887 \times 10^3 = 0.004 \times 10^3 = 4$

The true (exact) difference between the numbers is 3.8. These results show that, in the present problem, rounding gives a value closer to the true answer.

Round-off errors

- Are likely to occur
  - When the numbers that are involved in the calculations differ significantly in their magnitude.
  - When two numbers that are nearly identical are subtracted from each other.
- Example

$$x^2 - 100.0001x + 0.01 = 0$$

  - Exact solutions: $x_1 = 100 \ \ and \ x_2 = 0.0001$
  - Numerical solution: $x_1 = 100 \ \ and \ x_2 =$ 1.000000000033197e−004

$$x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \frac{(-b + \sqrt{b^2 - 4ac})}{(-b + \sqrt{b^2 - 4ac})} = \frac{2c}{-b + \sqrt{b^2 - 4ac}}$$

  - Numerical solution: $x_1 = 100 \ \ and \ x_2 =$ 1.000000000000000e−004

Round-off errors

- Example

Example 1-3: Round-off errors

Consider the function:

$$f(x) = x(\sqrt{x} - \sqrt{x-1}) \tag{1.12}$$

(*a*) Use MATLAB to calculate the value of $f(x)$ for the following three values of x: $x = 10$, $x = 1000$, and $x = 100000$.

(*b*) Use the decimal format with six significant digits to calculate $f(x)$ for the values of x in part (*a*). Compare the results with the values in part (*a*).

(*c*) Change the form of $f(x)$ by multiplying it by $\frac{\sqrt{x}+\sqrt{x-1}}{\sqrt{x}+\sqrt{x-1}}$. Using the new form with numbers in decimal format with six significant digits, calculate the value of $f(x)$ for the three values of x. Compare the results with the values in part (*a*).

$$f(100000) = 100000 * (\sqrt{100000} - \sqrt{100000-1}) = 158.1143$$

$$f(100000) = 100000(\sqrt{100000} - \sqrt{100000-1}) = 100000(316.228 - 316.226) = 200$$

$$f(100000) = \frac{100000}{\sqrt{100000} + \sqrt{100000-1}} = \frac{1000}{316.228 + 316.226} = 158.114$$

Truncation errors

- Ex) Taylor's series expansion

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \ldots$$

  - Exact value can be determined if an infinite number of terms are used.
  - The value can be approximated by using only a finite number of terms.
  - Truncation error = difference between the true value and an approximated value

- Ex) Derivative

$$\left.\frac{df(x)}{dx}\right|_{x = x_1} = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

- The truncation error is independent of round-off error.
- It exists even when the mathematical operations themselves are exact.

Total error

- Total error (true error) = true solution – numerical solution
- True relative error

$$TrueRelativeError = \left|\frac{TrueSolution - NumericalSolution}{TrueSolution}\right|$$

- Cannot actually be determined in problems
- Useful for evaluating the accuracy of different numerical methods
  - By solving problems that can be solved analytically

1.4 Computers and programming

Computer program

- A set of instructions
- Machine language is required.
- Early days of computers $\Rightarrow$ low level computer languages (assembler)

Operating system

- Interface or layers enabling easier contact and communication between users and machine language of the computer
- UNIX developed by Bell Lab. in the 1970s
- DOS (Disk Operating System) used by Microsoft Inc.

High level computer languages

- FORTRAN, C, C++
- MATLAB (in this course)

Algorithm

- Before a numerical method is programmed, it is helpful to plan out all the steps that have to be followed in order to implement the numerical method successfully.
- Such a plan is called an algorithm!

- Commands for input and output of data.
  - Importing data into the computer/ displaying on the monitor/ storing numerical results in files
- Commands for defining variables
- Commands that execute mathematical operations
  - Standard operations: addition, multiplication, power, etc.
  - Common functions: trigonometric, exponential, logarithmic, etc.
- Commands for control
  - Conditional statements: if-else
- Commands for repetition
  - Loop statement: for