ESO 208A: Computational Methods in Engineering

Richa Ojha

Department of Civil Engineering

IIT Kanpur

Acknowledgements: Profs. Abhas Singh and Shivam Tripathi (CE)

- **Copyright:**

Recap

- Computational methods cannot be studied in isolation of the problem

"The purpose of computing is insight, not numbers", Hamming

- Significant digits/figures are the numbers that one can use with confidence

- True error = True value – Measured/Computed value
  - approximate error
  - error bound

True error is never known

Recap

- Types of error
  - Model error
  - **Data error**
  - **Truncation error**
  - **Round-off error**

Computers are finite

Round-off error

- Round-off error originates from the fact that computers retain only a fixed number of significant figures during a calculation

- In addition, because computers use a base-2 representation, they cannot precisely represent certain base-10 numbers.

Round-off error

208

$= 11010000$

$2^7\ 2^6\ 2^5\ 2^4\ 2^3\ 2^2\ 2^1\ 2^0$

$= 1 \times 2^7 + 1 \times 2^6 + \ldots$

$= 208$

$208.625 = 11010000.101$

$2^{-1}\ 2^{-2}\ 2^{-3}$

$= 1 \times 2^{-1} + 0 + 1 \times 2^{-3} = 0.625$

Round-off error

Number representation in computers

- Integer
- Fixed point
- Floating point

Round-off error

Integer-unsigned (0,1,2), signed (-1,-2,1,2,0)

For a 4-bit machine how many integers one can store?

1-bit: one space in binary computer

4-bit: nibble

8-bit: one byte

32-bit: word

64-bit: double word

Round-off error

For a 4-bit machine how many integers one can store?

You can store 2^4 =16 numbers

- 0-15 in case of unsigned numbers
- In case of signed, the first bit holds the sign. The remaining 3 bits can hold binary number from 000 to 111, i.e. in decimal numbers from 0 to 7
- The range should be from -7 to 7, but the range is -8 to 7.
- The +ve limit in generalized form can be obtained by 2^{n-1}-1, n is number of bits.

Round-off error

- Fixed point- $\pi = 3.14, \frac{1}{\sqrt{2}} = 0.71$
  - Location of decimal point is fixed (two places after decimal)
  - Useful for hand calculator
- Floating point

Round-off error

Floating point numbers

$$x = \pm m b^{p}$$

m - mantissa

b - base — 10

p - exponent

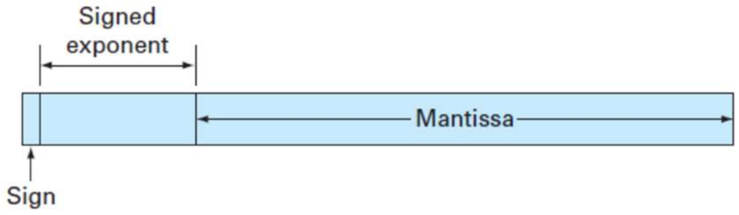
Example

$$\pi = 0.314 \times 10^{1}$$

$$1/\sqrt{2} = 0.707 \times 10^{0}$$

Round-off error

To store a floating point number, a computer word is divided into three parts



16-bit	1 bit	5 bit	10 bit
	sign	signed exponent	mantissa

IEEE 754 technical standard

32-bit	1	8 bit	23-bit	binary
64-bit	1	11 bit	52-bit	binary

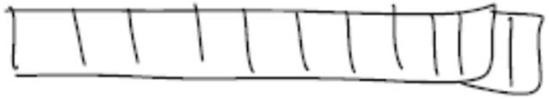
Round-off error

IEEE 754

64-bit	1 bit	11 bits	52 bits

IEEE754-r (2008)

11 bits



$\sim 2^{11}$ – 0 to 2047

Excess-p $\quad p = 2^{n-1} - 1 \quad$ n-bits

$p = 2^{10} - 1 = 1023$

-1023 to 1024

-1022 to 1024

1023

Round-off error

What we did so far was for binary, in case of decimal the maximum decimal power can be

$$m\,2^{1024} = \bar{m}\,10^{a}$$

$$a = \frac{\log(2)}{\log(10)}\,1024$$

$$= 308$$

$$10^{-308} \qquad 10^{308}$$

Summary

Number representation in Computers

- Integers
- Fixed Point
- Floating Point