Numerical Methods and Optimization

4/2/2024 – 3^{rd} Year

ENSIA — Spring 2024

Tutorial 1

- Number Representation and Accuracy -

Exercise 1: Binary/Floating-point representation

(1) Give the representation of the real numbers 234 and 12.625 in the binary system.

(2) Give the floating-point representation of the real number -3.625 in the binary system.

(3) What is the advantage of this representation?

Exercise 2: IEEE 754 Floating-Point Standard

(1) Represent the decimal numbers 3.14 and -8.625 using IEEE 754 Floating-Point Standard -Single Precision-

(2) Represent the decimal number 35.5 using IEEE 754 Floating-Point Standard -Double Precision-

(3) Give the binary representation in floating-point of the number A represented in IEEE 754 Floating-Point Standard SP:

$$A = 01000010001100110000000000000000$$

(4) Let A and B two numbers represented in IEEE 754 Floating-Point Standard SP,

a. Add the numbers A and B such that:

$$A = 01000000111000000000000000000000 \quad \text{and} \quad B = 01000000000100000000000000000000$$

b. Multiply the numbers A and B such that:

$$A = 01000000111000000000000000000000 \quad \text{and} \quad B = 01000001000110000000000000000000$$

(5) Determine the interval of the representable real numbers using IEEE 754 Floating-Point Standard SP

Exercise 3: Significant Digits

(1) Give the number of significant digits in the following numbers:

(a) 67.1 (b) 0.00173 (c) 2.30×10^{-9} (d) 3.0054 (e) 0.0004000

(2) Rewrite the following real numbers considering the number of significant digits given between parenthesis:

(a) 6243 (2) (b) 0.006738 (3) (c) 238.62 (3) (d) 6.345 (2)

(3) Give the results of the operations below if they are run on a computer storing the numbers in the decimal system with the following properties: $\epsilon_{machine} = 10^{-5}$, and $-20 < e < 20$ (e exponent).

(a) $1 + 10^{-6}$ (b) $1 + 10^{4}$ (c) $1 + 10^{6}$ (d) $10^{3} + 10^{6}$ (e) $10^{6}/10^{-12}$
(f) $10^{-9} \times 10^{-16}$ (g) $10^{28} + 10^{4}$ (h) $10^{5}/10^{-21}$ (i) $\sqrt{10^{10} - 10^{1}}$ (j) $\ln 10^{-25}$

Exercise 4: Errors and accuracy

(1) Evaluate the **Absolute True Error** and **Absolute Percent Relative Error** committed by approximating the number π by each of the following values:

(a) 3 (b) 3.14 (c) $\frac{22}{7}$

(2) The Area A of the earth is calculated using the following formula where r is the radius:

$$A = 4\pi r^2$$

a. Give 5 approximations that are used to calculate this area on a machine, and explain the impact of each approximation on the result.

b. If the radius $r = 6376$ is known with a precision of 2%, what is the absolute percent relative error on the area ?

Exercise 5: Rounding and Choping

(1)

a. Perform with two methods (using the distributivity of the multiplication on the addition) the operation $122 \times (133 + 695)$ using 3 significant digits and floating arithmetic with rounding and chopping.

b. What do you observe?

(2)

a. What (relative) error do we make when approximating $cos(1.57079)$ by $cos(1.57078)$?

b. Show that the absolute true error when approximating $cos(x)$ by $cos(x + h)$ where h is a small perturbation is given by $e_{abs} \simeq h$ for values of x close to $\pi/2$.

c. Conclude that the absolute percent relative error is $e_{rel} \simeq htan(x) \simeq \infty$. What do you conclude?