

Exercise session 1

Topics: finite arithmetic, cancellation error

1. Consider a computer which uses floating-point arithmetic with base $N = 10$, $t = 5$ digits for mantissa and rounding rule. Introduce the following numbers:

$$\begin{aligned} a &= 1.483593, \\ b &= 1.484111. \end{aligned}$$

Determine the result of $\bar{s} = \bar{a} \ominus \bar{b}$, where \bar{x} stands for machine-number corresponding to x in the used floating-point arithmetic and \ominus stands for the machine-subtraction. Compare \bar{s} with $c = a - b$ and then compute the relative error of \bar{s} .

2. Evaluate the functions:

$$\begin{aligned} f_1(x) &= \frac{1 - \cos(x)}{x^2}, & f_2(x) &= \frac{e^x - 1}{x}, \\ f_3(x) &= 1 - \sqrt{1 - x^2}, & f_4(x) &= \frac{(x + 1)^2 - 1}{x} \end{aligned}$$

in $x = 10^{-n}$, $n = 1, 2, \dots, 16$. After that, reformulate the functions in order to avoid cancellation error. Assuming the values obtained using reformulated functions as exact values, compute relative errors and compare them with machine precision. Print the relative error for every value of x and graphically represent it.

3. Consider the sequence

$$\begin{aligned} x_1 &= 2 \\ x_n &= 2^{n-1/2} \sqrt{1 - \sqrt{1 - 4^{1-n} x_{n-1}^2}}, \quad n \geq 2 \end{aligned}$$

which gives a π approximation. Compute the relative errors $|\pi - x_n|/\pi$ for $n = 1, \dots, 40$ and plot the relative errors behavior using logarithmic scale. After that, determine an expression \tilde{x}_n equivalent to x_n which avoids cancellation error. Graphically represent relative errors $|\pi - \tilde{x}_n|/\pi$ for $n = 1, \dots, 40$ using logarithmic scale. Compare this plot with the previous one and comment the results.

4. Given the function $f(x) = \sin(x)$, approximate $f'(x)$ using the following difference quotient

$$f'(x) \approx r(x) = \frac{\sin(x+h) - \sin(x)}{h},$$

with $h = 2^{-k}$, $k = 1, 2, \dots, 50$.

Fix for instance $x = \pi/4$; graphically represent the relative error $|f'(x) - r(x)|/|f'(x)|$ and comment the results. After that, reformulate $r(x)$ in order to avoid cancellation errors and graphically represent the relative error.