

NUMERISCHE METHODEN DER PHYSIK

SoSe 2021 – PROF. MARC WAGNER

LASSE MÜLLER: lmueller@itp.uni-frankfurt.de
LAURIN PANNULLO: pannullo@itp.uni-frankfurt.de

Exercise sheet 1

To be discussed in the week of 19.04.2021 - 23.04.2021

Exercise 1 [Floating-point numbers]

(1+1+1+1+2+1+1+2=10 pts.)

Let us assume we have a (rather primitive) computer that uses 8-bit floating-point arithmetic. The first bit represents the sign, the next 4 bits the exponent with bias $b = 7$ and the last 3 bits for the mantissa (normalized representation with leading 1 before the comma). With this we have the representation of a real number x as

$$x = (-1)^s \left[1 + \sum_{n=1}^3 m_n \cdot 2^{-n} \right] \cdot 2^{(\sum_{i=0}^3 e_{3-i} \cdot 2^{3-i}) - b}$$

which leads to a bit string $s e_3 e_2 e_1 e_0 m_1 m_2 m_3$. Assume that non-representable numbers are rounded to the nearest representable one (as usually it happens).

- Which number is represented by the bit-string 10111000?
- Which is the bit-string for the number -26 ? And for the number 0?
- How many different numbers can be exactly represented in this way? Which are the smallest and the largest positive ones?
- What is the result of the differences $(\frac{35}{32} - \frac{33}{32})$ and $(\frac{37}{32} - \frac{35}{32})$?
- Which number(s) have the largest absolute error? Which have the largest relative error in the interval between the smallest and the largest representable positive numbers?
- Repeat the task (iii) setting $b = 3$. Which role does the bias play? What happens varying it?
- How could you determine the smallest positive representable number on your computer? Try to write a simple program which prints to the screen the outcome using `single` and the `double` precision.
- Do you think it is a good idea, in a program, to check for equality between two floating-point numbers using the equality operator? When is it safe and when not? Which could be an alternative?

Exercise 2 [Golden ratio]

(3+2+5=10 pts.)

- It is well known, that the golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$ is the limit of the ratio of consecutive Fibonacci numbers $F(n-1)$ and $F(n)$. Write a short program to calculate

$$\delta(n) \equiv \frac{F(n)}{F(n-1)} - \varphi$$

and plot $\delta(n)$ as function of n . How does the plot changes using `single` and `double` precision? What happens for large n ?

- Prove that, for any value of n ,

$$\phi_{\pm}^{n+1} = \phi_{\pm}^{n-1} - \phi_{\pm}^n,$$

where

$$\phi_{\pm} = \frac{-1 \pm \sqrt{5}}{2}.$$

- Implement a short program to calculate the first 20 powers of ϕ_{\pm}

- (a) both using the iterative formula given above
- (b) and raising ϕ_{\pm} directly to the given power.

Repeat both strategies in **single** and **double** precision. Can you explain what happens?