Numerische Methoden der Physik

WiSe 2023-2024 – Prof. Marc Wagner

Michael Eichberg: eichberg@itp.uni-frankfurt.de
Lasse Müller: lmueller@itp.uni-frankfurt.de

Exercise sheet 1

Not to be handed in. To be discussed in the tutorials on 20.10.23 and 23.10.23

Exercise 1 [*Floating-point numbers*]

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

$$x = (-1)^{\mathtt{s}} \left[1 + \sum_{n=1}^{3} \mathtt{m}_n \cdot 2^{-n} \right] \cdot 2^{\left(\sum_{i=0}^{3} \mathtt{e}_{3-i} \cdot 2^{3-i} \right) - b}$$

which can be stored in a bit string $\mathtt{s}\,\mathtt{e}_3\mathtt{e}_2\mathtt{e}_1\mathtt{e}_0\,\mathtt{m}_1\mathtt{m}_2\mathtt{m}_3$. Assume that non-representable numbers are rounded to the nearest representable number (as usually it happens).

(i) Which number is represented by the bit-string `10111000`?

(ii) Which is the bit-string for the number -26? And for the number 0?

(iii) How many different numbers can be represented exactly in this way? Which is the smallest and which is the largest positive number?

(iv) What are the numerical results of the differences $\left(\frac{35}{32} - \frac{33}{32}\right)$ and $\left(\frac{37}{32} - \frac{35}{32}\right)$?

(v) 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 number?

(vi) Repeat (iii) setting $b = 3$. Which role does the bias play? What happens, when you vary the bias?

(vii) How could you determine the smallest positive representable number on your computer? Try to write a simple program which prints the result to the screen using `single` and the `double` precision.

(viii) 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? What could be an alternative?