Otto-von-Guericke-University Magdeburg
Max Planck Institute for Dynamics of Complex Technical Systems
Computational Methods in Systems and Control Theory

Dr. Jens Saak, Dipl.-Math. Martin Köhler
Website: http://www.mpi-magdeburg.mpg.de/csc/teaching/18ws/sc1/

Winterterm 2018/2019

Scientific Computing 1
7th Homework

Handout: 22nd Nov. 2018 **Return:** 30th Nov. 2018

Exercise 1: (4 Points)

Let a be a real number stored in IEEE double precision. Then a can be decomposed into

$$a_s + r_a, \tag{1}$$

where a_s is a truncated to single precision and the residual r_a is given by

$$r_a = a - a_s.$$

Thereby, r_a is computed in double precision and stored in single precision afterwards.

a.) In which range must a lie such that the expression (1) does not involve an infinite or NaN value?

b.) What is the maximum error between a and $a_s + r_a$ if the sum is evaluated in double precision?

Exercise 2: (4 Points)

Compute the forward error for the evaluation of the polynomial

$$P(x) = c_1 x + c_2 x^2$$

a.) using direct evaluation and

b.) using the Horner scheme.

Consider the case where x is close to a root of the polynomial and conclude which of those evaluation techniques is the more stable one.
In the IEE 754-2008 standard the fused-multiply-add operation $a \leftarrow a \pm (b \times c)$ is added to the set of basic instructions. What does that change (qualitatively and quantitatively) in the above considerations?

Exercise 3: (6 Points)

Determine the absolute and the relative condition numbers of

a.) $f(x) = \sin(x)$,

b.) $f(x) = \arctan(x)$,

c.) $f(x) = \sqrt{x \exp(x)}, \quad x > 0.$

Which values of x will lead to high condition numbers?

Exercise 4: (4 Points)

For all $x \in \mathbb{R}^n$ and a fixed $v \in \mathbb{R}^n$ we define the following mapping $f : \mathbb{R}^n \to \mathbb{R}$:

$$f(x) = \langle x, v \rangle = v^T x.$$

Determine the condition of this mapping. For which $x \in \mathbb{R}^n$ is the condition particularly small or particularly large?

Exercise 5: (4 Points)

Use a backward error analysis to determine numerical stability of:

a.) $f(x) = ax$ and

b.) $g(x) = a + x$.

Give conditions (where necessary) to guarantee stability. Assume that $a \in \mathbb{R}$ is fixed.

Exercise 6: (5 Points)

We want to compute the following two integrals

$$I_1 := \int_{-20}^{20} e^x \, \mathrm{d}x$$

and

$$I_2 := \int_{-20}^{20} e^{-x} \, \mathrm{d}x$$

using a C program. Create a naive implementation of the *midpoint rule*:

$$\int_a^b f(x) \, \mathrm{d}x \quad \approx \quad \sum_{i=0}^{n-1} h f\left(a + ih + \frac{1}{2}h\right),$$

where $h := \frac{b-a}{n}$, in **single precision** arithmetic.
The integrals I_1 and I_2 are now approximated by employing $n \in \{1024, 2048, 4096, 8192, 16384, 32768\}$ sampling points. Compare the results and prove them by computing the correct value of the integrals using their antiderivatives.
Analyze the reason behind the occuring errors and create a modified version of your implementation which avoids those errors **without** using double precision computations.
Compute the integral again with the sampling points $n \in \{1000, 2000, 4000, 8000, 16000, 32000\}$. What do you recognize in comparision to the previous results. Explain your observations.

Overall Points: 27