MAT801

# Numerik I – Assignment 1

Deadline: 01.03.2018, 12:00

**Exercise 1**

a) Let $x \in \mathbb{R}$ and denote by $\mathrm{rd}(x) \in \mathbb{R}(t,s)$ the symmetric rounding. Prove that

$$\left|\frac{x - \mathrm{rd}(x)}{x}\right| \leq 2^{-t}$$

holds. (3 points)

b) Suppose that, for $|x|$ small, one has an accurate value of $y = e^x - 1$ (obtained, e.g., by Taylor expansion). Use this value to compute accurately $\sinh(x) = \frac{e^x - e^{-x}}{2}$. (1 point)

**Exercise 2**

a) Let

$$y = \cos(x + \delta) - \cos(x). \tag{1}$$

Explain the difficulty of computing $y$ for small values of $\delta$. Find an alternative expression of (1) that does not exhibit these difficulties. (2 points)

b) Let $x, y, z \in \mathbb{R}(t,s)$. Use error analysis of first order (i.e. ignore quadratic and higher order error terms) to prove that the floating-point addition $\mathrm{fl}(\mathrm{fl}(x+y)+z)$ is more accurate than $\mathrm{fl}(x + \mathrm{fl}(y+z))$ if and only if $|x+y| < |y+z|$. (2 points)

**Exercise 3**

a) Write a program to compute

$$S_N = \sum_{i=1}^{N} \left[\frac{1}{i} - \frac{1}{i+1}\right] = \sum_{i=1}^{N} \frac{1}{i(i+1)} \tag{2}$$

once using the first summation and once using the (mathematically equivalent) second summation. For $N = 10^k$, $k = 1:7$, compute the respective absolute errors with respect to the value $\lim_{N\to\infty} S_N = 1$. Format your Matlab output so that the errors can be compared easily. Comment on the results. (2 points)

b) Sum the series

- $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^n}{(n!)^2}$,
- $\displaystyle\sum_{n=1}^{\infty} \frac{1}{(n!)^2}$

until there is no more change in the partial sums to within the machine precision. Generate the terms recursively. Print the number of terms required and the value of the sum. (2 points)

c) Let $f(x) = (n+1)x - 1$. The iteration

$$x_k = f(x_{k-1}), \quad k = 1, 2, \ldots, K, \quad x_0 = \frac{1}{n}$$

in exact arithmetic converges to the fixed point $1/n$ in one step. What happens in machine arithmetic? Run a program with $n = 1 : 5$ and $K = 10 : 10 : 50$ and explain what you observe. (2 points)