Prof. Ch. Schwab
B. Fitzpatrick
J. Zech

Spring Term 2016

# Numerische Mathematik I

ETH Zürich
D-MATH

## Homework Problem Sheet 1

For some problems, parts of the solution are already given. Fill in the gaps and complete the proofs where you see a red band at the left margin.
**Introduction.** Floating point arithmetic, rounding errors and SVD.

## Problem 1.1 Floating-Point Arithmetic

Every element $x \in \mathbb{F}$, where $\mathbb{F}$ denotes the (finite) set $\mathbb{F} = \mathbb{F}(\beta, t, e_{\min}, e_{\max})$ of floating point numbers with $\beta \in \mathbb{N}$, $\beta \geqslant 2$, $t \in \mathbb{N}$ and $e_{\min} \leqslant e_{\max} \in \mathbb{Z}$, is of the form

$$x = \pm\beta^e \sum_{i=1}^{t} \frac{d_i}{\beta^i}, \quad \text{where} \quad \begin{cases} \{d_i\}_{i=1}^t \subset \{0, 1, \ldots, \beta - 1\}, \\ x = 0 \iff d_1 = 0, \\ e \in \mathbb{Z} \cap [e_{\min}, e_{\max}]. \end{cases}$$

To approximate a real number $x \in \mathbb{R}$ by rounding to a floating point number $\mathrm{rd}(x) \in \mathbb{F}$, it is reasonable to take a "nearest" floating point number, i.e. a number $\mathrm{rd}(x) \in \mathbb{F}$ such that $|\mathrm{rd}(x) - x| = \min_{y\in\mathbb{F}} |y - x|$. If the latter minimum is not unique, i.e. if there are two $y \in \mathbb{F}$ minimizing $|y - x|$, $\mathrm{rd}(x)$ is defined as the one with $d_t$ *even*.

Now, consider $\mathbb{F} = \mathbb{F}(2, 3, -1, 1)$.

**(1.1a)** Determine all numbers in $\mathbb{F}$. Do that first by hand, and then write a MATLAB function that handles it.

**(1.1b)** Mark $\mathbb{F}$ on the real number line. You may want to do this in MATLAB.

**(1.1c)** Sketch the graphs of the functions

$$\mathrm{rd} : [x_{\min}; x_{\max}] \to \mathbb{F},\ x \mapsto \mathrm{rd}(x) \quad \text{and} \quad \mathrm{err} : [x_{\min}; x_{\max}] \to \mathbb{R},\ x \mapsto |\mathrm{rd}(x) - x|,$$

where $x_{\min} := \min\{x \in \mathbb{F} \mid x > 0\}$ and $x_{\max} := \max \mathbb{F}$.

Listing 1.1: Testcalls for Problem 1.1

```
F = ComputeF(2,3,-1,1)
```

Listing 1.2: Output for Testcalls for Problem 1.1

```
>> test_call

F =

```

```
    0.2500
    0.3750
    0.3125
    0.4375
    0.5000
    0.7500
    0.6250
    0.8750
    1.0000
    1.5000
    1.2500
    1.7500
```

## Problem 1.2 Round-off Error Analysis

This problem considers asymptotic round-off analysis as presented in [NMI, Sect. 1.3] and [NMI, Sect. 1.4]. The attribute "asymptotic" indicates that *you may assume all relative errors $\delta$ introduced by elementary operations to be so small that you can use linearization* (Taylor expansion) around zero and subsequently drop all "second order terms" of size $O(\delta^2)$.

Let $|x| < 1$, the MATLAB functions `asin(x)` and `atan(x)` compute $\arcsin(x)$ and $\arctan(x)$ respectively, with relative error $\leq u(\mathbb{F})$. It holds

$$f(x) := \arctan(x) = \arcsin\left(\frac{x}{\sqrt{1+x^2}}\right) =: g(x). \tag{1.2.1}$$

**(1.2a)** Implement a MATLAB routine that computes and print the values of the relative error

$$\left|\frac{g(x) - f(x)}{f(x)}\right|$$

with respect to the `atan`-function, for $x = 10^{-5}, 10^{-4}, \ldots, 1$ and for $x = 10^6, 10^7, \ldots, 10^{11}$. For which values of $x$ is formula (1.2.1) unstable?

**(1.2b)** Gauge the propagation of round-off errors introduced by the division in $f(x)$. Compute the relative error of

$$\widetilde{f}(x) = \arcsin\left(\frac{x}{\sqrt{1+x^2}}(1+\delta)\right)$$

with respect to $f(x)$. When is the error large for small values of $\delta$?

HINT: Use Taylor expansions.

**Solution:** The Taylor expansion of $\widetilde{f}(x + x\delta) = \widetilde{f}(x) + x\delta\widetilde{f}'(x) + \mathcal{O}(\delta^2)$ reads

Neglecting the terms $\mathcal{O}(\delta^2)$, the relative error is

**(1.2c)** Analyze the propagation of round-off errors in floating-point arithmetic by performing a complete round-off analysis of (1.2.1) as in [NMI, Sect. 1.4].

**Solution:** Let us denote by $\delta_1$, $\delta_2$, $\delta_3$ and $\delta_4$ the errors generated by the `arcsin` function, division, extraction of square root and square power operation respectively. Let $|\delta_i| \leq u(\mathbb{F})$ and

$$\widetilde{g}(x) = \arcsin\left(\frac{x}{\sqrt{1+x^2(1+\delta_4)}(1+\delta_3)}(1+\delta_2)\right)(1+\delta_1). \tag{1.2.2}$$

Using Taylor expansion and omitting higher order terms, one has

$$\sqrt{1+x^2(1+\delta_4)} =$$

Moreover, exploiting the fact that $\frac{1}{1+\eta} \approx 1-\eta$ for $|\eta| \leq u(\mathbb{F})$, yields

$$\frac{x}{\sqrt{1+x^2(1+\delta_4)}(1+\delta_3)}(1+\delta_2) =$$

The Taylor expansion of $\arcsin(x(1+\delta))$ reads

$$\arcsin(x+x\delta) = \arcsin x + \frac{x\delta}{\sqrt{1-x^2}} = \arcsin x\left(1+\frac{x\delta}{\arcsin x\sqrt{1-x^2}}\right).$$

Substituting the Taylor expansions of the single terms into (1.2.2), results in

## Problem 1.3 Summing the Harmonic Series

In analysis you have seen that the harmonic series diverges. On a computer this will not happen, of course!

The series $\sum_{k=1}^{+\infty} k^{-1}$ is called the harmonic series. The partial sums, $S_n = \sum_{k=1}^{n} k^{-1}$, can be computed recursively by setting $S_1 = 1$ and using $S_n = S_{n-1} + n^{-1}$. If this computation were carried out on your computer, what is the largest $S_n$ that would be obtained (approximately)? What is the according $n$ (approximately)? (Do not do this experimentally on the computer; it is too expensive.)

HINT: Find $n$ such that $|\frac{S_n - S_{n-1}}{S_n}| < u(\mathbb{F})$, where $u(\mathbb{F})$ is the unit round-off of the floating-point number system $\mathbb{F}$. To this end, first prove that $\sum_{k=1}^{n} \frac{1}{k} > \ln(n)$.

**Solution:** We show $\sum_{k=1}^{n} \frac{1}{k} > \ln(n+1) > \ln(n)$:

## Problem 1.4 Singular Value Decomposition and Matrix Norms

Let $\mathbf{A}$ be given by

$$\mathbf{A} := \begin{pmatrix} 5 & 3 \\ 0 & -4 \end{pmatrix}.$$

**(1.4a)** Find orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ in $\mathbb{R}^{2\times 2}$ and a diagonal Matrix $\mathbf{\Sigma}$ such that $A = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.

**Solution:** The matrices $\mathbf{U}$ and $\mathbf{V}$ result – up to the choice of sign – from the eigenvalue decomposition of the matrices $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$. We get

**(1.4b)** Compute the operator-2-norm, the Frobenius norm and the spectral radius of both $\mathbf{A}$ and $\mathbf{A}^{-1}$. For this, refer to [NMI, Prop 0.51] from the lecture notes.

**Solution:** The upper triangular matrix $\mathbf{A}$ obviously has eigenvalues $5$ and $-4$, so the spectrum is $\sigma(\mathbf{A}) = \{5, -4\}$.

## Problem 1.5 The Butterfly Effect

Let the function $f : \mathbb{R} \to \mathbb{R}$ be given by $x \mapsto \frac{1}{5}x^5 - \frac{2}{3}x^3 + x$.

**(1.5a)** Determine the sequence $x^{(n)}$, $n = 0, 1, \ldots$, of real numbers defined by

$$x^{(0)} := \sqrt{\frac{25 + 2\sqrt{55}}{27}} \quad \text{and} \quad x^{(n+1)} := \Psi(x^{(n)}), \quad \text{where} \quad \Psi(x) := x - \frac{f(x)}{\frac{\mathrm{d}f}{\mathrm{d}x}}. \tag{1.5.1}$$

**Remark:** The iteration $x \mapsto \Psi^{(n)}(x)$ is known as *Newton's method* and is used to find zeros of $f$.

**(1.5b)** Calculate the first $50$ terms of the sequence $x^{(n)}$ in MATLAB and plot them. How does the behaviour of the calculated sequence differ from the behaviour of the analytically analized sequence from subproblem (1.5a)? Why?

Published on February 26, 2016.

# References

[NMI] Lecture Notes for the course "Numerische Mathematik I".

Last modified on February 25, 2016