Stephan Stofer
Adrian Willi

Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN
Information Technology
FH Zentralschweiz

Applied Numerics - Exercise 1

Prof. Dr. Josef F. Bürgler

I.BA_IMATH, Semesterweek 08

The solution of the exercises should be presented in a clear and concise manner. Numerical results should be accurate to 4 digits. The exercises are accepted if You solve 75% of the exercises adequately. Please hand in the exercises no later than at the end of the last lecture in semesterweek 9.

1 Number representation (optional)

Review the slides on floating point numbers and show, that indeed $0 \leq m \leq b(1-b^{-p})$. Use this to find the bounds on m if $b = 10$ (decimal system) and $p = 5$ (5 significant digits).

2 Numerical Errors: Cancellation I

Consider the two equivalent expressions:

$$f_1(x) = \frac{(1+x)-1}{x}$$
$$f_2(x) = 1$$

Plot the error $\text{err} = f_2(x) - f_1(x)$ for $-1 \times 10^{-7} \leq x \leq 1 \times 10^{7}$ using single precision arithmetic.

3 Numerical Errors: Cancellation II (optional)

The fact that an error builds up in addition and subtraction depends on absolute accuracy, rather than relative accuracy, leads to a particular problem. Suppose we calculate $\sqrt{10} - \pi$ using single precision on an IBM System/370 ($\varepsilon_{\text{mach}} = 10^{-6}$):

$$\text{rd}(\sqrt{10}) = 3.16228$$
$$\text{rd}(\pi) = 3.14159$$
$$\text{rd}(\sqrt{10} - \pi) = 0.02069$$

The absolute and relative error of the subtraction are

$$\varepsilon_{\text{abs}} = \varepsilon_1 - \varepsilon_2 \approx 10^{-6} \quad \text{and} \quad \varepsilon_{\text{rel}} = \frac{\varepsilon_1 - \varepsilon_2}{\sqrt{10} - \pi} \approx \frac{10^{-6}}{10^{-2}} = 10^{-4} \quad \text{respectively,}$$

Aufgabe 1 (optional)

$b = 10 \quad p = 5$

$0 \leq m \leq b(1-b^{-p}) = 0 \leq m \leq 10(1-10^{-5}) = 9.9999$

Aufgabe 2

$f_1(x) = \frac{(1+x)-1}{x}$

$f_2(x) = 1$

$\Rightarrow$ z.B. $x = 1$: $1 - \left(\frac{(1+1)-1}{1}\right) = 1 - 1 = 0$

sollte in jedem Fall eigentlich 0 geben

Fehler plotten: $f_2(x) - f_1(x)$ für $-1 \times 10^{-7} \leq x \leq 1 \times 10^{7}$

Aufgabe 3 (optional)

$f(x) = x\left[\sqrt{x+1} - \sqrt{x}\right]$

x_0	Computed $f(x)$	True $f(x)$
1	0.414210	0.414214
10	1.54340	1.54347
100	4.99000	4.98756
1'000	15.8000	15.8074
10'000	50.0000	49.9988
100'000	100.000	158.113

Absolute error : ε_{abs}

$\varepsilon_{abs}(x_{100}) = 4.99000 - 4.98756$

$\stackrel{!}{=} 0.00244$

Grund:

$x = 100 \Rightarrow \sqrt{100} = 10.0000$ (genau) $\quad \sqrt{101} = 10.04999$ (gerundet)

$\Rightarrow \sqrt{x+1} - \sqrt{x} = \sqrt{100} - \sqrt{101} = 0.049900$, während das korrekte Wert 0.04998756 sein sollte

Relatives Fehler : ε_{rel}

$\varepsilon_{rel}(x_{100}) = \frac{0.00244}{4.98756} = 0.0004892$

kann wie folgt umgeformt werden, um loss of significance error zu vermeiden

$f(x) = x\left[\sqrt{x+1} - \sqrt{x}\right] \Rightarrow \frac{x}{\sqrt{x+1} - \sqrt{x}}$

where $\varepsilon_1 = \mathrm{rd}(\sqrt{10}) - \sqrt{10}$ and $\varepsilon_2 = \mathrm{rd}(\pi) - \pi$.

This problem is known as **loss of significance** or **canellation**. It can occur whenever two similar numbers of equal sign are subtracted (or two similar numbers of opposite sign are added), and is a major cause of inaccuracy in floating-point algorithms.

Find another example and calculate the absolute and relative error for Your example!

(siehe Seite vorher)

4 Numerical Errors: Cancellation III

A (not so efficient) possibility to compute π is to use a series (U_k) where U_k represents the circumference of a regular 2^k-gon, i.e. a regular polygon with 2^k vertices. It is possible to show that the following iteration formula holds [1]

$$U_{k+1} = 2^k \sqrt{2\left(1 - \sqrt{1 - (2^{-k}U_k)^2}\right)},\ k \geq 2.$$

Programm this forumula and compute the limit $\lim_{k\to\infty} U_k$ using

(i) the `float` data type, and

(i) the `double` data type.

You will notice that the sequence first starts to approach π. After some more iterations it starts to diverge from π. The reason is **cancellation**. Find the part in the formula, where cancellation occurs. Notice that at some point in the formula one number is subtracted from a number which is almost identical.

Find a way to avoid cancellation, by appropriate expansion of the critical term. Notice the fact

$$1 - \sqrt{1-a} = \left(1 - \sqrt{1-a}\right) \cdot \frac{1+\sqrt{1-a}}{1+\sqrt{1-a}} = \frac{1-(1-a)}{1+\sqrt{1-a}} = \frac{a}{1+\sqrt{1-a}}$$

As You can see, the numerator is no longer susceptible to cancellation. Improve Your programm appropriately and comment the convergence behaviour.

5 Direct Method: LU-Factorization

On the slides we computed the 3×3-matrices $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$. Show, that the $(3,1)$-Element of matrix product $\boldsymbol{L}_2\boldsymbol{L}_1$ is equal to $l_{2,1}l_{3,2} - l_{3,1}$. Compute $\boldsymbol{L}_2\boldsymbol{L}_1$ and multiply it with the matrix

$$(\boldsymbol{L}_2\boldsymbol{L}_1)^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ l_{2,1} & 1 & 0 \\ l_{3,1} & l_{3,2} & 1 \end{bmatrix}$$

to obtain the identity matrix.

[1] See the handwritten notes at the end of this ducument.

Aufgabe 4

$$u_{k+1} = 2^k \sqrt{2\left(1-\sqrt{1-(2^{-k}u_k)^2}\right)}, \quad k \geq 2$$

Dieser Teil verursacht den Auslöschungsfehler

$$\Rightarrow 1-\sqrt{1-(2^{-k}u_k)^2} = \left(1-\sqrt{1-(2^{-k}\cdot u_k)^2}\right)\cdot\frac{1+\sqrt{1-(2^{-k}\cdot u_k)^2}}{1+\sqrt{1-(2^{-k}\cdot u_k)^2}} = \frac{1-\left(1-(2^{-k}\cdot u_k)^2\right)}{1+\sqrt{1-(2^{-k}\cdot u_k)^2}}$$

$$= \frac{(2^{-k}\cdot u_k)^2}{1+\sqrt{1-(2^{-k}\cdot u_k)^2}}$$

→ siehe Octave Skript

$$u_{k+1} = 2^k \sqrt{\frac{2(2^{-k}u_k)^2}{1+\sqrt{1-(2^{-k}u_k)^2}}}$$

```
rel. error for bad (left) and good (right) method
2.55046415955671e-02 2.55046415955671e-02
6.41314885579401e-03 6.41314885579401e-03
1.60560696438116e-03 1.60560696438144e-03
4.01546850325250e-04 4.01546850320727e-04
1.00395783863265e-04 1.00395783858035e-04
2.50995129994164e-05 2.50995129497997e-05
6.27491342261881e-06 6.27491367437739e-06
1.56872974177577e-06 1.56873063332059e-06
3.92184886379381e-07 3.92182796684276e-07
9.80326701348349e-08 9.80457077232272e-08
2.45241179447573e-08 2.45114273902202e-08
6.40647820848329e-09 6.12785681221557e-09
3.87636533591722e-10 1.53196420305389e-09
2.63197006556094e-09 3.82991086102969e-10
1.47103964621716e-08 9.57477008467494e-11
8.19170108940481e-08 2.39368545326944e-11
4.68426546739940e-07 5.98410761468422e-12
1.24144517018555e-06 1.49585020618875e-12
1.24144517018555e-06 3.73750514568424e-13
2.59777256198812e-05 9.31549126704203e-14
7.54484510070300e-05 2.28999937065373e-14
2.73306885715936e-04 5.37160346202727e-15
2.73306885715936e-04 9.89505900899761e-16
```

↖ Fehler wird grösser

↖ Fehler wird kleines

Aufgabe 5

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -l_{3,2} & 1 \end{bmatrix} \quad L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -l_{2,1} & 1 & 0 \\ -l_{3,1} & 0 & 1 \end{bmatrix}$$

$$L_2L_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -l_{3,2} & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ -l_{2,1} & 1 & 0 \\ -l_{3,1} & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -l_{2,1} & 1 & 0 \\ l_{3,2}\cdot l_{2,1} - l_{3,1} & -l_{3,2} & 1 \end{bmatrix}$$

$$L_2L_1 \cdot (L_2L_1)^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -l_{2,1} & 1 & 0 \\ l_{3,2}\cdot l_{2,1} - l_{3,1} & -l_{3,2} & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ l_{2,1} & 1 & 0 \\ l_{3,1} & l_{3,2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \underline{\underline{I}}$$

6 Direct Method: Gaussian Elimination

Using Your favorite programming language write code to solve a system of linear equations of n equations in n unknowns. The coefficient matrix as well as the (possibly multiple) right hand sides are stored in a file in the following format:

```
3
2 -1 3
-4 6 -5
6 13 16
1
2
1
3
```

The first line defines n, then follow the n rows of the coefficient matrix $\boldsymbol{A}$, the number of right hand sides m and finally the m (here only one) right hand sides $\mathbf{b}$.

Make sure that You check the solution $\mathbf{x}$ by computing the $\|\boldsymbol{A}\mathbf{x}-\mathbf{b}\|$ (this should be a very small number). Test Your programm with several random coefficient matrices $\boldsymbol{A}$ and right hand sides $\mathbf{b}$.

7 Iterative Method: Gauss-Seidel I (optional) - ToDo

Write code (in any programming language) to solve systems of n equations in n unknowns. As a numerical example solve the following system:

$$\begin{bmatrix} 12 & 7 & 3 \\ 1 & 5 & 1 \\ 2 & 7 & -11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ -5 \\ 6 \end{bmatrix}$$

8 Iterative Method: Gauss-Seidel II (optional) - ToDo

Rewrite the following system of equations, such that the Gauss-Seidel method converges (diagonally dominant coefficient matrix):

$$\begin{aligned} 2x_1 + 7x_2 - 11x_3 &= 6 \\ x_1 + 2x_2 + x_3 &= -5 \\ 7x_1 + 5x_2 + 3x_3 &= 17 \end{aligned}$$

Have Fun!

Aufgabe 6

```
result =

   14.0833
    3.1667
   -7.6667

ans =

    0
    0
    0
```

A*result - b