Linnaeus University
Department of Mathematics
Sofia Eriksson

Written Exam on Numerical Methods, 1MA930/1MA931, 3 hp (7.5 hp/5 hp)
Thursday 1st of June 2023, 08.00–13.00.

1. (a) Which ones of the following numbers (i-iv) can be represented exactly in floating point arithmetics (IEEE double precision)? If a number can not be represented exactly, state to what value it is rounded by the computer:

   (i) $(1 + 2^{-58}) - 1$
   (ii) $(1 + 2^{-17}) - 1$
   (iii) 2^{-58}
   (iv) 2^{-17}

   (b) In (v-vi), identify for which values of x there is subtraction of nearly equal numbers, and find an alternate form that avoids the problem:

   (v) $\frac{\sqrt{1+x}-\sqrt{1-x}}{x}$
   (vi) $\frac{1-\cos(x)}{1+\cos(x)}$ (2p+3p)

   Suggested solution:

   (a)

   (i) $(1 + 2^{-58}) - 1 = 0$ can not be represented exactly since $2^{-58} < 2^{-52} = \varepsilon_{mach}$
   (ii) $(1 + 2^{-17}) - 1 = 2^{-17}$ can be represented exactly since $2^{-17} > 2^{-52} = \varepsilon_{mach}$
   (iii) 2^{-58} can be represented exactly since $2^{-58} > 2^{-1023} = "underflow"*$
   (iv) 2^{-17} can be represented exactly since $2^{-17} > 2^{-1023} = "underflow"*$

   * actually even smaller numbers can be represented if we consider subnormal representation

   (b)

   (v) $\frac{\sqrt{1+x}-\sqrt{1-x}}{x}$ has problems with subtraction of nearly equal numbers"/loss of significanceåhen $x \approx 0$. The problem is easiest handled as

$$\frac{\sqrt{1+x}-\sqrt{1-x}}{x} = \frac{(\sqrt{1+x}-\sqrt{1-x})(\sqrt{1+x}+\sqrt{1-x})}{x(\sqrt{1+x}+\sqrt{1-x})} = \frac{(1+x)-(1-x)}{x(\sqrt{1+x}+\sqrt{1-x})}$$
$$= \frac{2x}{x(\sqrt{1+x}+\sqrt{1-x})} = \frac{2}{(\sqrt{1+x}+\sqrt{1-x})}$$

   Alt: Can also be solved using Taylor expansions, leading to $\approx 1 + \frac{x^2}{8} + \mathcal{O}(x^4)$

   (vi) $\frac{1-\cos(x)}{1+\cos(x)}$ has problems in the numerator for $x \approx 2\pi n$, $n \in \mathbb{Z}$ and denominator for $x \approx \pi(2m+1)$, $m \in \mathbb{Z}$. Since the task is only worth 1p, it is enough to identify the problem at $x \approx 0$ as long as the solutionfits the problem identified. For example, for $x = 2\pi n$ (wherein $x \approx 0$)

$$\frac{1-\cos(x)}{1+\cos(x)} = \frac{(1-\cos(x))(1+\cos(x))}{(1+\cos(x))^2} = \frac{1-\cos^2(x)}{(1+\cos(x))^2} = \frac{\sin^2(x)}{(1+\cos(x))^2}$$

   Alt. Taylor

2. (a) Use the Newton-Raphson method to find approximations of all solutions to the equation $x^3 + 18x^2 - 39x + 11 = 0$. Answer with 6 correct decimals.

(b) If you instead would use the Bisection method to find one of the roots, and would start with an initial interval $[a, b]$ of width 1 (that is having $b = a + 1$): How many iterations would you have to do to obtain an answer with 6 correct decimals? (4p+1p)

Suggested solution:

(a) The Newton-Raphson method is $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$,
where we identify $f(x) = x^3 + 18x^2 - 39x + 11$ and compute $f'(x) = 3x^2 + 36x - 39$. Since it is a third degree polynomial, we expect it to have either one or three real roots. Either finding starting guesses graphically or by finding the positions of the max and min (at $x = -13$ and $x = 1$) we can get starting guesses. We find that there are roots close to -20, close to 0 and close to 2. Start iterating, to ensure 6 correct decimals, use as stopping criterion that $|\Delta x| = |x_k - x_{k-1}| < 0,5 \cdot 10^{-6}$

$$\begin{aligned}
x_0 &= -20 \\
x_1 &= -19,97959184 & |\Delta x| &= 0,020408163 > 0,5 \cdot 10^{-6} \\
x_2 &= -19,97955204 & |\Delta x| &= 3,98013 \cdot 10^{-5} > 0,5 \cdot 10^{-6} \\
x_3 &= -19,97955204 & |\Delta x| &= 1,51243 \cdot 10^{-10} < 0,5 \cdot 10^{-6}
\end{aligned}$$

answer: -19,979552

$$\begin{aligned}
x_0 &= 0 \\
x_1 &= 0,282051282 & |\Delta x| &= 0,282051282 > 0,5 \cdot 10^{-6} \\
x_2 &= 0,332890779 & |\Delta x| &= 0,050839497 > 0,5 \cdot 10^{-6} \\
x_3 &= 0,334721207 & |\Delta x| &= 0,001830428 > 0,5 \cdot 10^{-6} \\
x_4 &= 0,334723599 & |\Delta x| &= 2,392 \cdot 10^{-6} > 0,5 \cdot 10^{-6} \\
x_5 &= 0,334723599 & |\Delta x| &= 4,08568 \cdot 10^{-12} < 0,5 \cdot 10^{-6}
\end{aligned}$$

answer: 0,334724

$$\begin{aligned}
x_0 &= 2 \\
x_1 &= 1,711111111 & |\Delta x| &= 0,288888889 > 0,5 \cdot 10^{-6} \\
x_2 &= 1,648057584 & |\Delta x| &= 0,063053527 > 0,5 \cdot 10^{-6} \\
x_3 &= 1,644836836 & |\Delta x| &= 0,003220748 > 0,5 \cdot 10^{-6} \\
x_4 &= 1,644828436 & |\Delta x| &= 8,39981 \cdot 10^{-6} > 0,5 \cdot 10^{-6} \\
x_5 &= 1,644828436 & |\Delta x| &= 5,71188 \cdot 10^{-11} < 0,5 \cdot 10^{-6}
\end{aligned}$$

answer: 1,644828

(b)

The error is halved each iteration when using Bisection. Hence $error \approx (b-a)/2^{n+1}$ where n is the number of iterations and $b - a$ the size of the starting interval. We have $b - a = 1$ and we want $error < 0,5 \cdot 10^{-6}$. This leads to $1/2^{n+1} < 0,5 \cdot 10^{-6} \iff 2^{-n} < 10^{-6} \iff 2^n > 10^6 \iff n\log(2) > 6\log(10) \iff n > 6\log(10)/\log(2) \approx 19,93156857$ that is we need 20 iterations.

3. Consider an n-by-n system of linear equations where the coefficient matrix is in upper triangular form, as visualised below for a 4-by-4 system.

$$\left(\begin{array}{cccc|c} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{array}\right)$$

(a) Derive the number of operations needed to solve this n-by-n system (for an arbitrary positive integer n) using back-substitution. (*By operations we mean addition, subtraction, multiplication and division*)

(b) Explain why and in what situations it is advantageous to use LU-factorisation. (3p+2p)

Suggested solution: (a) To solve for the last unknown in row n, we need 1 division, as in $x_n = b_n/a_{nn}$.

To solve for the second last unknown in row $n-1$: we need 1 multiplication, 1 subtraction and 1 division, in total 3 operations, from $x_{n-1} = (b_{n-1} - a_{n-1,n}x_n)/a_{n-1,n-1}$.

For $n-2$ we need 5 operations=1 division, 2 multiplications and 2 subtractions.

In total we need $1 + 3 + 5 + \ldots + (2n-1) = \sum_{k=1}^{n}(2k-1)$.

We can compute $\sum_{k=1}^{n}(2k-1) = 2\sum_{k=1}^{n} k - \sum_{k=1}^{n} 1 = 2\frac{n(n+1)}{2} - n = n^2$

(b) It is good when solving a system $Ax = b$ repeatedly. To construct L and U using LU-factorization has complexity $\mathcal{O}(n^3)$ just as Gaussian elimination, however solving the resulting systems $Lz = b$ and $Ux = z$ has complexity $\mathcal{O}(n^2)$. If one wants to solve a system $Ax = b$ many times with different right-hand-sides b one can do the factorization once, and then solving the fast systems many times.

4. (a) Use the (composite) Simpson's method to compute the integral $I = \int_0^1 e^x \, \mathrm{d}x$ for two different step sizes h ($h = 1/2$ and $h = 1/4$). Then use Richardson extrapolation on the two results to further improve the approximation of the integral.

   (b) Show that the finite difference formula $D_h = \frac{f(x+h)-f(x-h)}{2h}$ is second order accurate.

   (3p+2p)

   Suggested solution:

   (a) With $h = 1/2$:

   $S_1 = \frac{h}{3}\left(e^0 + 4e^{0,5} + e^1\right) \approx 1,718861152$

   With $h = 1/4$:

   $S_2 = \frac{h}{3}\left(e^0 + 4e^{0,25} + 2e^{0,5} + 4e^{0,75} + e^1\right) \approx 1,718318842$

   Richardson extrapolation. We use that Simpsons is fourth order accurate:

$$RE = \frac{2^4 S_2 - S_1}{2^4 - 1} = \frac{16 \cdot 1,718318842 - 1,718861152}{15} = 1,718282688$$

   (b) We use Taylor expansion around x:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \ldots$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x) + \ldots$$

   Inserting these into the finite difference formula, we obtain

$$\begin{aligned} D_h &= \frac{f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \ldots - \left(f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x) + \ldots\right)}{2h} \\ &= \frac{2hf'(x) + 2\frac{h^3}{6}f'''(x) + \ldots}{2h} = f'(x) + \frac{h^2}{6}f'''(x) + \ldots \end{aligned}$$

   (Alternatively one includes more terms or is more detailed regarding rest terms.)

5. The differential equation $y' = -2y(1 + 4\cos(4t))$ with initial condition $y(0) = 1$ is solved using some of the numerical solvers mentioned in this course. (a) The figure [above] shows the resulting errors at time $t = 1.3$ (obtained using two different explicit methods). Given that information, answer the following:

   - What methods do "Method1" and "Method2" refer to? Motivate.
   - Explain the errors of "Method2" in terms of why the curve looks as it does for *fewer* time steps ($N \lessapprox 10^4$) and *more* time steps ($N \gtrapprox 10^4$), respectively.
   - Predict the behaviour of the curves as N increases even more, either by drawing a picture or by giving a crude guess for how large you think the errors produced by the two methods will be when the number of time steps are around $N \approx 10^{11}$.

   (b) Find the numerical solution to the above-mentioned initial value problem, at time $t = 0.1$. Use the Euler Backward method with time step 0.05. (3p+2p)

   Suggested solution:

   (a) The figure above shows the resulting errors at time $t = 1.3$ (obtained using two different explicit methods). Given that information, answer the following:

   - "Method1" is first order, and since it is mentioned that the method is *explicit* it must be Euler Forward.

     "Method2" is fourth order, the only fourth order method we have mentioned is RK4 (the classical fourth order Runge-Kutta)
   - For *fewer* time steps the truncation errors are dominating.

     For *more* time steps rounding errors are dominating.
   - Both methods will suffer from rounding errors, of about equal size. It is not possible to know exactly from the graph, but reasonable is anything in the range 10^{-6} up to 10^{-4}, most likely around 10^{-5} for both methods (best is to draw a picture, showing the $\mathcal{O}(1/h)$ behaviour we studied in the laborations).

   (b) Euler Backward: $y_{n+1} = y_n + hf(t_{n+1}, y_{n+1})$. We identify $f(t, y) = -2y(1 + 4\cos(4t))$

   We need to solve for y_{n+1} as

$$
\begin{aligned}
y_{n+1} &= y_n - 2hy_{n+1}(1 + 4\cos(4t_{n+1})) && \iff \\
y_{n+1} + 2hy_{n+1}(1 + 4\cos(4t_{n+1})) &= y_n && \iff \\
y_{n+1} &= \frac{y_n}{1 + 2h(1 + 4\cos(4t_{n+1}))}
\end{aligned}
$$

   We use the initial value $y_0 = y(0) = 1$ and note that $t_0 = 0$, $t_1 = h = 0.05$ and $t_2 = 2h = 0.1$.

   We get

$$
\begin{aligned}
y_1 &= \frac{y_0}{1 + 2h(1 + 4\cos(4t_1))} = \frac{1}{1 + 0.1(1 + 4\cos(0.2))} \approx 0,670229324 \\
y_2 &= \frac{y_1}{1 + 2h(1 + 4\cos(4t_2))} \approx \frac{0,670229324}{1 + 0.1(1 + 4\cos(0.4))} \approx 0,456427532
\end{aligned}
$$

   answer: $y(0,1) \approx y_2 = 0,456427532$.

6. Consider the boundary value problem (BVP)

$$\begin{aligned} \frac{d^2y}{dx^2} &= \frac{x+y}{25}, \qquad\qquad x \in [0,\ 20] \\ y(x=0) &= 1 \\ y(x=20) &= -7 \end{aligned}$$

(a) Approximate the boundary value problem described above as a finite difference problem with step size $\Delta x = h = 5$, and present the resulting system of equations in matrix form. *You don't need to solve the system!*

(b) Rewrite the BVP such that it could be solved using the shooting method. (3p+2p)

Suggested solution: (a) Let $x_{0,1,2,3,4} = 0, 5, 10, 15, 20$.

Let $\frac{d^2y}{dx^2}$ be approximated by the finite difference scheme $\frac{y(x+h)-2y(x)+y(x-h)}{h^2}$ and then let $y(x_i)$ be approximated by w_i (such that $y(x_i+h) = y(x_{i+1})$ is approximated by w_{i+1}).

At the boundaries we use $w_0 = y(0) = 1$ and $w_4 = y(20) = -7$.

For the interior points we have

$$\frac{w_{i+1} - 2w_i + w_{i-1}}{h^2} = \frac{x_i + w_i}{25}.$$

Use that $h^2 = 5^2 = 25$ such that we can simplify to

$$w_{i+1} - 2w_i + w_{i-1} = x_i + w_i \quad \Longleftrightarrow \quad w_{i+1} - 3w_i + w_{i-1} = x_i$$

leading to

$$\begin{aligned} i=1: &\quad w_2 - 3w_1 + w_0 = x_1 &\Longleftrightarrow&\quad w_2 - 3w_1 = x_1 - w_0 = 5 - 1 = 4 \\ i=2: &\quad w_3 - 3w_2 + w_1 = x_2 &\Longleftrightarrow&\quad w_3 - 3w_2 + w_1 = 10 \\ i=3: &\quad w_4 - 3w_3 + w_2 = x_3 &\Longleftrightarrow&\quad -3w_3 + w_2 = x_3 - w_4 = 15 + 7 = 22 \end{aligned}$$

written as a system on matrix form:

$$\begin{pmatrix} -3 & 1 & 0 \\ 1 & -3 & 1 \\ 0 & 1 & -3 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 10 \\ 22 \end{pmatrix}$$

(b) Rewrite as first order system by introducing $u = y'$ such that $u' = y''$. This gives $u' = y'' = \frac{x+y}{25}$.

For the boundary conditions we keep $y(0) = 1$, but replace $y(20) = -7$ by $u(0) = s$, where s must be found by the shooting method.

In total, our new system is

$$\begin{aligned} y' &= u \\ u' &= \frac{x+y}{25} \\ y(0) &= 1 \\ u(0) &= s \end{aligned}$$