Linnaeus University
Department of Mathematics
Sofia Eriksson

Written Exam on Numerical Methods, 2MA903, 1 hp (5 hp)
Tuesday 22th of March 2022, 14.00–19.00.

The solutions should be complete, correct, motivated, well structured and easy to follow.
Aids: Calculator (you may use a scientific calculator but *not* with internet connection).
Please begin each question on a new paper.
Preliminary grades: 15p-17p⇒E; 18p-20p⇒D; 21p-23p⇒C; 24p-26p⇒B; 27p-30p⇒A.

1. (a) Find the largest integer k for which $fl(35+2^{-k}) > fl(35)$ in double precision floating point representation.
(b) Find the roots of the quadratic equation $x^2+6x-7^{-14}=0$ with four significant digits accuracy (combining calculations by hand and evaluation on calculator). (5p)

Suggested solution:
(a) $35 = 32+2+1 = 2^5+2+2^0 = (100011)_2 = 1.00011\cdot 2^5$ we have

$$35+2^{-47} = 1.00011\cdot 2^5 + 0.\underbrace{000000\cdots0001}_{47\text{ bits}} = 1.00011\cdot 2^5 + 0.\underbrace{00000000000\cdots0001}_{52\text{ bits}}\cdot 2^5$$
$$= 1.\underbrace{00011000000\cdots0001}_{52\text{ bits}}\cdot 2^5$$

but

$$35+2^{-48} = 1.00011\cdot 2^5 + 0.\underbrace{000000\cdots0001}_{48\text{ bits}} = 1.00011\cdot 2^5 + 0.\underbrace{00000000000\cdots0000}_{52\text{ bits}}1\cdot 2^5$$
$$= 1.\underbrace{00011000000\cdots0000}_{52\text{ bits}}1\cdot 2^5$$

the last bit, bit number 53, is ignored (special case of rounding to nearest).
Thus $fl(35+2^{-47}) > fl(35)$ but $fl(35+2^{-48}) = fl(35)$
answer: $k=47$ is the largest integer for which $fl(35+2^{-k}) > fl(35)$.
(b)
The equation $x^2+6x-7^{-14}=0$ is solved by $x=-3\pm\sqrt{9+7^{-14}}$.
Executed (in double precision) this gives the roots
$-3+\sqrt{9+7^{-14}} \approx 2.455813330470846\cdot 10^{-13}$ and
$-3-\sqrt{9+7^{-14}} \approx -6.000000000000245$
Due to cancelling effects, the first of these roots is not very accurate. We rewrite the solution using its conjugate as
$x = \frac{(-3\pm\sqrt{9+7^{-14}})(-3\mp\sqrt{9+7^{-14}})}{-3\mp\sqrt{9+7^{-14}}} = \frac{-7^{-14}}{-3\mp\sqrt{9+7^{-14}}}$
which executed gives
$-7^{-14}/(-3-\sqrt{9+7^{-14}}) \approx 2.457401898265508\cdot 10^{-13}$
$-7^{-14}/(-3+\sqrt{9+7^{-14}}) \approx -6.003881160937727$ where the second is not very accurate.
answer: the roots are $2.457\cdot 10^{-13}$ and -6.000.

2. (a) The equation $x^3 - 6x^2 + 11x - 5 = 0$ has one real root, located in the interval $[0, 1]$. Do three iterations using the Bisection method. Report the answers and estimated errors in each iteration.

(b) How many iterations would we have to do using the Bisection method in order to guarantee four (4) correct decimals? (5p)

Suggested solution:

(a) We have $a_0 = 0$, $b_0 = 1$, $f(a_0) = -5 < 0$ and $f(b_0) = 1 > 0$. We have a change in sign so, yes there must be at least a root in the interval. Let $c_0 = (a_0 + b_0)/2 = 0.5$ be the initial approximation. Estimated error is $(b_0 - a_0)/2 = 0.5$.

Iteration 1: $f(c_0) = f(0.5) = -0.875 < 0$ $f(c)$ has the same sign as $f(a)$. Thus, let $a_1 = c_0 = 0.5$ be our new a and keep $b_1 = b_0$ as our b. Compute $c_1 = (a_1 + b_1)/2 = 0.75$. Estimated error is $(b_1 - a_1)/2 = 0.25$.

$f(c_1) = f(0.75) = 0.296875 > 0$. Since $f(c_1) > 0$, we replace b as $b_2 = c_1 = 0.75$ and keep $a_2 = a_1 = 0.5$. Compute $c_2 = (a_2 + b_2)/2 = 0.625$. Estimated error is $(b_2 - a_2)/2 = 0.125$.

$f(c_2) = -0.224609375 < 0$. We get $a_3 = c_2 = 0.625$ and keep $b_3 = b_2 = 0.75$. Our new approximation is $c_3 = 0.6875$, estimated error is $(b_3 - a_3) = (0.75 - 0.625)/2 = 0.125/2 = 0.0625$

alt: Tabular form:

iteration	a	b	$f(a_0)$	$f(b_0)$	c	$f(c)$	$\Delta x = (b-a)/2$
0	0	1	$-5 < 0$	$1 > 0$	0.5	$-0.875 < 0$	0.5
1	0.5	1	$-0.875 < 0$	$1 > 0$	0.75	$0.296\ldots > 0$	0.25
2	0.5	0.75	$-0.875 < 0$	$0.296\ldots > 0$	0.625	$-0.22\ldots < 0$	0.125
3	0.625	0.75	$-0.22\ldots < 0$	$0.296\ldots > 0$	0.6875		0.0625

(b) After one iteration the error is $0.25 = 2^{-2}$, after two iterations it is $0.125 = 2^{-3}$. The estimated error is $2^{-(i+1)}$ where i is the number of iterations. We have four correct decimals if the error is smaller than $0.5 \cdot 10^{-4}$, that is we seek an integer i such that

$2^{-(i+1)} < 0.5 \cdot 10^{-4}$

$2^{-i} < 10^{-4}$

$-i < \log_2(10^{-4}) \approx -13.287712379549449$

$i > 13.287712379549449$

answer: we need 14 iterations

3. (a) Interpolate the function $f(x) = \sin(x)$ at 4 equally spaced points on $[0, \pi/2]$.

   (b) Find an upper bound of the interpolation error at $x = \pi/4$. (5p)

   Suggested solution:
   See Sauer p.147-152

4. (a) Use the trapetzoidal method to calculate approximate values of the integral

$$I = \int_1^2 \ln(x^3)dx,$$

for 3 different step lengths: $h = 1, 0.5, 0.25$. (2p)

(b) Use Romberg's method on the approximate values of I obtained in a) to find an improved approximation of I. (3p)

Suggested solution:

(a)

$$R_{11} = T_{h=1} = \frac{h}{2}(\ln(1^3) + \ln(2^3)) = \frac{1}{2}(\ln(1) + \ln(8)) \approx 1.0397...$$

$$R_{21} = T_{h=0.5} = \frac{h}{2}(\ln(1^3) + 2\ln(1.5^3) + \ln(2^3)) = \frac{1}{4}(\ln(1) + 2\ln(1.5^3) + \ln(8)) \approx 1.1280...$$

$$R_{31} = T_{h=0.25} = \frac{1/4}{2}(\ln(1^3) + 2\ln(1.25^3) + 2\ln(1.5^3) + 2\ln(1.75^3) + \ln(2^3)) \approx 1.151098528$$

(b)

$$R_{22} = \frac{4R_{21} - R_{11}}{3} = \frac{4 \cdot 1.1280... - 1.0397...}{3} = 1.1575...$$

$$R_{32} = \frac{4R_{31} - R_{21}}{3} = \frac{4 \cdot 1.151098528 - 1.1280...}{3} = 1.1587...$$

$$R_{33} = \frac{16R_{32} - R_{22}}{15} = \frac{16 \cdot 1.1587... - 1.1575...}{15} = 1.1588...$$

answer: 1.1589

5. Let A be a 6×6 matrix with eigenvalues $\lambda_1 = -7$, $\lambda_2 = -6$, $\lambda_3 = -3$, $\lambda_4 = -2$, $\lambda_5 = 1$ and $\lambda_6 = 5$. Each eigenvalue λ_i is associated to a eigenvector $\mathbf{v}_i$, for $i = 1, 2, 3, 4, 5, 6$.

   To which eigenvector $\mathbf{v}_i$ (if any) does the algorithm converge to, when using

   (a) Power iteration,

   (b) Inverse Power Iteration,

   (c) Inverse Power Iteration with shift $s = 3$?

   Now let $\mathbf{v}$ be one of the eigenvectors of A such that $A\mathbf{v} = \lambda\mathbf{v}$, and assume that we have found an approximation $\widetilde{\mathbf{v}}$ of this eigenvector.

   (d) Derive the Rayleigh quotient, that is find the best approximation of λ in a least square sense. (5p)

   Suggested solution:

   (a) To the eigenvalue with biggest absolute value: Answer: -7

   (b) To the eigenvalue closest to s, that is, the method will converge to 1

   (c) The algorithm does not converge to any eigenvector because 3 is equally far from 1 and 5.

   (d) $A\tilde{v} = \lambda\tilde{v}$ can be viewed as an overdetermined equation system with one unknown; λ.

   This is easiest seen when written. $\tilde{v}\lambda = A\tilde{v}$. We form the normal equations by multiplying by the transpose of the coefficient matrix, i.e. by $\tilde{v}^T$. We get $\tilde{v}^T\tilde{v}\lambda = \tilde{v}^T A\tilde{v}$ and we solve for λ;

   $\lambda = \frac{\tilde{v}^T A \tilde{v}}{\tilde{v}^T \tilde{v}}$

Please turn, the questions continue on next page!

6. Let $y(x)$ be the solution of $y'(x) = t - ty$ for which $y(0) = 2$.

(a) Find an approximation of $y(2)$ using Euler backward with step length $h = 1$ and another approximation using $h = 0.5$. Answer using 4 correctly rounded decimals.

(b) Sketch the corresponding slope field for $y'(x) = t - ty$, for $x \in [0, 2]$. Include the two approximative solutions from (a) in your slope field picture.

(c) Using Richardson extrapolation, calculate an improved approximation of y(2) using the results obtained in (a). (5)

Suggested solution:
(a)

Euler backward: $y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}) = y_n + ht_{n+1}(1 - y_{n+1})$

Implicit method; solve for y_{n+1}: $y_{n+1} = \frac{y_n + ht_{n+1}}{1+ht_{n+1}}$

with $h = 1$ we have $t_0 = 0$, $t_1 = 1$ and $t_2 = 2$.

$y_0 = 2$

$y_1 = \frac{y_0+ht_1}{1+ht_1} = \frac{2+1}{1+1} = \frac{3}{2} = 1.5$

$y_2 = \frac{y_1+ht_2}{1+ht_2} = \frac{1.5+2}{1+2} = \frac{7}{6} = 1.1666\ldots$

such that $y(2) \approx 1.1667$.

with $h = 0.5$ we have $t_0 = 0$, $t_1 = 0.5$, $t_2 = 1$, $t_3 = 1.5$, $t_4 = 2$.

$y_0 = 2$

$y_1 = \frac{y_0+ht_1}{1+ht_1} = \frac{2+1/4}{1+1/4} = \frac{9}{5} = 1.8$

$y_2 = \frac{y_1+ht_2}{1+ht_2} = \frac{1.8+1/2}{1+1/2} = \frac{23}{15} = 1.5333\ldots$

$y_3 = \frac{y_2+ht_3}{1+ht_3} = \frac{1.5333...+3/4}{1+3/4} = \frac{137}{105} = 1.304761904761905\ldots$

$y_4 = \frac{y_3+ht_4}{1+ht_4} = \frac{1.30476......+1}{1+1} = \frac{121}{105} = 1.152380952380952\ldots$

such that $y(2) \approx 1.1524$.

(b)

(c) Euler backward is first order accurate. This means that $n = 1$ in the Richardson formula and we get the improved answer $\frac{2^1F(h/2)-F(h)}{2^1-1} = \frac{2\cdot1.152380952380952-1.1666...}{2-1} = 1.138095238095238\ldots$

answer: 1.1381