NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **MA2501 Numerical Methods**

Academic contact during examination: Markus Grasmair

Phone: 73 59 35 36

Examination date: 07th June 2014

Examination time (from–to): 09:00–13:00

Permitted examination support material:

- The textbook: Cheney & Kincaid, Numerical Mathematics and Computing, 6. or 7. edition.
- Rottmann, Mathematical formulae.
- Handouts on *Fixed point iterations* and *On the existence of a Cholesky factorization*.
- Approved basic calculator.

Other information:

- All answers should be justified and include enough details to make it clear which methods or results have been used.
- Some of the (sub-)problems will earn you more points than others — the total is 100 points.

Language: English

Number of pages: 10

Number pages enclosed: 0

Checked by:

Date Signature

Problem 1 Consider the linear system

$$\begin{pmatrix} 4 & 2 & 6 \\ 2 & 1 & 1 \\ -2 & 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \\ 5 \end{pmatrix}.$$

a) Compute the solution of this system using Gaussian elimination with scaled partial pivoting.
(10 points)

(The description of the method can be found in Cheney & Kincaid, pages 85ff; see in particular the 4×4 system[1] starting at the bottom of p. 87.) First, we compute for each line i its scale $\ell_i = \max_j |a_{ij}|$ (with a_{ij} being the ij-th entry of the matrix of the system). We obtain $\ell = (6, 2, 3)$. Next we compute for each line the relative size of its first entry, that is, the number $|a_{i1}|/\ell_i$. We obtain the scale vector $(2/3, 1, 2/3)$. Therefore we have to interchange the first and the second row and obtain the new system

$$\begin{pmatrix} 2 & 1 & 1 \\ 4 & 2 & 6 \\ -2 & 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 6 \\ 5 \end{pmatrix}.$$

We now subtract two times the first row from the second row and add the first row to the third row, and end up with

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & 0 & 4 \\ 0 & 4 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \\ 6 \end{pmatrix}.$$

Here it is obvious that we have to interchange the second and the third row. The resulting system is

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & 4 & 2 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 6 \\ 4 \end{pmatrix},$$

which is already in triangular form. From the last line of this system, we see that

$$z = 1.$$

Consequently,

$$y = \frac{1}{4}(6 - 2z) = 1.$$

Finally,

$$x = \frac{1}{2}(1 - y - z) = -\frac{1}{2}.$$

[1] Slightly exaggeratingly, this is called the "Sample 5×5 system."

Therefore

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 1 \end{pmatrix}.$$

b) Is it possible to solve this equation using Gaussian elimination without pivoting? Is it possible to apply Cholesky decomposition?
(5 points)

- It is not possible to solve this equation with Gaussian elimination without pivoting: If one tries to apply Gaussian elimination without any pivoting to the linear system, then one obtains after the first step the system

$$\begin{pmatrix} 4 & 2 & 6 \\ 0 & 0 & -2 \\ 0 & 4 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ -2 \\ 8 \end{pmatrix}.$$

  The next pivot element would be the second element in the second row, which, however, is zero. Therefore the algorithm breaks down at this point.
- Cholesky decomposition cannot be applied, because the matrix is not symmetric.

Problem 2 Consider the function

$$f(x) := 2x - \sin(x) + 2.$$

In order to solve the equation $f(x) = 0$, it is possible to apply a fixed point iteration of the form

$$x_{k+1} = x_k - \frac{1}{2} f(x_k).$$

a) Show that the equation $f(x) = 0$ has a unique solution $\hat{x}$, and that the iteration converges for every starting value $x_0 \in \mathbb{R}$ to $\hat{x}$.
(10 points)

We first note that $\hat{x}$ is a solution of the equation $f(x) = 0$, if and only if $\hat{x}$ is a fixed point of the mapping

$$x \mapsto \Phi(x) := x - \frac{1}{2} f(x) = \frac{1}{2}\sin(x) - 1.$$

Also, the iteration given in the problem is simply the fixed point iteration for the mapping Φ.

Now note that

$$\sup_{x\in\mathbb{R}}|\Phi'(x)| = \sup_{x\in\mathbb{R}}\Big|\frac{1}{2}\cos(x)\Big| = \frac{1}{2},$$

and therefore

$$|\Phi(x) - \Phi(y)| = \Big|\int_x^y \Phi'(z)\,dz\Big| \le \frac{1}{2}|x-y|$$

for all x, $y \in \mathbb{R}$. This shows that the mapping Φ is a contraction on $\mathbb{R}$ with contraction factor $\frac{1}{2}$. Therefore we infer using Banach's fixed point theorem (see the notes on fixed point iterations, Theorem 2) that Φ has a unique fixed point $\hat{x}$ and that the fixed point iteration $x_{k+1} = \Phi(x_k)$ converges for all starting values $x_0 \in \mathbb{R}$ to $\hat{x}$.

b) Compute one step of the fixed point iteration with a starting value $x_0 = 0$. Use your result to estimate, after how many steps we have $|x_k - \hat{x}| \le 2^{-20}$. *(5 points)*

After the first step of the fixed point iteration with $x_0 = 0$ we obtain

$$x_1 = x_0 - \frac{1}{2}(2x_0 - \sin(x_0) + 2) = \frac{1}{2}\sin(x_0) - 1 = \frac{1}{2}\sin(0) - 1 = -1$$

Since Φ is a contraction with contraction factor $1/2$, this implies that (see the notes on fixed point iterations, Theorem 2)

$$|x_k - \hat{x}| \le \frac{(1/2)^k}{1-(1/2)}|x_1 - x_0| = \frac{1}{2^{k-1}}.$$

For $k \ge 21$, the right hand side is smaller than or equal to 2^{-20}. Therefore, the required accuracy is reached after at most 21 steps.

Problem 3 Denote by f_n, $n \in \mathbb{N}$, the polynomial of degree n that interpolates the function $f(x) = e^x + e^{-x}$ in equidistant interpolation points in the interval $[0,1]$.

a) Show that $f_n(x) \to f(x)$ for every $x \in \mathbb{R}$. *(10 points)*

For every $x \in \mathbb{R}$ and $n \in \mathbb{N}$ there exists ξ (depending on both x and n) lying either in the interval $[0,1]$ or between x and the interval $[0,1]$ such that

$$f(x) - f_n(x) = \frac{1}{(n+1)!}f^{(n+1)}(\xi)\prod_{i=0}^{n}\Big(x - \frac{i}{n}\Big)$$

(see Cheney & Kincaid, p. 181, Theorem 1 [with the interpolation points $x_i = i/n$]). Moreover, we have

$$f^{(n+1)}(\xi) = e^{\xi} + (-1)^{n+1}e^{-\xi}.$$

Since ξ lies either in $[0,1]$ or between x and this interval, we have

$$e^{\xi} \le \max\{e^x, e^1\}, \qquad \text{and} \qquad e^{-\xi} \le \max\{e^{-x}, e^0\}.$$

Thus

$$|f^{(n+1)}(\xi)| \le \max\{e^x, e\} + \max\{e^{-x}, 1\} =: C$$

with C independent of ξ (and thus n). Moreover, for $0 \le x \le 1$ we have

$$\Big|\prod_{i=0}^{n}\Big(x - \frac{i}{n}\Big)\Big| \le 1,$$

for $x > 1$ we have

$$\Big|\prod_{i=0}^{n}\Big(x - \frac{i}{n}\Big)\Big| \le x^{n+1},$$

and for $x < 0$ we have

$$\Big|\prod_{i=0}^{n}\Big(x - \frac{i}{n}\Big)\Big| \le (-x+1)^{n+1}.$$

In total,

$$\Big|\prod_{i=0}^{n}\Big(x - \frac{i}{n}\Big)\Big| \le (|x|+1)^{n+1}.$$

Therefore

$$|f(x) - f_n(x)| \le \frac{C}{(n+1)!}(|x|+1)^{n+1}.$$

Since $s^{n+1}/(n+1)! \to 0$ for every $s \in \mathbb{R}$, this shows that $f_n(x) \to f(x)$ for every $x \in \mathbb{R}$.

b) Provide an estimate for

$$\sup_{0\le x\le 1} |f_5(x) - f(x)|.$$

(10 points)

For equidistant interpolation points on the interval $[0,1]$ we have the estimate

$$\sup_{0\le x\le 1} |f(x) - f_n(x)| \le \frac{h^{n+1}}{4(n+1)} \sup_{0\le x\le 1} |f^{(n+1)}(x)|$$

(see Cheney & Kincaid, p 183, Theorem 2). Here $h = 1/n$. With $n = 5$ we therefore have

$$\sup_{0\le x\le 1} |f(x) - f_5(x)| \le \frac{1}{5^6 \cdot 4 \cdot 6} \sup_{0\le x\le 1} |e^x + e^{-x}|.$$

Since the function $x \mapsto e^x + e^{-x}$ is convex, it attains its maximum on an interval on the interval's boundary. Thus

$$\sup_{0\le x\le 1} |e^x + e^{-x}| = \max\{e^0 + e^{-0}, e^1 + e^{-1}\} = e + e^{-1}.$$

Hence

$$\sup_{0\le x\le 1} |f(x) - f_5(x)| \le \frac{e + e^{-1}}{5^6 \cdot 4 \cdot 6} \approx 8.23 \cdot 10^{-6}.$$

Problem 4 We are given a function $f\colon \mathbb{R} \to \mathbb{R}$ at the following points:

x_i	-2	-1	$-\frac{1}{2}$	$\frac{1}{2}$	1	2
$f(x_i)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{2}$	2	4	16

Compute from these values the best possible approximation of $f'(0)$ using central finite differences and Richardson extrapolation.
(10 points)

(The method can be found in Cheney & Kincaid, p. 190ff. A rather efficient way of solving this problem is to simply "run" the *Derivative* Psudocode on page 193 with $x = 0$, $n = 2$, $h = 2$.) The central finite difference of f at 0 with step size $h > 0$ is given by

$$f_h'(0) := \frac{f(h) - f(-h)}{2h}$$

(in the notation of Cheney & Kincaid, this would be $\varphi(h)$). Using the points given in the table, we can compute central differences with step sizes 2, 1, and 1/2. We obtain

$$\begin{aligned} f_2'(0) &:= \frac{f(2) - f(-2)}{4} = \frac{16 - \frac{1}{16}}{4} = 3.984375,\\ f_1'(0) &:= \frac{f(1) - f(-1)}{2} = \frac{4 - \frac{1}{4}}{2} = 1.875,\\ f_{1/2}'(0) &:= \frac{f(1/2) - f(-1/2)}{1} = 2 - \frac{1}{2} = 1.5, \end{aligned}$$

We now start Richardson extrapolation and define

$$\begin{aligned} D_{00} &:= f_2'(0) = 3.984375, \\ D_{10} &:= f_1'(0) = 1.875, \\ D_{20} &:= f_{1/2}'(0) = 1.5. \end{aligned}$$

Next we compute the approximations D_{11} and D_{21} of $f'(0)$ using the formula

$$D_{k,1} = D_{k0} + \frac{1}{3}(D_{k0} - D_{k-1,0}).$$

We obtain

$$\begin{aligned} D_{11} &= D_{10} + \frac{1}{3}(D_{10} - D_{00}) = 1.171875, \\ D_{21} &= D_{20} + \frac{1}{3}(D_{20} - D_{10}) = 1.375. \end{aligned}$$

Finally, we compute

$$D_{22} = D_{21} + \frac{1}{15}(D_{21} - D_{11}) \approx 1.388542,$$

which is the best result we can get with this method from the given data.

Problem 5 Consider a quadrature rule of the form

$$Q(f, -1, 1) := 2\Big(c_0 f(-1) + c_1 f(-2/3) + c_2 f(0) + c_3 f(2/3) + c_4 f(1)\Big)$$

for the approximation of a definite integral $\int_{-1}^{1} f(x)\,dx$.

a) Find weights $c_0, \ldots, c_4 \in \mathbb{R}$ such that all polynomials of degree 4 are integrated exactly, that is,

$$Q(P, -1, 1) = \int_{-1}^{1} P(x)\,dx$$

whenever P is a polynomial of degree 4.
(15 points)

First one notes that the nodes are symmetric around 0 (the midpoint of the integration interval), and thus the weights have to be symmetric as well. That is, we have

$$c_0 = c_4 \qquad \text{and} \qquad c_1 = c_3.$$

Next we derive equations for the remaining weights by applying the quadrature formula to the monomials 1, x^2, and x^4 (because of the symmetry, all odd monomials are automatically integrated exactly).[2] For the monomial 1 we obtain

$$\int_{-1}^{1} 1\,dx = 2,$$
$$Q(1,-1,1) = 2(c_0 + c_1 + c_2 + c_1 + c_0).$$

For x^2 we have

$$\int_{-1}^{1} x^2\,dx = \frac{2}{3},$$
$$Q(x^2,-1,1) = 2\Big(c_0 + \frac{4}{9}c_1 + \frac{4}{9}c_1 + c_0\Big).$$

For x^4 we have

$$\int_{-1}^{1} x^4\,dx = \frac{2}{5},$$
$$Q(x^4,-1,1) = 2\Big(c_0 + \frac{16}{81}c_1 + \frac{16}{81}c_1 + c_0\Big).$$

We thus obtain the linear system

$$\begin{aligned} 2c_0 + 2c_1 + c_2 &= 1, \\ 2c_0 + \frac{8}{9}c_1 &= \frac{1}{3}, \\ 2c_0 + \frac{32}{81}c_1 &= \frac{1}{5}. \end{aligned}$$

Subtracting the third from the second equation, we obtain

$$\Big(\frac{8}{9} - \frac{32}{81}\Big)c_1 = \frac{1}{3} - \frac{1}{5},$$

[2] The problem is also solvable without using this shortcut. Comparing $Q(x^k,-1,1)$ with $\int_{-1}^{1} x^k\,dx$ for $k = 0,1,2,3,4$, leads (similar as in the proposed solution) to the linear system

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & -2/3 & 0 & 2/3 & 1 \\ 1 & 4/9 & 0 & 4/9 & 1 \\ -1 & -8/27 & 0 & 8/27 & 1 \\ 1 & 16/81 & 0 & 16/81 & 1 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1/3 \\ 0 \\ 1/5 \end{pmatrix}.$$

Now subtracting the fourth from the second line we obtain

$$-10/27c_1 + 10/27c_3 = 0 \qquad \text{or} \qquad c_1 = c_3.$$

Using this equation in the fourth line shows that

$$-c_0 - \frac{8}{27}c_1 + \frac{8}{27}c_1 + c_4 = 0 \qquad \text{and therefore} \qquad c_0 = c_4.$$

The rest of the equations are now the same as in the proposed solution.

which simplifies to

$$\frac{40}{81}c_1 = \frac{2}{15},$$

and therefore

$$c_1 = \frac{27}{100}.$$

Using the second equation in the linear system, we deduce that

$$c_0 = \frac{1}{2}\Big(\frac{1}{3} - \frac{8}{9}c_1\Big) = \frac{1}{2}\Big(\frac{1}{3} - \frac{6}{25}\Big) = \frac{7}{150}.$$

Finally,

$$c_2 = 1 - 2c_0 - 2c_1 = 1 - \frac{7}{75} - \frac{27}{50} = \frac{11}{30}.$$

To summarise, the quadrature formula reads

$$Q(f,-1,1) = 2\Big(\frac{7}{150}f(-1) + \frac{27}{100}f(-2/3) + \frac{11}{30}f(0) + \frac{27}{100}f(2/3) + \frac{7}{150}f(1)\Big).$$

b) Using the weights computed in the first part of the exercise, find the smallest integer $k \in \mathbb{N}$ for which $Q(x^k,-1,1) \neq \int_{-1}^{1} x^k\,dx$.
(5 points)

By construction we have $Q(x^k,-1,1) = \int_{-1}^{1} x^k\,dx$ for $k = 0,1,2,3,4$. Furthermore, the symmetry of the nodes and coefficients implies that

$$Q(x^5,-1,1) = 0 = \int_{-1}^{1} x^5\,dx.$$

Next we compute

$$\int_{-1}^{1} x^6\,dx = \frac{2}{7}$$

and

$$Q(x^6,-1,1) = 2\Big(\frac{7}{150} + \frac{27}{100}\frac{64}{729} + \frac{27}{100}\frac{64}{729} + \frac{7}{150}\Big) = 4\Big(\frac{7}{150} + \frac{32}{50\cdot 27}\Big) = \frac{38}{135}.$$

Since $\frac{2}{7} \neq \frac{38}{135}$, it follows that $k = 6$ is the smallest integer for which $Q(x^k,-1,1) \neq \int_{-1}^{1} x^k\,dx$.

Problem 6 Consider the initial value problem

$$\begin{aligned} y' &= \cos(y) - 2y \\ y(0) &= 0. \end{aligned}$$

a) Apply two steps of the explicit Euler method with a step size of $h = 1$ for the solution of this equation.
(5 points)

The explicit Euler method is given by the iteration $y_{k+1} = y_k + hf(y_k)$. In the given problem, $f(y) = \cos(y) - 2y$, and we use the step size $h = 1$. Thus we have $y_{k+1} = \cos(y_k) - y_k$. Moreover, we are given the initial value $y_0 = y(0) = 0$. We obtain therefore

$$\begin{aligned} y_1 &= \cos(0) - 0 = 1, \\ y_2 &= \cos(1) - 1 \approx -0.4596977. \end{aligned}$$

b) Apply two steps of the implicit Euler method with a step size of $h = 1$ for the solution of this equation. In each step, use two steps of Newton's method (with a resonable starting value of your choice) for the solution of the non-linear equation you have to solve.
(15 points)

The implicit Euler method is given by the iteration $y_{k+1} = y_k + hf(y_{k+1})$. With $h = 1$ we obtain in this case $y_{k+1} = y_k + \cos(y_{k+1}) - 2y_{k+1}$, or

$$3y_{k+1} - \cos(y_{k+1}) - y_k = 0.$$

Newton's method for the solution of this equation therefore reads as

$$y_{k+1,j+1} = y_{k+1,j} - \frac{3y_{k+1,j} - \cos(y_{k+1,j}) - y_k}{3 + \sin(y_{k+1,j})}$$

(note that the iteration index for Newton's method is j and k is fixed in this iteration). As starting value for this iteration, it makes sense to apply one step of the explicit Euler method,[3] that is,

$$y_{k+1,0} = y_k + \cos(y_k) - 2y_k = \cos(y_k) - y_k.$$

We thus obtain:

[3] Starting with the last iterate is also reasonable, that is, setting $y_{k+1,0} = y_k$. Starting all iterations with, for instance, $y_{k+1,0} = 0$ does not make much sense (although it works in this particular example).

- *Initialisation:*
$$y_0 = 0.$$

- *First step:*
$$y_{1,0} = \cos(0) - 0 = 1,$$
and then
$$y_{1,1} = 1 - \frac{3 - \cos(1)}{3 + \sin(1)} \approx 0.359699$$
and
$$y_1 := y_{1,2} \approx 0.359699 - \frac{3 \cdot 0.359699 - \cos(0.359699)}{3 + \sin(0.359699)} \approx 0.3170097.$$

- *Second step:*
$$y_{2,0} = \cos(0.3170097) - 0.3170097 \approx 0.6331621,$$
and then
$$y_{2,1} = 0.6331621 - \frac{3 \cdot 0.6331621 - \cos(0.6331621) - 0.3170097}{3 + \sin(0.6331621)} = 0.4170202$$
and
$$y_2 = y_{2,2} = 0.4170202 - \frac{3 \cdot 0.4170202 - \cos(0.4170202) - 0.3170097}{3 + \sin(0.4170202)} = 0.4112197.$$