NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **MA2501 Numerical Methods**

Academic contact during examination: Eirik Hoel Høiseth

Phone: 73 55 02 81

Examination date: 4. June 2015

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

Permitted examination support material: Support material code C

- Approved basic calculator.
- The textbook: Cheney & Kincaid, Numerical Mathematics and Computing, 6th or 7th edition, including the list of errata.
- Rottmann, Mathematical formulae.
- Handout: Fixed point iterations.

Other information:
All answers should be justified and include enough details to make it clear which methods and/or results have been used.

Some of the (sub-)problems are worth more points than others. The total value is 100 points

Language: English

Number of pages: 11

Number pages enclosed: 0

Checked by:

Date Signature

Problem 1 Approximate the value of the integral

$$\int_0^1 e^{-x}\,dx,$$

using the composite Simpson's rule with $n = 4$ subintervals. Determine an upper bound for the absolute error using the error term. Verify that the absolute error is within this bound.
`(10 points)`

Suggested solution:
Applying composite Simpson's rule with $n = 4$ subintervals, i.e. step-size $h = 1/4$, to this integral we get

$$\int_0^1 e^{-x}\,dx \approx \frac{1}{3\cdot 4}\left(e^{-0} + 4e^{-0.25} + 2e^{-0.5} + 4e^{-0.75} + e^{-1}\right) \approx 0.6321342.$$

We want to find an upper bound for the absolute error using the general error term for the composite Simpson's rule applied to $\int_a^b f(x)\,dx$

$$-\frac{1}{180}(b-a)h^4 f^{(4)}(\xi),$$

where $\xi \in (a, b)$. Taking absolute values, and inserting $a = 0$, $b = 1$ and $h = 1/4$, the absolute error E is

$$E = \frac{1}{180\cdot 4^4}\left|f^{(4)}(\xi)\right|.$$

Since $|f^{(4)}(x)| = e^{-x}$ is a decreasing positive function, we have the bound $|f^{(4)}(\xi)| \le e^{-0} = 1$. Insertion gives the absolute error bound

$$E \le \frac{1}{46080} \approx 2.2\times 10^{-5}.$$

To verify that the bound holds here, we easily compute the exact value of the integral

$$\int_0^1 e^{-x}\,dx = \left[-e^{-x}\right]_0^1 = -e^{-1} - (-e^{-0}) = 1 - e^{-1} \approx 0.6321206$$

Thus the actual absolute error is (correct to the digits used)

$$|0.6321342 - 0.6321206| \approx 1.4\times 10^{-5},$$

so the absolute error is within our bound. The bound quite closely bounds the actual absolute error in this case.

Problem 2

a) Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 9 & -3 & -3 \\ -3 & 10 & 1 \\ -3 & 1 & 5 \end{bmatrix}.$$

Show that $\mathbf{A}$ has a unique Cholesky factorization, without computing it.
`(5 points)`

Suggested solution:
$\mathbf{A}$ is real and symmetric by inspection. This also means the eigenvalues of $\mathbf{A}$ are real. Furthermore $\mathbf{A}$ is strictly diagonally dominant with positive diagonal elements. It follows from Gerschgorin's Theorem that all the eigenvalues of $\mathbf{A}$ are positive. This in turn implies that $\mathbf{A}$ is positive definite. $\mathbf{A}$ is thus symmetric positive definite (SPD) and consequently has a Cholesky factorization.

b) Compute the Cholesky factorization of $\mathbf{A}$, and use it to solve the linear system $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{b} = [-9, -1.5, 5]^T$.
`(12.5 points)`

Suggested solution:
We apply the algorithm for the Cholesky factorizations of $\mathbf{A} = [a_{ij}]$ as $\mathbf{A} = \mathbf{LL}^T$ where $\mathbf{L} = [l_{ij}]$, $i, j = 1, 2, 3$.

$$\begin{aligned}
l_{11} &= \sqrt{a_{11}} = 3, \\
l_{21} &= a_{12}/l_{11} = -1, \\
l_{31} &= a_{13}/l_{11} = -1, \\
l_{22} &= \sqrt{a_{22} - l_{21}^2} = \sqrt{10 - (-1)^2} = 3, \\
l_{32} &= \frac{a_{32} - l_{21}l_{31}}{l_{22}} = \frac{1 - (-1)(-1)}{3} = 0, \\
l_{33} &= \sqrt{a_{33} - l_{31}^2 - l_{32}^2} = \sqrt{5 - (-1)^2 - 0^2} = 2.
\end{aligned}$$

Thus

$$\mathbf{L} = \begin{bmatrix} 3 & 0 & 0 \\ -1 & 3 & 0 \\ -1 & 0 & 2 \end{bmatrix}.$$

$\mathbf{A} = \mathbf{LL}^T$ implies $\mathbf{L}(\mathbf{L}^T\mathbf{x}) = \mathbf{b}$ which means we can solve the linear system by doing two triangular solves. First we solve the lower triangular system

$$\mathbf{Ly} = \mathbf{b},$$

for $\mathbf{y} = [y_1, y_2, y_3]^T$ using forward substitution

$$\begin{aligned}
y_1 &= \frac{b_1}{l_{11}} = \frac{-9}{3} = -3,\\
y_2 &= \frac{b_2 - l_{21}y_1}{l_{22}} = \frac{-1.5 - (-1)(-3)}{3} = -1.5,\\
y_3 &= \frac{b_3 - l_{31}y_1 - l_{32}y_2}{l_{33}} = \frac{5 - (-1)(-3) - 0(-1.5)}{2} = 1.
\end{aligned}$$

Second we solve the upper triangular system

$$\mathbf{L}^T\mathbf{x} = \mathbf{y},$$

for $\mathbf{x} = [x_1, x_2, x_3]^T$ using back substitution.

$$\begin{aligned}
x_3 &= \frac{y_3}{l_{33}} = \frac{1}{2} = 0.5,\\
x_2 &= \frac{y_2 - l_{32}x_3}{l_{22}} = \frac{-1.5 - 0(0.5)}{3} = -0.5,\\
x_1 &= \frac{y_1 - l_{31}x_3 - l_{21}x_2}{l_{11}} = \frac{-3 - (-1)(0.5) - (-1)(-0.5)}{3} = -1.
\end{aligned}$$

The solution of the linear system is therefore $\mathbf{x} = [-1, -0.5, 0.5]^T$.

c) Perform 1 iteration of the SOR method with relaxation parameter $\omega = 1.1$ for the linear system $\mathbf{Ax} = \mathbf{b}$ from **b)**. Use the starting point $\mathbf{x}^{(0)} = [0, 0, 0]^T$.

Does it look like the iterations will converge towards the solution? Will the iterations converge for an arbitrary starting point?
`(10 points)`

Suggested solution:
Using the notation $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, x_3^{(i)}]^T$ and doing the iterations componentwise, we get

$$\begin{aligned}
x_1^{(1)} &= \omega \frac{b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)}}{a_{11}} + (1-\omega)x_1^{(0)} \\
&= 1.1 \frac{-9 - (-3)0 - (-3)0}{9} + (1-1.1)0 = -1.1, \\
x_2^{(1)} &= \omega \frac{b_2 - a_{21}x_1^{(1)} - a_{23}x_3^{(0)}}{a_{22}} + (1-\omega)x_2^{(0)} \\
&= 1.1 \frac{-1.5 - (-3)(-1.1) - 1 \cdot 0}{10} + (1-1.1)0 = -0.528, \\
x_3^{(1)} &= \omega \frac{b_3 - a_{31}x_1^{(1)} - a_{32}x_2^{(1)}}{a_{33}} + (1-\omega)x_3^{(0)} \\
&= 1.1 \frac{5 - (-3)(-1.1) - 1(-0.528)}{5} + (1-1.1)0 = 0.49016.
\end{aligned}$$

Though based on very limited results, it certainly looks like the iterations are converging towards the solution $\mathbf{x} = [-1, -0.5, 0.5]^T$ found in **b)**. The relative error in every component is at most 10% after just a single iteration, and the error decreases in the later components, where we use more updated values.

The iterations will in fact converge regardless of the starting point. We've determined in **a)** that $\mathbf{A}$ is SPD with positive diagonal elements, which implies, according to the Theorem on the convergence of the SOR method in Cheney & Kincaid, that the SOR method will converge for all $\omega \in (0, 2)$ and any starting point $\mathbf{x}^{(0)}$.

Problem 3 Let $f(x)$ be a polynomial of degree at most 3, i.e.

$$f(x) = Ax^3 + Bx^2 + Cx + D,$$

for some constants $A, B, C, D \in \mathbb{R}$. Suppose furthermore we are given the $n+1$ distinct knots $a = t_0 < t_1 < \ldots < t_n = b$ on the interval $[a, b]$.

Let $S(x)$ be the clamped cubic interpolating spline for the table of values

x	t_0	t_1	t_2	$\cdots$	t_n
y	$f(t_0)$	$f(t_1)$	$f(t_2)$	$\cdots$	$f(t_n)$

that satisfies $S'(a) = f'(a)$ and $S'(b) = f'(b)$. For what values of A, B, C, D, if any, will $S(x)$ equal $f(x)$ on $[a, b]$?

What if $S(x)$ is instead the natural cubic interpolating spline for this table of values?

Hint: $S(x)$ is uniquely defined in both cases.
(10 points)

Suggested solution:
It is clear that for any $A, B, C, D \in \mathbb{R}$, $f(x)$ will be an interpolating cubic spline here. In particular, it is a smooth function, which is a polynomial of degree at most 3 on any interval, and clearly interpolates itself at any set of points. What is left to check is the additional conditions that uniquely determines the spline.

For the clamped cubic spline, the criterion $S'(a) = f'(a)$ and $S'(b) = f'(b)$, is again obviously satisfied if $S(x) = f(x)$ on $[a, b]$. Thus the clamped cubic spline $S(x)$ will equal $f(x)$ on $[a, b]$ for any values of A, B, C, D.

For the natural cubic spline, the two additional criteria are $S''(a) = 0$ and $S''(b) = 0$. From

$$f''(x) = 6Ax + 2B,$$

we see that for f to satisfy these criteria, we must require $A = B = 0$, i.e. f must be a straight line. Thus the natural cubic spline $S(x)$ will equal $f(x)$ on $[a, b]$ for any values of C, D with $A = B = 0$.

Problem 4

a) The function $f(x) = \sin x$ has the unique zero $x^* = \pi$ on the interval $[3, 4]$. Perform 3 iterations of Newton's method to approximate this zero. Use $x_0 = 4$.

Note: For the sake of task **b)** do not round off the calculated values.
(7.5 points)

Suggested solution:
Since $f'(x) = \cos x$ the Newton iteration becomes

$$x_{n+1} = x_n - \frac{\sin x_n}{\cos x_n} = x_n - \tan x_n,$$

for $n = 0, 1, 2, \cdots$. Thus

$$\begin{aligned}
x_1 &= x_0 - \tan x_0 = 4 - \tan 4 = 2.842178718,\\
x_2 &= x_1 - \tan x_1 = 2.842178718 - \tan 2.842178718 = 3.150872940,\\
x_3 &= x_2 - \tan x_2 = 3.150872940 - \tan 3.150872940 = 3.141592387.
\end{aligned}$$

b) Determine the absolute error for x_i, $i = 0, 1, 2, 3$, from **a)**. Estimate the order of convergence α. Explain this behaviour.

Hint: You can assume $\alpha \in \mathbb{N}$. A method with order of convergence $\alpha \in \mathbb{N}$ will behave like

$$|x_{n+1} - x^*| \approx M|x_n - x^*|^\alpha,$$

for some positive constant M when $|x_n - x^*|$ becomes sufficiently small. `(12.5 points)`

Suggested solution:
Using the notation, $e_n \equiv |x_n - x^*|$ we compute the absolute errors

$$\begin{aligned}
e_0 &\equiv |x_0 - x^*| = |4 - \pi| \approx 8.584 \times 10^{-1},\\
e_1 &\equiv |x_1 - x^*| = |2.842178718 - \pi| \approx 2.994 \times 10^{-1},\\
e_2 &\equiv |x_2 - x^*| = |3.150872940 - \pi| \approx 9.280 \times 10^{-3},\\
e_3 &\equiv |x_3 - x^*| = |3.141592387 - \pi| \approx 2.666 \times 10^{-7}.
\end{aligned}$$

Since the number of correct digits in the approximation appears to roughly triple for each iteration, we suspect that the order of convergence is cubic, i.e $\alpha = 3$. To test this we follow the hint and compute the ratios

$$\frac{e_{n+1}}{e_n^3}.$$

This is done below

$$\begin{aligned}
\frac{e_1}{e_0^3} &\approx \frac{2.994 \times 10^{-1}}{(8.584 \times 10^{-1})^3} \approx 0.473,\\
\frac{e_2}{e_1^3} &\approx \frac{9.280 \times 10^{-3}}{(2.994 \times 10^{-1})^3} \approx 0.346,\\
\frac{e_3}{e_2^3} &\approx \frac{2.666 \times 10^{-7}}{(9.280 \times 10^{-3})^3} \approx 0.334.
\end{aligned}$$

These ratios do indeed appear to tend towards a constant, and we estimate the order of convergence to be cubic. This behaviour is immediately a bit surprising, since Newton's method normally gives quadratic convergence for sufficiently good initial guesses. To explain this we look at this iteration as a fixed point iteration

$$x_{n+1} = g(x_n)$$

with

$$g(x) = x - \tan x.$$

From Theorem 10 in the note on fixed point iterations, we expect precisely cubic order of convergence provided $g'(x^*) = g''(x^*) = 0$, but $g'''(x^*) \neq 0$. Computing derivatives of g is straightforward

$$\begin{aligned} g'(x) &= 1 - \frac{1}{\cos^2 x}, \\ g''(x) &= -2\frac{\sin x}{\cos^3 x}, \\ g'''(x) &= -2\frac{\cos^4 x + 3\sin^2 x \cos^2 x}{\cos^6 x} = -2\frac{\cos^2 x + 3\sin^2 x}{\cos^4 x} = -2\frac{1 + 2\sin^2 x}{\cos^4 x}. \end{aligned}$$

Evaluation at $x = x^* = \pi$, gives $g'(x^*) = g''(x^*) = 0$ and $g'''(x^*) = -2 \neq 0$, so the observed cubic convergence is indeed as expected from the theory. In fact the proof of the aforementioned theorem indicates that we should expect

$$M = \frac{|g'''(x^*)|}{3!} = \frac{1}{3},$$

which is consistent with our earlier ratio computations.

Problem 5

a) We want to approximate $f''(x)$ with a computer, by using the forward difference rule

$$f''(x) \approx \frac{f(x) - 2f(x+h) + f(x+2h)}{h^2},$$

with $h > 0$. Assume that, in the above formula, all function values are subject to round off error, so that the actual values used by the computer are

$$\begin{aligned} \tilde{f}(x) &= f(x)(1+\delta_1), \\ \tilde{f}(x+h) &= f(x+h)(1+\delta_2), \\ \tilde{f}(x+2h) &= f(x+2h)(1+\delta_3), \end{aligned}$$

where $|\delta_i| \leq \epsilon$, $i = 1, 2, 3$. Show the following upper bound for the total approximation error

$$\left| f''(x) - \frac{\tilde{f}(x) - 2\tilde{f}(x+h) + \tilde{f}(x+2h)}{h^2} \right| \leq \frac{5h}{3} M_1 + \frac{4\epsilon}{h^2} M_2,$$

where $M_1 = \max_{y \in [x, x+2h]} |f'''(y)|$ and $M_2 = \max_{y \in [x, x+2h]} |f(y)|$.
`(12.5 points)`

Suggested solution:
We first compute an expression for the truncation error. Taylor expanding $f(x+h)$ and $f(x+2h)$ gives

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2!} f''(x) + \frac{h^3}{3!} f'''(\xi_1), \\ f(x+2h) &= f(x) + 2hf'(x) + \frac{2^2 h^2}{2!} f''(x) + \frac{2^3 h^3}{3!} f'''(\xi_2). \end{aligned}$$

with $\xi_1 \in (x, x+h)$ and $\xi_2 \in (x, x+2h)$. Inserting this in the rule and grouping in powers of h

$$\begin{aligned} &\frac{f(x) - 2f(x+h) + f(x+2h)}{h^2} = \frac{[f(x) - 2f(x) + f(x)] + h\,[-2f'(x) + 2f'(x)]}{h^2} \\ &+ \frac{h^2\,[-f''(x) + 2f''(x)] + h^3\,[-2f''(\xi_1) + 8f''(\xi_2)]\,/6}{h^2} \\ &= f''(x) + h\frac{-2f''(\xi_1) + 8f''(\xi_2)}{6}. \end{aligned}$$

Now we include the rounding error

$$\begin{aligned} &\frac{\tilde{f}(x) - 2\tilde{f}(x+h) + \tilde{f}(x+2h)}{h^2} = \\ &= \frac{f(x) - 2f(x+h) + f(x+2h)}{h^2} + \frac{f(x)\delta_1 - 2f(x+h)\delta_2 + f(x+2h)\delta_3}{h^2} \\ &= f''(x) + h\frac{-2f''(\xi_1) + 8f''(\xi_2)}{6} + \frac{f(x)\delta_1 - 2f(x+h)\delta_2 + f(x+2h)\delta_3}{h^2}. \end{aligned}$$

From this we determine the bound

$$\begin{aligned}
&\left| f''(x) - \frac{\tilde{f}(x) - 2\tilde{f}(x+h) + \tilde{f}(x+2h)}{h^2} \right| = \\
&= \left| h\frac{-2f''(\xi_1) + 8f''(\xi_2)}{6} + \frac{f(x)\delta_1 - 2f(x+h)\delta_2 + f(x+2h)\delta_3}{h^2} \right| \\
&\leq \left| h\frac{-2f''(\xi_1) + 8f''(\xi_2)}{6} \right| + \left| \frac{f(x)\delta_1 - 2f(x+h)\delta_2 + f(x+2h)\delta_3}{h^2} \right| \\
&\leq h\frac{2\left|f''(\xi_1)\right| + 8\left|f''(\xi_2)\right|}{6} + \frac{|f(x)|\,|\delta_1| + 2\,|f(x+h)|\,|\delta_2| + |f(x+2h)|\,|\delta_3|}{h^2} \\
&\leq h\frac{5}{3}M_1 + \frac{4\epsilon}{h^2}M_2.
\end{aligned}$$

b) Find the positive value of h, h_{min}, that minimizes the upper bound from **a)**, taking M_1, M_2 and ϵ to be known and positive constants.
`(5 points)`

Suggested solution:
It is obvious that the bound tends to infinity when $h \to 0$ or $h \to \infty$, so there must be some global minimum. Since the function is differentiable for $h > 0$ this minimum happens when the derivative with respect to h is 0. Thus

$$\frac{5}{3}M_1 - \frac{8\epsilon}{h_{min}^3}M_2 = 0.$$

Isolating h_{min} gives

$$h_{min} = \sqrt[3]{\frac{24\epsilon M_2}{5M_1}}.$$

c) Compute an approximation of h_{min} for $f(x) = \ln x$ and $x = 2$, using $\epsilon = 1.1 \times 10^{-16}$ (double precision), $M_1 \approx |f'''(x)|$ and $M_2 \approx |f(x)|$.

Estimates of $f''(x)$ in this case, using the forward difference rule, were calculated on a computer with double precision. The corresponding absolute errors are given in the table below. Do these results agree with our analysis?

h	Absolute error
10^{-1}	0.022985146546089
10^{-2}	0.002478310901527
10^{-3}	0.000249781382827
10^{-4}	0.000024981537422
10^{-5}	0.000001089537932
10^{-6}	0.000133247448048
10^{-7}	0.016853164828717
10^{-8}	0.250000000000000

(5 points)

Suggested solution:
For the specific case we have $M_1 \approx 2/2^3 = 1/4$ and $M_2 \approx \ln 2$. Insertion of this and the value for ϵ gives

$$h_{min} \approx \sqrt[3]{\frac{4 \cdot 24 \cdot 1.1 \times 10^{-16} \cdot \ln 2}{5}} \approx 1.1 \times 10^{-5}$$

Regarding the table, we see that the approximation error initally decreases as $\mathcal{O}(h)$ (truncation error dominating), reaches a minimum around h_{min}, before roughly increasing as $\mathcal{O}(1/h^2)$ (round off error dominating). Thus the results mesh well with what is expected from our analysis. Note that in the last line h has become so small that the numerator in the estimate for $f''(x)$, and consequently the estimate itself, evaluated to 0.

Problem 6 Consider the following initial value problem for $x(t) : \mathbb{R} \to \mathbb{R}$ and $y(t) : \mathbb{R} \to \mathbb{R}$

$$\begin{aligned} x'' &= e^{-x'} + x - \cos t, \\ y' &= \sqrt[3]{y} - tx', \\ x(0) &= -2, \quad x'(0) = 0, \quad y(0) = 8. \end{aligned}$$

Convert this problem to an equivalent system of first-order differential equations in autonomous vector form, with initial values. Take 2 steps with Euler's method with step-size $h = 0.5$ for this system.
(10 points)

Suggested solution:
By introducing new variables

$$\mathbf{X} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} t \\ x \\ x' \\ y \end{bmatrix},$$

we arrive at the system

$$\mathbf{X}' = \begin{bmatrix} x_0' \\ x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} 1 \\ x_2 \\ e^{-x_2} + x_1 - \cos(x_0) \\ \sqrt[3]{x_3} - x_0 x_2 \end{bmatrix} = \mathbf{F}(\mathbf{X}),$$

with initial conditions

$$\mathbf{X}(0) = \begin{bmatrix} x_0(0) \\ x_1(0) \\ x_2(0) \\ x_3(0) \end{bmatrix} = \begin{bmatrix} 0 \\ -2 \\ 0 \\ 8 \end{bmatrix} = \mathbf{X}_0.$$

We now take two steps with Euler's method, denoting by $\mathbf{X}_n$ the result after n steps.

$$\begin{aligned} \mathbf{X}_1 = \mathbf{X}_0 + h\mathbf{F}(\mathbf{X}_0) &= \begin{bmatrix} 0 \\ -2 \\ 0 \\ 8 \end{bmatrix} + 0.5 \begin{bmatrix} 1 \\ 0 \\ -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ -2 \\ -1 \\ 9 \end{bmatrix}, \\ \mathbf{X}_2 = \mathbf{X}_1 + h\mathbf{F}(\mathbf{X}_1) &= \begin{bmatrix} 0.5 \\ -2 \\ -1 \\ 9 \end{bmatrix} + 0.5 \begin{bmatrix} 1 \\ -1 \\ e^1 - 2 - \cos 0.5 \\ \sqrt[3]{9} - 0.5(-1) \end{bmatrix} \approx \begin{bmatrix} 1.00000 \\ -2.50000 \\ -1.07965 \\ 10.2900 \end{bmatrix}. \end{aligned}$$

We remark that the exact solution is $\mathbf{X}(1) = [1.00000, -2.66886, -1.05234, 10.5098]^T$, accurate to the digits given. Therefore the approximation is not very accurate. This should be expected, since we are using a first order method with a fairly large step-size.