NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **MA2501 Numerical Methods**

Academic contact during examination: Elena Celledoni
Phone: 48238584

Examination date: May 13, 2019
Examination time (from–to): 15:00-19:00
Permitted examination support material: C: Approved calculator.

Language: English
Number of pages: 9
Number of pages enclosed: 2

Checked by:

Informasjon om trykking av eksamensoppgave
Originalen er:
1-sidig ☒ **2-sidig** ☐
sort/hvit ☒ **farger** ☐
skal ha flervalgskjema ☐

Date Signature

This exam set includes an appendix with formulae and theorems useful for the solution of the exam questions.

Problem 1 Interpolate the function $\log x$ with a quadratic polynomial in the interpolation nodes $x_0 = 10$, $x_1 = 11$, $x_2 = 12$.

a) Write down the interpolation polynomial.

b) Find an error bound for error committed in $x = 11.1$ using the error formula of the interpolation polynomial.

Solution

a) Using Newton divided differences and the Newton form of the interpolation polynomial we obtain:

$$p_2(x) = \log(10)+(\log(11)-\log(10))(x-10)+\frac{\log(12)-2\log(11)+\log(10)}{2}(x-10)(x-11).$$

b) Using the formula for the interpolation error

$$p_2(x) - \log(x) = \frac{1}{6}\,\frac{d^3}{dx^3}\log(x)\bigg|_{x=\xi}(x-10)(x-11)(x-12), \qquad \xi \in (10,12)$$

we obtain

$$|p_2(11.1) - \log(11.1)| \leq \frac{1}{6}\left(\max_{\xi\in(10,12)}\frac{2}{\xi^3}\right)\,(1.1\cdot 0.1\cdot 0.9) = 3.3\cdot 10^{-5}$$

while the true error is $p_2(11.1) - \log(11.1) = -2.474830\cdot 10^{-5}$. And we see that the error formula gives a rather accurate bound of the error in this case.

Problem 2 The Newton-Cotes formula with $n = 3$ on the interval $[-1, 1]$ is

$$\int_{-1}^{1} f(x)\,dx \approx w_0 f(-1) + w_1 f(-1/3) + w_2 f(1/3) + w_3 f(1).$$

Show that

$$\begin{aligned} 2w_0 + 2w_1 &= 2 \\ 2w_0 + \frac{2}{9}w_2 &= \frac{2}{3}, \end{aligned}$$

and find the values of the weights w_0, w_1, w_2 and w_3. You can use the fact that this formula is to be exact for all polynomials of degree 3.

Solution

Imposing the quadrature formula being exact for the basis of monomials of degree less than or equal to three, that is 1, x, x^2 and x^3, we obtain the following four equations:

$$\begin{aligned} w_0 + w_1 + w_2 + w_3 &= \int_{-1}^{1} 1\,dx = 2 \\ w_0(-1) + w_1(-\frac{1}{3}) + w_2(\frac{1}{3}) + w_3 &= \int_{-1}^{1} x\,dx = 0 \\ w_0(-1)^2 + w_1(-\frac{1}{3})^2 + w_2(\frac{1}{3})^2 + w_3 &= \int_{-1}^{1} x^2\,dx = \frac{2}{3} \\ w_0(-1)^3 + w_1(-\frac{1}{3})^3 + w_2(\frac{1}{3})^3 + w_3 &= \int_{-1}^{1} x^3\,dx = 0 \end{aligned}$$

the second and fourth equation lead to $w_3 - w_0 = \frac{1}{3}(w_1 - w_2)$ and $w_3 - w_0 = \frac{1}{27}(w_1 - w_2)$, and so $w_3 = w_0$ and $w_1 = w_2$. The system is therefore simplified into

$$\begin{aligned} 2w_0 + 2w_1 &= 2 \\ 2w_0 + \frac{2}{9}w_1 &= \frac{2}{3} \end{aligned}$$

where $w_1 = w_2$, and we see that this linear system coincides with the one of the text of the exercise. Solving this system we obtain $w_1 = w_2 = \frac{3}{4}$ and $w_3 = w_0 = \frac{1}{4}$.

Problem 3

a) Consider the approximation of the derivatives of a differentiable function using central differences, i.e.

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}. \tag{1}$$

Assuming f is three times differentiable with continuous third derivative on $(x-h, x+h)$ show that

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + E(f,x), \qquad E(f,x) = -\frac{h^2}{6}f'''(\eta), \qquad \eta \in (x-h, x+h).$$

Solution

Using Taylor thorem around x we get

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{3!}f'''(\xi), \qquad \xi \in (x, x+h),$$

and similarly

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{3!}f'''(\gamma), \qquad \gamma(x-h, x).$$

Subtracting the second from the first and dividing by $2h$ we get

$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{h^2}{3!}\frac{f'''(\xi) + f'''(\gamma)}{2}$$

since f''' is continuous on $(x-h, x+h)$ by the intermediate value theorem there exist $\eta \in (\xi, \gamma) \subset (x-h, x+h)$ such that $\frac{f'''(\xi)+f'''(\gamma)}{2} = f'''(\eta)$. And so

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6}f'''(\eta), \qquad \eta \in (x-h, x+h),$$

as we wanted to show.

b) We want to approximate the derivative of $\sinh(x)$ in $x = 0.400$. We are given the following values of the function with a precision of 8 digits (error 10^{-9})

x	$f(x)$
0.398	0.40859100
0.399	0.40967146
0.400	0.41075233
0.401	0.41183360
0.402	0.41291529

and we know that the exact value of the derivative in $x = 0.4$ is 1.0810724 (also with a precision of 8 digits).

Using central differences (1) we obtain the following results

h	approximation	error
0.001	1.0810700	$2.4 \cdot 10^{-6}$
0.002	1.0810725	-10^{-7}

Can you explain why contrary to the result in **a**) the most accurate value is the one obtained with the largest value of h?

Hint: Make an analysis of the propagation of rounding errors. Start from

$$\frac{(f(x+h) + \varepsilon_+) - (f(x-h) + \varepsilon_-)}{2h},$$

where ε_+ and ε_- are the rounding errors committed in the evaluation of sinh (here $|\varepsilon_+| < 10^{-9}$ and $|\varepsilon_-| < 10^{-9}$), $x = 0.4$, and $h = 0.001$ or $h = 0.002$.

Solution

We have

$$\frac{(f(x+h)+\varepsilon_+)-(f(x-h)+\varepsilon_-)}{2h} = \frac{f(x+h)-f(x-h)}{2h} + \frac{\varepsilon_+-\varepsilon_-}{2h},$$

and we see that while

$$\lim_{h\to 0}\frac{f(x+h)-f(x-h)}{2h} = f'(x),$$

we have

$$\lim_{h\to 0}\frac{|\varepsilon_+-\varepsilon_-|}{2h} = +\infty.$$

because $|\varepsilon_+ - \varepsilon_-|$ however small does not tend to zero as h tends to zero.

In our particular example we have $E(x,h) = -\frac{h^2}{6}\cosh(x)$ and using the approximation $\frac{|\varepsilon_+-\varepsilon_-|}{2h} \approx \frac{10^{-9}}{h}$ we obtain a table consistent with the one given in the text of the exercise

h	$E(x,h)$	$\frac{10^{-9}}{h}$
0.001	$1.6\cdot 1.08\cdot 10^{-7}$	10^{-6}
0.002	$6.6\cdot 1.08\cdot 10^{-7}$	$5\cdot 10^{-7}$

showing that for $h = 0.001$ (or smaller) the error is dominated by the roundoff error and not by the truncation error.

Problem 4

a) Find the QR factorisation of the matrix

$$A = \begin{pmatrix} 9 & -6 \\ 12 & -8 \\ 0 & 20 \end{pmatrix}.$$

Solution

We need two Householder transformations, P_1 and P_2 for obtaining the QR factorization. Using the results in the Appendix we obtain

$$P_1 = I - 2ww^T = \begin{bmatrix} 0.6 & 0.8 & 0 \\ 0.8 & -0.6 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$P_1 A = \begin{bmatrix} 15 & -10 \\ 0 & 0 \\ 0 & 20 \end{bmatrix}.$$

It is clear that to obtain the upper triangular matrix R we need only to permute the last and second last row of $P_1 A$, this is achieved by multiplying by

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

One obtains the same P_2 considering the Hauseholder transformation obtained by applying the procedure from the Appendix to the vector $[0,\, 20]^T$ in the second column of $P_1 A$. Then we get

$$Q = P_1 P_2 = \begin{bmatrix} 0.6 & 0 & 0.8 \\ 0.8 & 0 & -0.6 \\ 0 & 1 & 0 \end{bmatrix}, \qquad R = \begin{bmatrix} 15 & -10 \\ 0 & 20 \\ 0 & 0 \end{bmatrix}.$$

b) Find the least squares solution of the system of linear equations

$$\begin{aligned} 9x - 6y &= 300, \\ 12x - 8y &= 600, \\ 20y &= 900. \end{aligned}$$

Solution

The solution can be found by using the obtained QR factorisation or by solving the normal equations. Applying $Q^T = P_2 P_1$ to both sides of the system we obtain

$$\begin{aligned} 15x - 10y &= 660, \\ 20y &= 900. \end{aligned}$$

with solution $x = 74$, $y = 45$. This is also the least squares solution of the system (see Theorem 2.13 in Süli and Mayers) and the solution of the normal equations $A^T A\mathbf{x} = A^T\mathbf{b}$ with $\mathbf{x} := [x\,, y]^T$ and $\mathbf{b} := [300,\, 600,\, 900]^T$.

Problem 5 We consider the pendulum equations

$$\theta''(t) + \sin(\theta(t)) = 0, \quad 0 < t < 1, \quad \theta(0) = \alpha,\ \theta(1) = \beta, \tag{2}$$

with numerical discretization on the grid $t_m = m\,h$, $m = 0, \ldots, M+1$ and $h = \frac{1}{M+1}$ leading to the discretized equations

$$\frac{1}{h^2}(\Theta_{m-1} - 2\Theta_m + \Theta_{m+1}) + \sin(\Theta_m) = 0, \quad m = 1, 2, \ldots, M+1.$$

The system of equations is nonlinear and we write it in a vector form as

$$G_h(\Theta) = \mathbf{0}, \quad \Theta = \begin{bmatrix} \Theta_1 \\ \vdots \\ \Theta_M \end{bmatrix},$$

where $G_h(\Theta) = A_h\Theta + \sin(\Theta)$, $\sin(\Theta) := [\sin(\Theta_1), \ldots, \sin(\Theta_M)]^T$ and A_h is the $M \times M$ matrix

$$A_h := \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}.$$

A numerical solution of this system can be obtained iteratively, for example using a Newton method:

$$\Theta^{[k+1]} = \Theta^{[k]} - J_h(\Theta^{[k]})^{-1} G_h(\Theta^{[k]}),$$

where $J_h(\Theta)$ is the Jacobian of $G_h(\Theta)$.

a) Find the Jacobian matrix $J_h(\Theta)$, i.e. the matrix with entries

$$(J_h)_{i,j}(\Theta) = \frac{\partial G_{h,i}(\Theta)}{\partial \Theta_j}, \qquad i, j = 1, \ldots, M$$

where we have denoted with $G_{h,i}(\Theta)$ the i-th component of G_h.

Solution

$J_h(\Theta)$ is a tridiagonal symmetric matrix such that

$$(J_h)_{i,j}(\Theta) = \begin{cases} \frac{1}{h^2} & \text{for} \quad j = i-1 \text{ or } j = i+1, \\ \frac{-2}{h^2} + \cos(\Theta_i) & \text{for} \quad j = i, \\ 0 & \text{otherwise}, \end{cases}$$

so

$$J_h(\Theta) = A_h + C(\Theta), \qquad C(\Theta) := \operatorname{diag}(\cos(\Theta_1), \ldots, \cos(\Theta_M)).$$

b) The truncation error is

$$\overrightarrow{\tau}_h := G_h(\overrightarrow{\theta}), \qquad \overrightarrow{\theta} := \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}, \quad \theta_j := \theta(t_j).$$

Show using Taylor theorem that the components of $\overrightarrow{\tau}_h$ satisfy

$$\tau_m = \frac{1}{12}h^2\theta^{(4)}(t_m) + \mathcal{O}(h^4), \quad m = 1, \ldots, M.$$

This ensures that the method is consistent of order 2. What does this mean for the 2-norm of the vector $\overrightarrow{\tau}_h$? Interpret now the vector $\overrightarrow{\tau}_h$ as a piecewise constant function τ_h defined by

$$\tau_h(x) := \tau_m, \qquad x \in [t_m, t_{m+1}), \qquad m = 1, \ldots, M.$$

What can you conclude about the 2-norm of this piecewise constant function?

Solution

From the equation we can see that the solution of the problem is smooth and has bounded derivatives. We write out $\tau_m = G_{h,m}(\overrightarrow{\theta})$, this is

$$\tau_m = \frac{1}{h^2}(\theta_{m-1} - 2\theta_m + \theta_{m+1}) + \sin(\theta_m).$$

Since $\frac{1}{h^2}(\theta_{m-1} - 2\theta_m + \theta_{m+1}) = \theta''(t_m) + \frac{h^2}{12}\theta^{(4)}(\hat{t}_m)$ with $\hat{t}_m \in (t_m, t_m + h)$ then

$$\tau_m = \theta''(t_m) + \frac{h^2}{12}\theta^{(4)}(\hat{t}_m) + \sin(\theta_m) = \frac{h^2}{12}\theta^{(4)}(\hat{t}_m).$$

We will use the notation $\overrightarrow{\tau}_h$ to mean the vector with components τ_m and τ_h to mean the corresponding piecewise constant function on $[0, 1]$. We have proved in the course that there is a scaling factor $h^{\frac{1}{2}}$ between the 2-norm of τ_h interpreted as a piecewise constant function and as a vector:

$$\|\tau_h\|_2 = h^{\frac{1}{2}} \|\overrightarrow{\tau}_h\|_2.$$

For the norm τ_h as a vector we have

$$\begin{aligned} \|\overrightarrow{\tau}_h\|_2 &= \sqrt{\sum_{m=1}^{M} \tau_m^2} = \sqrt{\sum_{m=1}^{M} \left(\frac{h^2}{12}\right)^2 (\theta^{(4)}(\hat{t}_m))^2} = \sqrt{\left(\frac{h^2}{12}\right)^2 M (\max_{t\in[0,1]} |\theta^{(4)}(t)|)^2} \\ &\leq \frac{h^2}{12}\sqrt{M} \max_{t\in[0,1]} |\theta^{(4)}(t)| \end{aligned}$$

so, since $\sqrt{M+1} = \left(\frac{1}{h}\right)^{\frac{1}{2}}$, we have

$$\|\overrightarrow{\tau}_h\|_2 = \mathcal{O}(h^{\frac{3}{2}}).$$

On the other hand for the 2-norm of the piecewise constant function τ_h we have

$$\|\tau_h\|_2 \leq h^{\frac{1}{2}} \frac{h^2}{12} \sqrt{M} \max_{t\in[0,T]} |\theta^{(4)}(t)| = \mathcal{O}(h^2).$$

In both cases $\lim_{h\to 0} \|\tau_h\|_2 = 0$, $\lim_{h\to 0} \|\overrightarrow{\tau}_h\|_2 = 0$.

c) Finally we want to prove convergence. To this aim, you first need to find the equation relating the error $\overrightarrow{E}_h := \Theta - \overrightarrow{\theta}$ to the truncation error $\overrightarrow{\tau}_h$. Then derive an appropriate bound of the 2-norm of the error by means of the 2-norm of the truncation error and prove convergence.

Hint

- Use Taylor theorem to express $\sin(\Theta)$ by means of $\sin(\overrightarrow{\theta})$ and the error E_h.
- You may use a result that we have proven in the lectures that A_h is invertible for h sufficiently small and $\|A_h^{-1}\|_2 = \frac{1}{\pi^2} + K\,h^2$, with K a constant independent on h.

Solution

Using that $G_h(\Theta) = \mathbf{0}$ and $G_h(\overrightarrow{\theta}) = \tau_h$ we obtain the equation

$$A_h\,\overrightarrow{E}_h + \sin(\Theta) - \sin(\overrightarrow{\theta}) = -\overrightarrow{\tau}_h,$$

since $\sin(\Theta) = \sin(\overrightarrow{\theta} + \overrightarrow{E}_h)$ and for each component $\sin(\Theta_m) = \sin(\theta_m + E_{h,m}) = \sin(\theta_m) + E_{h,m}\cos(\hat{\theta}_m)$ with $\hat{\theta}_m \in (\theta_m, \theta_m + E_{h,m})$, we can rewrite the equation in the form

$$A_h\,\overrightarrow{E}_h + C(\hat{\theta})\overrightarrow{E}_h = -\overrightarrow{\tau}_h, \qquad \hat{\theta} := \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_M \end{bmatrix},$$

and since A_h is invertible for sufficiently small h we have

$$\overrightarrow{E}_h = -A_h^{-1}C(\hat{\theta})\overrightarrow{E}_h - A_h^{-1}\overrightarrow{\tau}_h.$$

For the piecewise constant functions E_h and τ_h on $[0,1]$, taking 2-norms we obtain

$$\begin{aligned}\|E_h\|_2 = h^{\frac{1}{2}}\|\overrightarrow{E}_h\|_2 &\leq h^{\frac{1}{2}}\|A_h^{-1}\|_2\left(\|C(\hat{\theta})\overrightarrow{E}_h\|_2 + \|\overrightarrow{\tau}_h\|_2\right)\\ &\leq h^{\frac{1}{2}}\ \|A_h^{-1}\|_2\left(\max_{1\leq m\leq M}|\cos(\hat{\theta}_m)|\|\overrightarrow{E}_h\|_2 + \|\overrightarrow{\tau}_h\|_2\right)\\ &= \|A_h^{-1}\|_2\left(\max_{1\leq m\leq M}|\cos(\hat{\theta}_m)|\|E_h\|_2 + \|\tau_h\|_2\right)\end{aligned}$$

and so

$$\|E_h\|_2 \leq \|A_h^{-1}\|_2\left(\|E_h\|_2 + \|\tau_h\|_2\right).$$

Since $\|A_h^{-1}\|_2 = \frac{1}{\pi^2} + \mathcal{O}(h^2)$, for h sufficiently small $\|A_h^{-1}\|_2 < 1$. We can then write

$$\|E_h\|_2 \leq \frac{\|A_h^{-1}\|_2}{1-\|A_h^{-1}\|_2}\|\tau_h\|_2.$$

Using $\|A_h^{-1}\|_2 = \frac{1}{\pi^2} + \mathcal{O}(h^2)$ and Taylor theorem we find

$$\frac{\|A_h^{-1}\|_2}{1-\|A_h^{-1}\|_2} = \frac{1}{\pi^2-1} + \mathcal{O}(h^2),$$

so

$$\|E_h\|_2 \leq \frac{1}{\pi^2-1}\|\tau_h\|_2 + \mathcal{O}(h^2).$$

By the consistency of the method proved in point **b**) $\lim_{h\to 0}\|\tau_h\|_2 = 0$ and so by the bound we have found $\lim_{h\to 0}\|E_h\|_2 = 0$ and this concludes the proof of convergence.

APPENDIX

This appendix contains useful formulae and theorems to solve the exam questions.

Householder transformations

An Householder transformation is a matrix of the form

$$P = I - 2ww^T,$$

where I is the identity matrix $n \times n$, $w \in \mathbb{R}^n$ with w of 2-norm equal to 1. Given $x \in \mathbb{R}^n$ we can define an Householder transformation such that

$$Px = \gamma\, \mathbf{e}_1, \qquad \gamma \in \mathbb{R}$$

and $\mathbf{e}_1$ the first canonical vector. This can be achieved by taking $w = \tilde{w}/\|\tilde{w}\|_2$ and

$$\tilde{w} = x \pm \|x\|_2 \mathbf{e}_1. \tag{3}$$

To solve the exercises, both plus and minus can be used. For the sake of simplicity, it is advisable to choose the sign that gives the simplest calculations.

Intermediate value theorem

Theorem *Suppose that f is a real-valued function, defined and continuous on the closed interval $[a, b]$ of $\mathbb{R}$. Then, f is a bounded function on the interval $[a, b]$ and, if y is any number such that*

$$\inf_{x\in[a,b]} f(x) \leq y \leq \sup_{x\in[a,b]} f(x)$$

then there is a number $\xi \in [a, b]$ such that $f(\xi) = y$. In particular the infimum and supremum of f are achieved, and can be replaced by $\min_{x\in[a,b]}$ and $\max_{x\in[a,b]}$, respectively.

Taylor's Theorem

Theorem *Suppose that n is a nonnegative integer, and f is a real-valued function, defined and continuous on the closed interval $[a, b]$ of $\mathbb{R}$, such that the derivatives of f of order up to and including n are defined and continuous on the closed interval $[a, b]$. Suppose further that $f^{(n)}$ is differentiable on the open interval (a, b). Then,*

for each value of x in $[a, b]$, there exist a number $\xi = \xi(x)$ in the open interval (a, b) such that

$$f(x) = f(a) + (x-a)f'(a) + \cdots + \frac{(x-a)^n}{n!} f^{(n)}(a) + \\ + \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(\xi).$$

Error formula for the interpolation polynomial

Suppose x_i, $i = 0, \ldots, n$ are distinct real numbers. Suppose $p_n(x)$ is the interpolation polynomial of degere less than or equal to n interpolating the data $(x_0, f(x_0)), \ldots, (x_n, f(x_n))$ on the interpolation nodes $x_0, \ldots, x_n$.

Theorem *Suppose that $n \geq 0$, and that f is a real valued function, defined and continuous on the closed real interval $[a, b]$, such that the derivative of f and of order $n+1$ exists and is continuous on $[a, b]$. Then, given that $x \in [a, b]$, there exists $\xi = \xi(x)$ in (a, b) such that*

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \pi_{n+1}(x),$$

where

$$\pi_{n+1}(x) = (x - x_0) \cdot (x - x_1) \cdots (x - x_n).$$