NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for

TMA4212 Numerical solution of differential equations with difference methods

Academic contact during examination: Elena Celledoni

Phone: 48238584

Examination date: 28. May 2016

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

Permitted examination support material: C: Approved simple pocket calculator is allowed. Rottman is allowed.

Language: English

Number of pages: 8

Number of pages enclosed: 2

Checked by:

Date Signature

The learning outcome has been published on the course webpage and on the official description of the course. The seven learning goals **L1** to **L7** are reported in the appendix. Learning outcome **L6**, **L3** and to some extent **L4** have been tested through the project work. We here test further the achievement of **L4** as well as **L1, L2, L5 L7**. All answers must be properly argued for.

Problem 1 (**L2**)

We are solving the Poisson equation

$$\Delta u = u_{xx} + u_{yy} = f, \qquad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega,$$

with the finite element method. $\Omega \subset \mathbf{R}^2$ is a rectangular domain with the sides aligned with the x and y axes and one corner in the origin. We use square elements and quadratic basis functions. We consider the element K with vertices $(0,0)$, $(h,0)$, $(0,h)$ and (h,h) where h is a discretization parameter.

a) Find an expression for the four quadratic, finite element basis functions $\varphi_1 = \varphi_{(0,0)}$, $\varphi_2 = \varphi_{(h,0)}$, $\varphi_3 = \varphi_{(0,h)}$ and $\varphi_4 = \varphi_{(h,h)}$ on K, by combining appropriately the linear polynomials

$$\frac{h-x}{h}, \quad \frac{x}{h}, \quad \frac{h-y}{h}, \quad \frac{y}{h}.$$

Solution The quadratic basis functions on K are

$$\varphi_1 = \frac{(x-h)(y-h)}{h^2}, \quad \varphi_2 = \frac{x(h-y)}{h^2}, \quad \varphi_3 = \frac{(h-x)y}{h^2}, \quad \varphi_4 = \frac{xy}{h^2}.$$

b) Find the bilinear function α arising in the Galerkin formulation of the Poisson equation. The element stiffness matrix is

$$A^K = \{\alpha_K(\varphi_i, \varphi_j)\}_{i,j=1,\ldots,4},$$

where α_K denotes the restriction of α to the element K. Find the elements $A^K_{2,4}$ and $A^K_{4,2}$ of this matrix.

Solution The bilinear function arising in the weak formulation of the Poisson equation with homogeneous Dirichlet boundary conditions is

$$-\alpha(u,v) = \int_0^1 \int_0^1 \nabla u \cdot \nabla v \, dx \, dy,$$

and α_K is its restriction to the element K. Since α is symmetric, we have

$$A^K_{2,4} = \alpha_K(\varphi_4, \varphi_2) = \alpha_K(\varphi_2, \varphi_4) = A^K_{4,2}.$$

We also have

$$\alpha_K(\varphi_4, \varphi_2) = -\int_0^h \int_0^h \nabla\varphi_4 \cdot \nabla\varphi_2 \, dx \, dy,$$

and

$$\nabla\varphi_4 = \begin{bmatrix} \frac{y}{h^2} \\ \frac{x}{h^2} \end{bmatrix}, \quad \nabla\varphi_2 = \begin{bmatrix} \frac{h-y}{h^2} \\ -\frac{x}{h^2} \end{bmatrix}, \qquad \nabla\varphi_4 \cdot \nabla\varphi_2 = \frac{yh - y^2}{h^4} - \frac{x^2}{h^4}.$$

So after integration

$$A^K_{2,4} = A^K_{4,2} = \alpha_K(\varphi_4, \varphi_2) = \frac{1}{6}.$$

Problem 2 (**L1, L3**)

Consider the linear advection equation

$$u_t + a u_x = 0, \quad x \in \mathbf{R}, \quad u(x,0) = u_0(x)$$

with a constant. Consider the two schemes

$$\frac{u^{n+1}_m - u^n_m}{\Delta t} + a\frac{u^n_{m+1} - u^n_m}{\Delta x} = 0, \tag{1}$$

$$\frac{u^{n+1}_m - u^n_m}{\Delta t} + a\frac{u^n_m - u^n_{m-1}}{\Delta x} = 0. \tag{2}$$

a) Which one of the two schemes would you use to approximate this equation when $a > 0$ and which one when $a < 0$ and why?

Solution If $a < 0$ we would use (1). If $a > 0$ we would use (2). This is because according to the CFL condition it is necessary that the characteristics of the equation (i.e. the lines $x(t) = x_0 + ta$) lie in the domain of dependence of the method in order to have convergence. Fixed a point (x^*, t^*) in the (x,t)-plane, for (1) the domain of dependence is a triangle with vertices

(x^*, t^*), $(x^*, 0)$ and $(x^* + t^* \frac{\Delta x}{\Delta t}, 0)$, and for (2) the domain of dependence is a triangle with vertices (x^*, t^*), $(x^*, 0)$ and $(x^* - t^* \frac{\Delta x}{\Delta t}, 0)$. Let $p = \frac{\Delta t}{\Delta x}$. For (1) and with $a < 0$, if $0 \leq (-a)p \leq 1$, the characteristic line through (x^*, t^*) is contained in the domain of dependence. Similarly, for (2) and for $a > 0$, if $0 \leq ap \leq 1$, the characteristic line through (x^*, t^*) is contained in the triangle with vertices (x^*, t^*), $(x^*, 0)$ and $(x^* - t^* \frac{\Delta x}{\Delta t}, 0)$. So the necessary condition for convergence is satisfied.

b) Perform a Von Neumann stability analysis for the scheme (1).

Solution. See chapter 7.4 in the note.

Problem 3 (**L5, L7**)

a) The 2×2 matrix A is symmetric and positive definite. Show that the Jacobi iteration for A converges. For the Jacobi iteration see the appendix.

Solution We assume the 2×2 symmetric matrix A has the form

$$A = \begin{bmatrix} a_{1,1} & a \\ a & a_{2,2} \end{bmatrix}.$$

Since A is positive definite $0 < \det(A) = a_{1,1}a_{2,2} - a^2$, and $a^2 < a_{1,1}a_{2,2}$. If we split $A = D - R$ where D is the diagonal of A, the iteration matrix for the Jacobi method is

$$D^{-1}R = \begin{bmatrix} 0 & \frac{a}{a_{1,1}} \\ \frac{a}{a_{2,2}} & 0 \end{bmatrix}.$$

To ensure convergence we need to have $\rho(D^{-1}R) < 1$ where $\rho(D^{-1}R)$ is the spectral radius of $D^{-1}R$. The eigenvalues of $D^{-1}R$ are

$$\lambda = \pm \frac{a}{\sqrt{a_{1,1}a_{2,2}}}$$

and since $a^2 < a_{1,1}a_{2,2}$, $|\lambda| < 1$ and $\rho(D^{-1}R) < 1$. So convergence of the Jacobi iteration is always guaranteed for a symmetric positive definite 2×2 matrix.

b) The $N \times N$ matrix E has all its elements equal to 1. Show that one of the eigenvalues of E is N, and all the others are zero. Construct a matrix $A = I + \kappa E$, where κ is a constant to be determined, such that A is symmetric and positive definite, but in general the Jacobi method diverges.

Solution We can write $E = \mathbf{e}\,\mathbf{e}^T$. With $\mathbf{e}$ the vector with N components equal to 1. So E is a rank one matrix, $\mathbf{e}$ is an eigenvector of E with eigenvalue N (because $\mathbf{e}\,\mathbf{e}^T\mathbf{e} = N\mathbf{e}$) and all the other eigenvalues are zero. The eigenvalues of $I + \kappa\, E$ are

$$\lambda(I + \kappa\, E) = \begin{cases} 1 + \kappa N \\ 1 \end{cases}$$

and for $I + \kappa\, E$ to be positive definite, κ needs to be either positive or if κ is negative $\frac{1}{N} > |\kappa|$. The iteration matrix of the Jacobi method is simply

$$D^{-1}R = \frac{\kappa}{1+\kappa}\,(I - E)$$

with eigenvalues

$$\lambda\left(\frac{\kappa}{1+\kappa}\,(I - E)\right) = \begin{cases} +\frac{\kappa}{1+\kappa}(1 - N) \\ \frac{\kappa}{1+\kappa} \end{cases}$$

so

$$\rho\left(\frac{\kappa}{1+\kappa}\,(I - E)\right) = \left|\frac{\kappa}{1+\kappa}\right|(N - 1).$$

Excluding the case $\kappa = 0$ for which $A = I + \kappa E$ is the identity matrix and the solution of the linear system is already given, for negative and positive values of κ we have

(i) when $\kappa < 0$ and $|\kappa| < \frac{1}{N}$ then the spectral radius is always less than 1

$$\rho(\frac{\kappa}{1+\kappa}\,(I - E)) = \left|\frac{1}{\frac{1}{k} + 1}\right|(N - 1) < 1$$

because $|\frac{1}{k} + 1| > N - 1$, and the Jacobi method converges without any further restrictions on κ;

(ii) when $\kappa > 0$

$$\rho(\frac{\kappa}{1+\kappa}\,(I - E)) = \frac{\kappa}{1+\kappa}(N - 1) < 1 \Leftrightarrow \kappa < \frac{1}{N-2},$$

so for $k \geq \frac{1}{N-2}$ and $N > 2$ the Jacobi method does not converge (for any x^0) even though $I + \kappa E$ is symmetric and positive definite (this provides an answer to the exam question **b**)).

c) Explain why the result of **b)** does not contradict point **a)** above in this exercise.

Solution The result of **b)** (ii) is valid for $N > 2$. If $N = 2$ the symmetric matrix $I + \kappa E$ is positive definite when

$$1 + 2\kappa > 0.$$

So either for $\kappa > 0$ or for κ negative and $|\kappa| < \frac{1}{2}$. For the spectral radius of the iteration matrix we have

$$\rho(\frac{\kappa}{1+\kappa}(I - E)) = \left|\pm\frac{\kappa}{1+\kappa}\right| = \left|\pm\frac{1}{\frac{1}{\kappa}+1}\right| < 1$$

is always less than one both in the case when $\kappa > 0$ or when κ negative and $|\kappa| < \frac{1}{2}$ (ie. whenever $I + \kappa E$ is positive definite). So if $N = 2$, we cannot find any κ such that the matrix $I + \kappa E$ is symmetric and positive definite and at the same time the Jacobi method does not converge (this is consistent with point **a)**).

Problem 4 (**L1, L2, L4, L7**)
Consider the boundary value problem

$$-p_0 u'' + r_0 u = f(x), \quad u(0) = 0, \quad u(1) = 0, \tag{3}$$

on the interval $[0, 1]$, where p_0 and r_0 are positive constants and $f \in C^4[0, 1]$. Use equally spaced points

$$x_i = ih, \quad i = 0, 1, \ldots, n, \quad \text{with } h = \frac{1}{n},\ n \geq 2,$$

and the standard piecewise linear finite element basis functions (hat functions) φ_i, $i = 1, 2, \ldots, n-1$.

a) State the weak formulation of the problem and the Galerkin method and show that the finite element equations for $u_i = u^h(x_i)$ become

$$-p_0\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + r_0\frac{u_{i-1} + 4u_i + u_{i+1}}{6} = \frac{1}{h}\langle f, \varphi_i\rangle \tag{4}$$

for $i = 1, 2, \ldots, n-1$, with $u_0 = 0$ and $u_n = 0$.

Solution Weak formulation: for $v \in H_0^1$ consider

$$-p_0\int_0^1 u'' v dx + r_0\int_0^1 uv dx = \int_0^1 fv dx$$

using integration by parts we get

$$p_0 \int_0^1 u^{'} v^{'} dx + r_0 \int_0^1 uvdx = \int_0^1 fvdx = \langle f, v \rangle,$$

and defining

$$A(u,v) := p_0 \int_0^1 u^{'} v^{'} dx + r_0 \int_0^1 uvdx, \quad \forall u, v \in H_0^1$$

we get the weak formulation. Which is:
Find $u \in H_0^1$ such that

$$A(u,v) = \langle f, v \rangle, \quad \forall v \in H_0^1.$$

Denoting with $\varphi_1, \ldots, \varphi_{n-1}$ the linear finite element basis, where

$$\varphi_j = \begin{cases} \frac{x - x_{j-1}}{h} & x_{j-1} \leq x \leq x_j \\ \frac{x_{j+1} - x}{h} & x_j \leq x \leq x_{j+1} \\ 0 & \text{otherwise.} \end{cases}$$

we can state the Galerkin method: Find $u^h = \sum_{j=1}^{n-1} u_j^h \varphi_j \in H_0^1$ such that

$$A(u^h, \varphi_i) = \langle f, \varphi_i \rangle, \quad i = 1, \ldots, n-1.$$

This is equivalent to the linear system

$$MU = b$$

where M is $(n-1) \times (n-1)$ symmetric and with entries

$$M_{i,j} = A(\varphi_j, \varphi_i)$$

and $U = [u_1^h, \ldots, u_{n-1}^h]^T$, $b = [\langle f, \varphi_1 \rangle, \ldots, \langle f, \varphi_{n-1} \rangle]^T$. We need to find the elements of M for row i. We have

$$\varphi_j^{'} = \begin{cases} \frac{1}{h} & x_{j-1} \leq x \leq x_j \\ -\frac{1}{h} & x_j \leq x \leq x_{j+1} \\ 0 & \text{otherwise.} \end{cases}$$

Using the formula for $A(\cdot, \cdot)$ and computing the integrals we obtain

$$M_{i,i} = A(\varphi_i, \varphi_i) = 2p_0 \frac{1}{h} + r_0 h \frac{2}{3}, M_{i,i+1} = M_{i-1,i} = -p_0 \frac{1}{h} + r_0 \frac{1}{6} h,$$

while $M_{i,j} = 0$ if $j \geq i+2$, $j \leq i-2$. Substituting these values in

$$\sum_{j=1}^{n-1} M_{i,j} u_j^h = b_i,$$

we obtain (4).

b) By expanding in Taylor series we have obtained that

$$\frac{1}{h}\langle f, \varphi_i \rangle = f(x_i) + \frac{1}{12}h^2 f''(x_i) + \mathcal{O}(h^4). \tag{5}$$

Interpreting (4) as a finite difference approximation to the boundary value problem, and using (5), show that the corresponding local truncation error τ_i satisfies

$$\tau_i = \frac{1}{12}h^2 r_0 u''(x_i) + \mathcal{O}(h^4), \quad i = 1, \ldots, n-1.$$

Solution The local truncation error τ_i satisfies

$$-p_0 \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} + r_0 \frac{u(x_{i-1}) + 4u(x_i) + u(x_{i+1})}{6} = f(x_i) + \frac{1}{12}h^2 f''(x_i) + \mathcal{O}(h^4) + \tau_i \tag{6}$$

using (3) at $x = x_i$ and rearranging the terms, this simplifies to

$$-p_0 \frac{h^2}{12} u''''(x_i) + r_0 \frac{h^2}{6} u''(x_i) - \frac{1}{12}h^2 f''(x_i) + \mathcal{O}(h^4) = \tau_i.$$

Differentiating (3) twice we obtain $-p_0 u'''' + r_0 u'' = f''$, which we can substitute in the previous expression to finally obtain

$$\tau_i = r_0 \frac{h^2}{12} u''(x_i) + \mathcal{O}(h^4).$$

c) Show finally the following bound for the error

$$\max_{0 \leq i \leq n} |u(x_i) - u^h(x_i)| \leq M h^2,$$

where M is a positive constant.

Solution By subtracting (4) from the equation for the local truncation error, we obtain the equation for the error at x_i, i.e. the equation for $e_i = u(x_i) - u_i^h$, this is

$$-p_0 \frac{e_{i-1} - 2e_i + e_{i+1}}{h^2} + r_0 \frac{e_{i-1} + 4e_i + e_{i+1}}{6} = -\tau_i$$

$i = 1, \ldots, n-1$. We can rewrite it as

$$(12p_0 + 4r_0h^2)e_i = (6p_0 - r_0h^2)(e_{i-1} + e_{i+1}) - 6h^2\tau_i$$

taking maxima over i at the left hand side we have

$$(12p_0 + 4r_0h^2)|e_i| \leq 2|6p_0 - r_0h^2| \max_i |e_i| + 6h^2 \max_i |\tau_i|$$

for all i, and then

$$(12p_0 + 4r_0h^2) \max_i |e_i| \leq 2|6p_0 - r_0h^2| \max_i |e_i| + 6h^2 \max_i |\tau_i|$$

and

$$(12p_0 + 4r_0h^2) \max_i |e_i| \leq 2(6p_0 + r_0h^2) \max_i |e_i| + 6h^2 \max_i |\tau_i|$$

leading to

$$(12p_0 + 4r_0h^2 - 12p_0 - 2r_0h^2) \max_i |e_i| \leq 6h^2 \max_i |\tau_i|$$

and

$$r_0h^2 \max_i |e_i| \leq 3h^2 \max_i |\tau_i|.$$

Finally, using the obtained expression for τ_i and assuming boundedness of the derivatives of u, we get

$$\max_i |e_i| \leq \frac{1}{4} \max_{x \in [0,1]} |u''(x_i)|h^2 = Kh^2.$$

Appendix

- $$\frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} = u''(x_i) + \frac{h^2}{12}u''''(x_i) + \mathcal{O}(h^4)$$

- **Jacobi iteration**: given the linear system of equations

  $$Ax = b$$

  with A $n \times n$ matrix and b a vector with n components, we split A as the sum of its diagonal D minus a matrix R:

  $$A = D - R.$$

  The Jacobi iteration is an iterative method to approximate the solution of the linear system, and is given by the iteration

  $$x^{k+1} = D^{-1}(R\,x^k + b), \tag{7}$$

  with x^0 a given initial guess. Note that (7) this is a fixed point iteration to solve the fixed point equation $x = D^{-1}(R\,x + b)$, whose solution is the same of the linear system.

Learning outcome:

Knowledge	**L1**	Understanding of error analysis of difference methods: consistency, stability, convergence of difference schemes.
	L2	Understanding of the basics of the finite element method.
Skills	**L3**	Ability to choose and implement a suitable discretization scheme given a particular PDE, and to design numerical tests in order to verify the correctness of the code and the order of the method.
	L4	Ability to analyze the chosen discretization scheme, at least for simple PDE-test problems.
	L5	Ability to attack the numerical linear algebra challenges arising in the numerical solution of PDEs.
General competence	**L6**	Ability to present in oral and written form the numerical and analytical results obtained in the project work.
	L7	Ability to apply acquired mathematical knowledge in linear algebra and calculus to achieve the other goals of the course.