Degree Master of Science in Mathematical Modelling and Scientific Computing

Numerical Solution of Differential Equations & Numerical Linear Algebra

Friday 17th January 2014, 9:30 p.m. – 11:30 a.m.

Candidates should submit answers to a maximum of four questions that include an answer to at least one question in each section.

Please start the answer to each question on a new page.

All questions will carry equal marks.

Do not turn over until told that you may do so.

Section A —Numerical Solution of Differential Equations

Question 1

(a) Suppose that $\theta \in [0, 1]$ and let N be an integer, $N \geq 2$. Consider the one-step method

$$y_{n+1} = y_n + hf(x_n + \theta h, y_n + \theta hf(x_n, y_n)), \qquad n = 0, 1, \ldots, N-1, \qquad (*)$$

for the numerical solution of the initial-value problem $y' = f(x, y)$, $y(x_0) = y_0$, over the uniform computational mesh $\{x_n : x_n := nh, n = 0, 1, \ldots, N\}$ of spacing $h := X/N > 0$, with $X > 0$, contained in the closed interval $[0, X]$ of the real line.

Define the truncation error T_n of the method. Show that there exists a $\theta_0 \in [0, 1]$ such that

$$T_n = \frac{1}{24}h^2 \left(y'''(x_n) + 3f_y(x_n, y(x_n))y''(x_n)\right) + \mathcal{O}(h^3).$$

By considering $f(x, y) \equiv \lambda y$ where λ is a *nonzero* real number, show that there is no $r > 2$ such that $T_n = \mathcal{O}(h^r)$ as $h \to 0$, $n \to \infty$, with $nh = x_n$. Hence deduce that for $\theta = \theta_0$ the method $(*)$ is second order accurate.

[12 + 4 marks]

(b) Consider the initial-value problem

$$y' = \lambda y, \qquad y(0) = y_0, \qquad (**)$$

where $\lambda \in \mathbb{C}$ with $\mathrm{Re}(\lambda) < 0$, and $y_0 \in \mathbb{C}$. Determine $\lim_{x\to+\infty} y(x)$.

Apply the method $(*)$ to the initial value problem $(**)$ over the computational mesh $\{x_n : x_n := nh, n = 0, 1, \ldots\}$, $h > 0$, with $y_n \in \mathbb{C}$ denoting the numerical approximation to $y(x_n) \in \mathbb{C}$ at $x = x_n$. Show that there exists a complex number $z = z(\lambda h, \theta)$ such that $y_{n+1} = z(\lambda h, \theta)y_n$ for $n = 0, 1, \ldots$.

Express, in terms of $|z(\lambda h, \theta)|$, the region of absolute stability

$$\mathcal{H} := \{\lambda h \in \mathbb{C} : \lim_{n\to+\infty} y_n = 0 \quad \text{for any } y_0 \in \mathbb{C}\}.$$

[2 + 3 + 4 marks]

TURN OVER

Question 2

State the general form of a linear k-step method for the numerical solution of the initial-value problem $y' = f(x, y)$, $y(x_0) = y_0$ on a nonempty, closed interval $[x_0, X]$ of the real line. **[2 marks]**

(a) Define the *truncation error* T_n of the method.

[2 marks]

(b) What does it mean to say that the method is:

(i) *consistent*;

[2 marks]

(ii) *zero-stable*;

[2 marks]

(iii) *convergent*?

[2 marks]

(c) State the *root condition*, relating zero-stability of a linear k-step method to the roots of a certain kth degree polynomial. **[2 marks]**

(d) State Dahlquist's theorem (without proof). Using this theorem, show that there is unique value of the parameter $\alpha \in \mathbb{R}$ such that the three-step method

$$y_{n+3} = y_{n+2} + \frac{h}{12}\left[\, 23f(x_{n+2}, y_{n+2}) - \alpha f(x_{n+1}, y_{n+1}) + 5f(x_n, y_n)\,\right]$$

is convergent.

[13 marks]

Question 3

Consider the initial-value problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - u, \qquad -\infty < x < \infty, \quad 0 < t \leq T,$$
$$u(x,0) = u_0(x), \qquad -\infty < x < \infty,$$

where T is a fixed real number and u_0 is a real-valued, bounded and continuous function of $x \in (-\infty, \infty)$.

(a) Formulate the forward Euler scheme for the numerical solution of this initial-value problem on a mesh with uniform spacings $\Delta x > 0$ and $\Delta t := T/M$ in the x and t co-ordinate directions, respectively, where M is a positive integer.

[7 marks]

(b) Let U_j^m denote the numerical approximation to $u(j\Delta x, m\Delta t)$ computed by the explicit Euler scheme, $0 \leq m \leq M$, $j \in \mathbb{Z}$, where $\mathbb{Z}$ denotes the set of all integers. Suppose that $\|U^0\|_{\ell_2} := \left(\Delta x \sum_{j\in\mathbb{Z}} |U_j^0|^2\right)^{1/2}$ is finite. By taking the (semi-discrete) Fourier transform of the explicit Euler scheme, show that

$$\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}$$

for all m, $1 \leq m \leq M$, provided that $\Delta t \leq \frac{2(\Delta x)^2}{4+(\Delta x)^2}$.

Deduce that the explicit Euler scheme is *conditionally stable* in the $\|\cdot\|_{\ell_2}$ norm.

[9 marks]

(c) Let U_j^m denote the numerical approximation to $u(j\Delta x, m\Delta t)$ computed by the explicit Euler scheme, $0 \leq m \leq M$, $j \in \mathbb{Z}$, where $\mathbb{Z}$ denotes the set of all integers. Suppose that $\|U^0\|_{\ell_\infty} := \max_{j\in\mathbb{Z}} |U_j^0|$ is finite. Show that

$$\|U^m\|_{\ell_\infty} \leq \|U^0\|_{\ell_\infty}$$

for all m, $1 \leq m \leq M$, provided that $\Delta t \leq \frac{(\Delta x)^2}{2+(\Delta x)^2}$.

Deduce that the explicit Euler scheme is *conditionally stable* in the $\|\cdot\|_{\ell_\infty}$ norm.

[9 marks]

Question 4

Suppose that $u_0 : \mathbb{R} \to \mathbb{R}$ is a bounded continuous function and $f : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function such that f' is monotonic increasing.

(a) Formulate the upwind finite difference scheme for the numerical solution of the nonlinear hyperbolic partial differential equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}[f(u)] = 0, \qquad x \in \mathbb{R}, \quad t > 0,$$

subject to the initial condition $u(x, 0) = u_0(x)$ for $x \in \mathbb{R}$, on a uniform finite difference mesh of spacing Δx in the x-direction and spacing Δt in the t-direction.

[5 marks]

(b) Let $\mathbb{Z}$ denote the set of all integers and let U_j^m denote the upwind finite difference approximation to u at the mesh point (x_j, t_m), where $x_j = j\Delta x$, $j \in \mathbb{Z}$ and $t_m = m\Delta t$, $m = 0, 1, 2, \ldots$. Show by induction that, if

$$\left(\max_{x \in \mathbb{R}} |f'(u_0(x))| \right) \frac{\Delta t}{\Delta x} \leq 1,$$

then the following inequalities hold:

(i)

$$\left(\max_{j \in \mathbb{Z}} |f'(U_j^m)| \right) \frac{\Delta t}{\Delta x} \leq 1 \qquad \text{for all } m = 0, 1, 2, \ldots \, ;$$

[10 marks]

(ii)

$$\max_{j \in \mathbb{Z}} |U_j^{m+1}| \leq \max_{j \in \mathbb{Z}} |U_j^m| \qquad \text{for all } m = 0, 1, 2, \ldots \, .$$

[10 marks]

Section B — Numerical Linear Algebra

Question 5

(a) Let u_i be linearly independent nonzero $m \times 1$ vectors for $i = 1, 2$ and let v_i be linearly independent nonzero $n \times 1$ vectors for $i = 1, 2$. Let $A_1 = u_1 v_1^T$ and $A_2 = u_1 v_1^T + u_2 v_2^T$. What is the QR factorization of A_1? What condition on v_1 and v_2 is needed so that the QR decomposition of A_2 gives rise to a matrix R with nonzero diagonal entries.

[2+3 marks]

(b) State the algorithm for Householder triangularization for computing the R matrix in the QR factorization of an $m \times n$ matrix A, and derive an asymptotic estimate of the number of floating point operations needed to compute R (as a function of m and n).

[4+4 marks]

(c) Let x be a fixed nonzero vector and let P and Q be unitary matrices. If $Px = Qx$, does it follow that $P = Q$? Provide a proof or construct a counter-example.

[5 marks]

(d) Consider the floating point multiplication of an invertible matrix A by an invertible matrix B that is stored exactly in floating point format. Instead of the matrix product AB a computer will multiply B by $A + \delta A$. Bound the relative error of the resulting matrix product in terms of the condition numbers of the matrices and the relative size of δA compared to the size of A. Also, show that the result of $(A + \delta A)B$ is equal to $A(B + \delta B)$ for some δB whose relative size compared to B can be bounded in terms of matrix condition numbers.

[4+3 marks]

Question 6

(a) Let B and C be full rank $m \times n$ matrices with $m > n/2$ which satisfy $(Bx)^T(Cy) = 0$ for all $x, y \in \mathbb{R}^n$. Let $F = [B\ C]$ be the $m \times 2n$ matrix formed by appending the matrix C to B. Explain any special structure in the R matrix of the QR factorization of F. How might this be used in practice for a faster QR factorization of F?

[7 marks]

(b) Let the $m \times m$ matrix A be decomposed into its upper triangular, diagonal, and lower triangular components $A = U + D + L$; that is, $D_{ii} = A_{ii} \neq 0$ and $D_{ij} = 0$ otherwise, $U_{ij} = A_{ij}$ for $j > i$ and $U_{ij} = 0$ otherwise, and $L_{ij} = A_{ij}$ for $i > j$ and $L_{ij} = 0$ otherwise. Assume that $(U + D)$ and $(D + L)$ are both invertible and let

$$\alpha_i = \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|}$$

and

$$\beta_i = \sum_{j=i+1}^{m} \frac{|a_{ij}|}{|a_{ii}|}.$$

State the Gauss-Seidel and Jacobi methods for iteratively computing approximate solutions to the linear system $Ax = b$.

Let $e_n^J = x - x_n^J$ be the error of the Jacobi method. Show that $\|e_{n+1}^J\|_\infty \leq \|e_n^J\|_\infty \max_i(\alpha_i + \beta_i)$. Show that if $\max_i(\alpha_i + \beta_i) < 1$ then the error for the Gauss-Seidel method, $e^{GS} = x - x_n^{GS}$, satisfies the relation $\|e_{n+1}^{GS}\|_\infty \leq \gamma \|e_n^{GS}\|_\infty$ for some $\gamma \leq \max_i(\alpha_i + \beta_i)$.

[3+7 marks]

(c) Let an $m \times m$ matrix L_k have entries that are zero except for: $L_k(i,i) = 1$ for all i and $L_k(i,k)$ which may be nonzero for $i > k$. What are the entries of L_k^{-1}? What are the entries of $L_1^{-1}L_2^{-1} \cdots L_{m-1}^{-1}$? Explain why this is important in the floating point operation cost of computing the LU factorization of an $m \times m$ matrix A. Derive the leading term of the floating point operation cost of computing the LU factorization of a matrix that does not require pivoting at any stage.

[2+2+4 marks]