MA3650 - Numerical Methods for Differential Equations

Luke Dando

March 26, 2021

Contents

1 Revision of Taylor series and Newton's method

1.1 Lecture 1

1.1.1 Newton's method

Let x_0 be an initial guess of a root to a function $f(x)$. Then

$$x_{n+1} = x_n - \frac{f(x)}{f'(x)}, \quad n \geq 0. \tag{1}$$

1.2 Lecture 2

1.2.1 IVT: Intermediate Value Theorem

Theorem 1 *Let $f : [a, b] \to \mathbb{R}$ be a continuous function, then f achieves all values of the interval $[f(a), f(b)]$. As a corollary, if $f(a) \cdot f(b) < 0$, then f has a root on (a, b) (in the interior of $[a, b]$).*

1.2.2 Complex convergence

Lemma 1 *The map $g : (0, \infty) \to \mathbb{R}$, $g(x) = ln(e/x)$, maps (e, ∞) into $(-\infty, 0)$ and $(0, e)$ to $(0, e)$.*

Corollary 1 *Newton's method for $f(x) = ln(x)$ is possible for for any $x_0 \in (0, e)$, and impossible when $x_0 > e$.*

2 Convergence of Newton's method

2.1 Lecture 3

2.1.1 Babylonian iteration

Let $\alpha > 0$. To get an approximate value for $\sqrt{\alpha}$, start with a crude guess x_1 and improve on the approximation using

$$x_{n+1} = \frac{(x_n + \alpha/x_n)}{2}, \quad n \geq 1. \tag{2}$$

2.1.2 Error estimates

Lemma 2 *Let x_n be given by eq. (2). Then,*

$$x_{n+1} - \sqrt{\alpha} = (x_n - \sqrt{\alpha})^2/2x_n. \tag{3}$$

2.1.3 Convergence close to the square root

When $x_n \approx \sqrt{\alpha}$,

$$(x_{n+1} - \sqrt{\alpha}) = \frac{(x_n - \sqrt{\alpha})^2}{2x_n} \approx \frac{(x_n - \sqrt{\alpha})^2}{2\sqrt{\alpha}} \tag{4}$$

We define the error e_n (at iterate n as

$$e_n = x_n - \sqrt{\alpha}, \tag{5}$$

allowing us to rewrite the estimate eq. (3) as

$$e_{n+1} \approx \frac{e_n^2}{2\sqrt{\alpha}} \tag{6}$$

2.2 Lecture 4

2.2.1 Newton's method theory

Theorem 2 *Let $I = [a, b]$ and $f : I \to \mathbb{R}$ be twice continuously differentiable. Suppose that*

$$f(a) \cdot f(b) < 0$$

and that there are constants m and M, such that

$$0 < m \leq |f'(x)| \quad \textit{and} \quad |f''(x)| \leq M \tag{7}$$

for all $x \in I$. Let $K = M/2m$. Then, choose a root r of f in I and $0 < \delta < 1/K$ such that $J = [r-\delta, r+\delta] \subseteq I$. Then, for any $x_0 \in J$, the sequence defined by Newton's method in eq. (1) belongs to J and $x_n{}_{n=1}^{\infty}$ converges to r. Moreover

$$|x_{n+1} - r| \leq K|x_n - r|^2, \quad n \geq 0, \tag{8}$$

so the convergence is quadratic.

3 Wrapping up chapter 1

3.1 Lecture 6

3.1.1 Absolute and relative error

Definition 1 *If a is a number and $\hat{a}$ is an approximation to a, the absolute error is $|a - \hat{a}|$ and the relative error is $\frac{|a-\hat{a}|}{|a|}$ provided $a \neq 0$.*

4 Finite differences

4.1 Lecture 7

4.1.1 Euler scheme

For every $0 \leq i \leq n$, we denote by v_i the approximation of $u(t_i)$, the value of the exact solution at $t = t_i$. We approximate by the forward difference

$$u'(t_i) \approx \frac{v_{i+1} - v_i}{h} \tag{9}$$

4.1.2 Two dimensional grids

Consider a function $u : \Omega \to \mathbb{R}$ where $\Omega = I \times J$ is the cartesian product of two real intervals, say $x \in I$ and $t \in J$. Although I and J can be infinite, numerical schemes consider finite grids, although they can be of arbitrary length. Let $I = [a, b]$ and $J = [c, d]$. We approximate u at grid points (x_i, t_j) such that

1. $x_i = a + i(\delta x)$, $0 \leq i \leq N$ with $N \cdot (\delta x) = b - a$, that is, $\delta x = (b - a)/N$. The quantity δx is the spatial step-size.

2. $t_j = c + j(\delta t)$, $0 \leq j \leq M$ with $m \cdot (\delta t) = d - c$, that is, $\delta t = (d - c)/M$. The quantity δt is the temporal step-size.

4.1.3 Standard approximations

We denote by $v_{i,j}$ the numerical scheme approximation of $u(x_i, t_j)$. We shall need the following approximations.

1. **Forward difference**:
$$\left(\frac{\partial u}{\partial x}\right)_{i,j} \approx \frac{v_{i+1,j} - v_{i,j}}{\delta x}, \tag{10}$$

2. **Backward difference**:
$$\left(\frac{\partial u}{\partial x}\right)_{i,j} \approx \frac{v_{i,j} - v_{i-1,j}}{\delta x}, \tag{11}$$

3. **Central (or symmetric) differences**:
$$\left(\frac{\partial u}{\partial x}\right)_{i,j} \approx \frac{v_{i+1,j} - v_{i-1,j}}{2(\delta x)}, \tag{12}$$

4. **Second order difference**:
$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \approx \frac{v_{i+1,j} + v_{i-1,j} - 2v_{i,j}}{(\delta x)^2}. \tag{13}$$

   This is obtained by using the forward difference of $\frac{\partial^2 u}{\partial x^2}$, where we use the backward difference for the first derivative at (x_{i+1}, t_j) and the forward difference for the derivative at (x_i, t_j).

4.2 Lecture 8

4.2.1 Errors for the standard approximation

The error for the forward difference approximation is

$$\left| f'(a) - \frac{f(a+h) - f(a)}{h} \right| = |f''(\xi)|\frac{h}{2} \leq \max_{x\in[a,a+h]}(|f''(x)|)\frac{h}{2}. \tag{14}$$

The error is linear in h.
For the central difference approximation, the error is

$$\left| f'(a) - \frac{f(a+h) - f(a-h)}{2h} \right| = |f'''(\xi) - f'''(\zeta)|\frac{h^2}{12} \leq \max_{x\in[a-h,a+h]}(|f'''(x)|)\frac{h^2}{6}. \tag{15}$$

The error is now quadratic in h.

4.2.2 Order symbol

Definition 2 *Let $f : I_0 \to \mathbb{R}$ be a real valued function, we say that f is of order not exceeding, or bounded with respect to, x^α as x tends to 0, denoted by*

$$f = O(x^\alpha),$$

if there exists $\alpha \in \mathbb{R}$, $C > 0$ such that

$$|f(x)| \leq C|x|^\alpha \tag{16}$$

for all $x \in I_0$.

4.2.3 Local error of standard finite differences

Consider a grid (x_i, t_j) of step-sizes δx and δt, respectively. Let u_j represent $u(x_i, t_j)$. The following relations are exact:

1. $\frac{\partial u}{\partial x}(x_i, t_j) = \frac{u_{i+1,j} - u_{i,j}}{\delta x} + O(\delta x)$,
2. $\frac{\partial u}{\partial x}(x_i, t_j) = \frac{u_{i,j} - u_{i-1,j}}{\delta x} + O(\delta x)$,
3. $\frac{\partial u}{\partial x}(x_i, t_j) = \frac{u_{i+1,j} - u_{i-1,j}}{2(\delta x)} + O((\delta x)^2)$,
4. $\frac{\partial u}{\partial x}(x_i, t_j) = \frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{(\delta x)^2} + O(\delta x)$.

5 Explicit numerical schemes for the heat equation

5.1 Lecture 9

5.1.1 Explicit scheme of the heat equation

$$\frac{v_{i,j+1} - v_{i,j}}{(\delta t)} = \frac{v_{i+1,j} + v_{i-1,j} - 2v_{i,j}}{(\delta x)^2},$$

can be rearranged to give

$$v_{i,j+1} = \rho v_{i-1,j} + (1-2\rho)v_{i,j} + \rho v_{i+1,j}, \quad 1 \le i \le N-1,\, 1 \le j \le M-1, \tag{17}$$

where the values of u at t_{j+1} are given from the values at t_j and ρ is the Courant number $\frac{\delta t}{(\delta x)^2}$.

5.2 Lecture 10

5.2.1 Matrix form for the explicit heat equation

We can rewrite eq. (17) as

$$\underline{v}_{j+1} = A\underline{v}_j, \quad 0 \le j \le M-1,$$

where A is the tridiagonal matrix

$$A = \begin{pmatrix} 1-2\rho & \rho & 0 & \cdots & 0 & 0 \\ \rho & 1-2\rho & \rho & \cdots & 0 & 0 \\ 0 & \rho & 1-2\rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \rho & 1-2\rho & \rho \\ 0 & 0 & 0 & \cdots & \rho & 1-2\rho \end{pmatrix}. \tag{18}$$

6 Further numerical schemes for the heat equation

6.1 Lecture 11

6.1.1 Implicit scheme of the heat equation

$$\frac{v_{i,j} - v_{i,j-1}}{(\delta t)} = \frac{v_{i+1,j} + v_{i-1,j} - 2v_{i,j}}{(\delta x)^2},$$

can be rearranged to give

$$-\rho v_{i-1,j} + (1+2\rho)v_{i,j} - \rho v_{i+1,j} = v_{i,j-1}, \quad 1 \le i \le N-1,\, 1 \le j \le M-1, \tag{19}$$

where the values of u at t_j are given from the values at t_{-1j} and ρ is the Courant number $\frac{\delta t}{(\delta x)^2}$.

6.1.2 Matrix form for the implicit heat equation

The scheme in eq. (19) can be written as

$$B\underline{v}_j = \underline{v}_{j-1}, \quad 1 \le j \le M,$$

where B is the tridiagonal matrix

$$B = \begin{pmatrix} 1+2\rho & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+2\rho & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+2\rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & -\rho & 1+2\rho & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1+2\rho \end{pmatrix}. \tag{20}$$

6.1.3 Crank-Nicholson Scheme

$$\frac{v_{i,j+1}-v_{i,j}}{(\delta t)}=\frac{1}{(\delta x)^2}[\theta(v_{i-1,j+1}+2v_{i,j+1}+v_{i+1,j})+(1-\theta)(v_{i-1,j}-2v_{i,j}+v_{i+1,j})], \tag{21}$$

where θ is a weight. The Crank-Nicholson scheme is written in matrix form by

$$C\underline{v}_{j+1}=D\underline{v}_j$$

where matrices C and D are determined as before.

7 Numerical schemes for complex boundary conditions

7.1 Lecture 13

7.1.1

8 Numerical schemes for complex boundary conditions 2

8.1 Lecture 14

8.2 Lecture 15

8.3 Lecture 16

9 Matrix norms

9.1 Lecture 17

9.1.1 Vector Norms

Definition 3 *A norm on $\mathbb{R}^n$ is a mapping $\|\ \|:\mathbb{R}^n\to\mathbb{R}$ satisfying the following axioms.*

1. $\|\underline{x}\|\geq 0$ *for all vectors* $\underline{x}\in\mathbb{R}^n$ *with* $\|\underline{x}\|=0$ *iff* $\underline{x}=\underline{0}$ *(separation axiom).*
2. $\|\alpha\underline{x}\|=|\alpha|\cdot\|\underline{x}\|,\ \forall\alpha\in\mathbb{R},\ \forall\underline{x}\in\mathbb{R}^n$*, (scaling axiom).*
3. $\left\|\underline{x}+\underline{y}\right\|\leq\|\underline{x}\|+\left\|\underline{y}\right\|,\ \forall\underline{x},\underline{y}\in\mathbb{R}^n$*, (triangle inequality).*

9.1.2 Matrix norms

Definition 4 *A function $\|\ \|:M_n(\mathbb{R})\to\mathbb{R}$ is a matrix norm provided the following conditions are satisfied:*

1. $\|A\|\geq 0,\ \forall A\in M_n(\mathbb{R})$*, with* $\|A\|=0$ *iff* $A=0$ *(separation axiom).*
2. $\|A+B\|\leq\|A\|+\|B\|,\ \forall A,B\in M_n(\mathbb{R})$*, the triangle inequality.*
3. $\|\alpha A\|=|\alpha|\cdot\|A\|,\ \forall A\in M_n(\mathbb{R}),\ \forall\alpha\in\mathbb{R}$*, (scaling axiom).*
4. $\|AB\|\leq\|A\|\cdot\|B\|,\ \forall A,B\in M_n(\mathbb{R})$*, the sub-multiplicative condition.*

9.1.3 Compatible norm conditions

Definition 5 *We say that a given matrix $\|\ \|_M$ is compatible with a vector norm $\|\ \|_{\underline{v}}$ on $\mathbb{R}^n$ if*

$$\|A\underline{x}\|_{\underline{v}}\leq\|A\|_M\|\underline{x}\|_{\underline{v}},\quad\forall A\in M_n(\mathbb{R}),\ \forall\underline{x}\in\mathbb{R}^n.$$

Proposition 1 *For any compatible matrix norm $\|\ \|$, $\rho(A)\leq\|A\|$.*

9.1.4 Subordinate norm condition

Definition 6 *Given any vector norm $\| \ \|_{\underline{v}}$, we can define a corresponding matrix norm $\| \ \|_M$ which is said to be subordinate to $\| \ \|_{\underline{v}}$ by*

$$\|A\|_M = \max_{\|x\|_{\underline{v}}=1} \|A\underline{x}\|_{\underline{v}}. \tag{22}$$

Proposition 2 *The spectral norm (2-norm) of a matrix is subordinate to the Euclidean norm on $\mathbb{R}^n$.*

Proposition 3 *1. In general, $\|I_n\| \geq 1$.*

2. If a matrix norm $\| \ \|_M$ is subordinate to a vector norm, $\|I_n\|_M = 1$.

3. The Frobenius norm is not subordinate to any vector norm, but it is compatible with $\| \ \|_2$.

10 Matrix perturbations and condition numbers

10.1 Lecture 18

10.1.1 Small norm perturbation of the identity

Theorem 3 *Let $\| \ \|$ be a matrix norm subordinate to a vector norm. Let $E \in M_n(\mathbb{R})$ with $\|E\| < 1$. Then $I_n - E$ is invertible and*

$$\frac{1}{1+\|E\|} \leq \left\|(I_n - E)^{-1}\right\| \leq \frac{1}{1-\|E\|}, \tag{23}$$

as well as

$$\frac{\|E\|}{1+\|E\|} \leq \left\|(I_n - E)^{-1} - I_n\right\| \leq \frac{\|E\|}{1-\|E\|}. \tag{24}$$

10.1.2 Perturbations of Linear Systems

Theorem 4 *Given a subordinate matrix norm $\| \ \|$, a non-singular matrix A, two vectors $\underline{b}, \delta\underline{b} \neq \underline{0}$ and a matrix δA such that $\|\delta A\| < 1/\left\|A^{-1}\right\|$, then $\underline{x}$ such that $A\underline{x} = \underline{b}$ and $(\underline{x}+\delta\underline{x})$ such that $(A+\delta A)(\underline{x}+\delta\underline{x}) = \underline{b}+\delta\underline{b}$ satisfy*

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \quad \leq \quad \frac{\|A\| \cdot \left\|A^{-1}\right\|}{1 - \|\delta A\| \cdot \|A^{-1}\|} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\underline{b}\|}{\|\underline{b}\|} \right). \tag{25}$$

10.2 Lecture 19

10.2.1 Condition numbers

Definition 7 *Given a matrix norm $\| \ \|$ and a non-singular matrix A, the condition number of A relative to the norm $\| \ \|$ is defined by*

$$\kappa(A) = \|A\| \cdot \left\|A^{-1}\right\|.$$

Proposition 4 *1. When $\| \ \|$ is a subordinate to some other vector norm, $\kappa(A) \geq 1$. In particular, $\kappa(I_n) = 1$.*

2. When O is orthogonal ($O^{-1} = O^T$), $\kappa_2(O) = 1$, so there are non-identity matrices with condition number κ_2 equal to 1.

Definition 8 *A matrix A is perfectly conditioned if $\kappa(A) = 1$, it is well-conditioned if $\kappa(a)$ is 'close to 1' and is ill-conditioned if $\kappa(A)$ is 'much larger than 1'.*

Proposition 5 *Let A be an invertible matrix and $\| \ \|$ a matrix norm.*

1. $\kappa(\alpha A) = \kappa(A)$ for any $\alpha \neq 0$.

2. $\kappa(A^{-1}) = \kappa(A)$.

3. Let $\| \ \|$ be compatible with a vector norm a vector norm, let $|\lambda_{\max}| = \rho(A)$ be the largest eigenvalue of A in modulus and $|\lambda_{\min}| > 0$ be the smallest eigenvalue of A in modulus. Then,

$$\kappa(A) \geq \frac{|\lambda_{\max}|}{|\lambda_{\min}|}.$$

4. If $\|\ \|$ is subordinate ($\|I_n\|$=1) and $\|A - I_n\| < 1$, then

$$\kappa(A) \leq \frac{\|A\|}{1 - \|A - I_n\|}. \tag{26}$$

10.2.2 Spectral condition number and singular values of matrices

Proposition 6 *Let A be a normal matrix ($AA^T = A^TA$), in particular symmetric, then*

$$\kappa_2(A) = \frac{\max\{|\lambda| : \lambda \in \sigma(A)\}}{\min\{|\lambda| : \lambda \in \sigma(A)\}}.$$

Proposition 7 *Let A be an invertible matrix.*

1. *There exist two orthonormal matrices O_1 and O_2 such that*

$$A = O_1 D O_2^T \tag{27}$$

where D is a diagonal matrix with positive entries.

2. *Moreover, let $\sigma_{\max}$, $\sigma_{\min}$ be the largest, respectively smallest, singular value of A:*

$$\kappa_2(A) = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

11 Interpolation, Lagrange polynomials

11.1 Lecture 21

11.1.1 Construction of Lagrange Polynomials

For the Lagrange polynomials, y_i and $f(x_i)$ can be used interchangeably.

Lemma 3 *Given $n+1$ points $(x_i, y_i) \in \mathbb{R}^2$, $0 \leq i \leq n$, with $x_i \neq x_j$ when $i \neq j$. The polynomials*

$$L_i(x) = \prod_{j=0,\neq i}^{n} \left(\frac{x - x_j}{x_i - x_j}\right), \quad 0 \leq i \leq n, \tag{28}$$

satisfy $L_i(x_i) = 1$ and $L_i(x_j) = 0$ for $j \neq i$.

Definition 9 *Given $n+1$ points $(x_i, y_i) \in \mathbb{R}^2$, $0 \leq i \leq n$, with $x_i \neq x_j$ when $i \neq j$. The Lagrange polynomial p is a polynomial of degree up to n equal to*

$$p(x) = \sum_{i=0}^{n} y_i L_i(x). \tag{29}$$

This is linear if $n = 1$ and quadratic if $n = 2$.

11.1.2 Vandermonde Method

Given $p(x) = ax^2 + bx + c$, with a, b and c to be determined from the conditions $p(x_i) = y_i$, $i = 0, 1, 2$, we can input this into a matrix know as the Vandermonde matrix and solve it for values a, b and c.

$$\begin{pmatrix} x_0^2 & x_0 & 1 \\ x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}. \tag{30}$$

This can be extended for larger polynomials trivially.

11.2 Lecture 22

11.2.1 Error estimates for Lagrange interpolation

Proposition 8 *Suppose $f \in C^{n+1}[a, b]$. Then, the error function e between f and it's Lagrange interpolant p at $\{x_i\}_{i=0}^n$ is given by*

$$e(x) = f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0) \cdots (x - x_n),$$

where $\xi = \xi(x)$ lies in (a, b).

12 Interpolation, splines

12.1 Lecture 23

12.1.1 Linear Splines

The rule of S_n on the interval $[x_i,\, x_{i+1}]$ is

$$S_n(x) = y_i + \frac{x - x_i}{x_{i+1} - x_i}(y_{i+1} - y_i) = \left(\frac{x - x_i}{x_{i+1} - x_i}\right) y_{i+1} + \left(\frac{x_{i+1} - x}{x_{i+1} - x_i}\right) y_i.$$

We can also re-cast S_n as

$$S_n(x) = \left(\frac{x_{i+1}y_i - x_k y_{i+1}}{x_{i+1} - x_i}\right) + x\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right), \quad x \in [x_i,\, x_{i+1}] \tag{31}$$

to identify the gradient easily.

12.2 Lecture 24

Assume we have splines on n intervals. The global regularity of the spline and the degree of the local polynomials satisfy the following relations. The notation $f \in C^k[a,\, b]$ means that f is k-times continuously differentiable. Recall that this means that the function and its first k derivatives are continuous at any $z \in]a,\, b]$, that is,

$$\lim_{x \to z^-} f^{(j)}(x) = \lim_{x \to z^+} f^{(j)}(x), \quad 0 \le j \le k.$$

polynomial	coefficients	C^0	C^1	C^2	equations	free coefficients
linear	$2n$	$2(n-1)+2$			$2n$	0
quadratic	$3n$	$2(n-1)+2$	$n-1$		$3n-1$	1
cubic	$4n$	$2(n-1)+2$	$n-1$	$n-1$	$4n-2$	2

Table 1: Number of variables (coefficients) and equations for splines.

12.2.1 Natural Cubic Splines

Given a function f defined on $[a, b]$ and a set of points $a = x_0 < x_1 < \ldots < x_n = b$, a function S is called a natural cubic spline if there exists n cubic polynomials S_i such that:

1. $S(x) = S_i(x)$ for x in $[x_i,\, x_{i+1}]$ and $0 \le i \le n-1$;
2. $S_i(x) = f(x_i)$ and $S_i(x_{i+1}) = f(x_{i+1})$, $0 \le i \le n-1$;
3. $S'_{i+1}(x_{i+1}) = S'_i(x_{i+1})$ for $0 \le i \le n-1$;
4. $S''_{i+1}(x_{i+1}) = S''_i(x_{i+1})$ for $0 \le i \le n-1$;
5. $S''(x_0) = S''(x_n) = 0$.

Denote $c_i = S''(x_i)$, $0 \le i \le n$. Then $c_n = c_0 = 0$. Then

$$S''_i = c_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + c_{i+1} \frac{x - x_i}{x_{i+1} - x_i}. \tag{32}$$

Integrate twice using boundary conditions to fix the integration constants.

13 Different schemes for the heat equation, consistency

13.1 Lecture 25

13.1.1 Consistency

A scheme is said to be consistent if it solves the PDE problem.

13.1.2 Stability

A numerical algorithm is said to be stable provided small errors in arithmetic remain bounded independently of the number of numerical steps used to reach the same value estimates.

13.1.3 Convergence

A scheme converges if it is both consistent and stable as the temporal and spatial time-steps are reduced.

13.1.4 Four Finite Difference Schemes for the Heat Equation

The performance of the four schemes is summarised as follows:

1. The Euler explicit scheme converges if δt is sufficiently small with respect to $(\delta x)^2$ but otherwise will blow-up.
2. The Richardson scheme seems to be unstable for all choices of the Courant number r, so it will not converge.
3. The Dufort-Frankel scheme converges but not necessarily to the correct solution as it is not consistent with the heat equation.
4. The Crank-Nicolson scheme seems to exhibit stable behaviour for all choices of the Courant number and converges to the correct solution as spatial resolution is improved.

14 Consistency

14.1 Lecture 26

Refer to lecture notes.

15 Convergence

15.1 Lecture 27

15.1.1 Local Error

denote u the exact solution and v the solution of the explicit Euler scheme to the equation

$$u_t = u_{xx}, \quad u(0,t) = u(1,t) = 0, \quad u(x,0) = u_0(x). \tag{33}$$

Proposition 9 *Let*

$$e_{i,j} = v_{i,j} - u(x_i, t_j)$$

be the local error at each grid point for the current set-up. Recall that $(\delta t) = r(\delta x)^2$*, where* r *is a constant. Then, the local error satisfies the following approximate iterative process:*

$$\begin{aligned} e_{i,j+1} &= (1-2r)e_{i,j} + re_{i+1,j} + re_{i-1,j} \\ &\quad + \underbrace{\frac{(\delta t)(\delta x)^2}{12}(1-6r)(u_{tt})_{i,j}}_{\textit{leading approximation term}} + O((\delta t)^3, (\delta t)(\delta x)^4). \end{aligned}$$

We can see that choosing $r = 1/6$ improves the accuracy by removing the leading error term. Proof in lecture notes.

15.2 Lecture 28

15.2.1 Error Along Time Layers for the Explicit Euler Scheme for the Heat Equation

Define E_j to be the spatial discretion error at a time t_j, that is,

$$E_j = \max\{|e_{i,j}| : 0 \le i \le M\} = \|e_{i,j}\|_\infty$$

and let $\|u_{tt}\|_\infty$ be the maximum value of u_{tt} across space and time on the rectangle $[0,\,1]\times[0,\,T]$:

$$\|u_{tt}\| = \{|u_{tt}(x,t) : (x,t) \in [a,b]\times[0,T]\}. \tag{34}$$

Our main result is as follows:

Proposition 10 *With the previous notation, when $r \le 1/2$, the error of the time-layer at time T is*

$$E_N \le T\left(\frac{(\delta x)^2|1-6r|\cdot\|u_{tt}\|_\infty}{12} + O((\delta t)^2,(\delta x)^4)\right).$$

15.2.2 Consequences for Convergence

Corollary 2 *Provided that $r \le 1/2$:*

1. *The maximum spatial discretisation error in the Euler scheme grows in direct proportion to T, but*
2. *For a fixed T, the error tends to 0 as δx tends to 0.*
3. *Remark that the explicit Euler scheme is more accurate than might otherwise have been anticipated when $r = 1/6$ because the first part of the error vanishes.*

16 Stability Fourier method

16.1 Lecture 29

Use lecture notes.

16.2 Lecture 30

Use lecture notes.

16.3 Lecture 31

Use lecture notes.

17 Stability matrix method

17.1 Lecture 32

17.1.1 Eigenvalues of Banded Matrices

A matrix is said to be banded (or Toeplitz) if the values along each diagonal are constant.

17.1.2 Circulant Matrices

An $m\times m$-banded matrix is circulant if $a_i = a_{i-m}$, $\forall 1 \le i \le m$. The rows of a circulant $n\times n$-matrix are obtained by shifting a sequence of n numbers by m places.

Proposition 11 *The tri-diagonal circulant matrix $(m+1)\times(m+1)$-matrix*

$$A = \begin{pmatrix} a_0 & a_1 & & & a_{-1} \\ a_{-1} & a_0 & a_1 & & \\ & a_{-1} & a_0 & a_1 & \\ & & \vdots & \vdots & a_1 \\ a_1 & & & a_{-1} & a_0 \end{pmatrix}$$

has eigenvalues

$$\lambda_p = a_{-1}e^{-\frac{2\pi p}{m+1}i} + a_0 + a_1 e^{\frac{2\pi p}{m+1}i}, \quad 1 \le p \le m+1, \tag{35}$$

corresponding to the eigenvector $\underline{v}_p$ having its j-th component equal to $e^{\frac{2\pi jp}{m+1}i}$, $1 \le j \le m+1$.

Corollary 3 *The tri-diagonal circulant $(m+1)\times(m+1)$-matrix*

$$A=\begin{pmatrix} a & b & & & b \\ b & a & b & & \\ & b & a & b & \\ & & \ddots & \ddots & b \\ b & & & b & a \end{pmatrix},$$

has eigenvalues

$$\lambda_p = a + 2b\cos\left(\frac{2\pi p}{m+1}\right), \quad 1 \le p \le m+1. \tag{36}$$

17.1.3 Tri-diagonal Matrices

Proposition 12 *Let $bc > 0$. The tri-diagonal $(m+1)\times(m+1)$-matrix*

$$\begin{pmatrix} a & b & 0 & 0 & \cdots & 0 \\ c & a & b & 0 & \cdots & 0 \\ 0 & c & a & b & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c & a & b \\ 0 & \cdots & 0 & 0 & c & a \end{pmatrix}$$

has eigenvalues

$$\lambda_k = a + 2\sqrt{bc}\cos\theta_k, \quad k = 1, \ldots, n,$$

where $\lambda_k = k\pi/(n+1)$, with corresponding eigenvector

$$\underline{v}_k = (\sqrt{c/b}\sin\theta_k,\ (c/b)\sin(2\theta_k), \ldots,\ (c/b)^{j/2}\sin(j\theta_k), \ldots,\ (c/b)^{n/2}\sin(n\theta_k)).$$

Corollary 4 *The tri-diagonal symmetric $(m+1)\times(m+1)$-matrix*

$$A=\begin{pmatrix} a & b & & & \\ b & a & b & & \\ & b & a & b & \\ & & \ddots & \ddots & b \\ & & & b & a \end{pmatrix},$$

has eigenvalues

$$\lambda_p = a + 2b\cos\left(\frac{p\pi}{m+1}\right), \quad 1 \le p \le m, \tag{37}$$

corresponding to the eigenvector having its j-th component equal to $\sin\left(\frac{jp\pi}{m+1}\right)$, $1 \le j \le m+1$.

17.2 Lecture 33

17.2.1 Statement of Gershgorin's Circle Theorem

Theorem 5 *Let A be an $n\times n$-matrix and the disks*

$$D_k = \{z \in \mathbb{C} : |z - A_{kk}| \le \sum_{j=1\neq k} |A_{kj}|\}.$$

Then, all eigenvalues of A are contained within the region $D = \bigcup_{k=1}^{n} D_k$. The disk D_k is called a Gershgorin's disk. Its boundary is a Gershgorin's circle, denoted by C_k.
In particular, if $m \le n$ of these disks form a subset of D which does not intersect the remaining $(n-m)$ disks, then precisely m eigenvalues are contained within such a subset.

17.2.2 Gershgorin's Circle Theorem and Transpose Matrix

Lemma 4 *Given an $n\times n$-real matrix A, then $\sigma(A) = \sigma(A^T)$.*

17.3 Lecture 34

17.3.1 Diagonally Dominant Matrices

Definition 10 *A matrix such that* $|A_{kk}| > \sum_{j=1\neq k}^{n} |A_{kj}|$ *for all* $1 \leq k \leq n$*, is said to be strictly diagonally dominant.*

Proposition 13 *Strictly dominant diagonal matrices are invertible.*

17.3.2 Lax Theorem

A boundary value problem is properly posed if:

1. The solution is unique when it exists (it has at most one solution);
2. The solution depends continuously on the initial data;
3. When a solution does not exist for some initial data, there exists arbitrarily close initial data for which a solution does exist.

Theorem 6 *Given a properly posed linear boundary value problem and a linear finite difference approximation scheme that is consistent with the BVP and stable, then the scheme is convergent.*