Scientific Computing (MA322)

Satyajit Pramanik[1]
Assistant Professor
Department of Mathematics
Indian Institute of Technology Guwahati

[1]satyajitp@iitg.ac.in

Contents

1 Root findings **5**
1.1 Problems . . . 5

2 Polynomial interpolations **7**
2.1 Problems . . . 7

3 Numerical integrations or quadratures **9**
3.1 Problems . . . 14

4 Numerical differentiation and initial value problems for ODEs **17**
4.1 Introduction . . . 17
4.2 Consistency of Euler Method . . . 17
4.3 Modified Euler Method . . . 19
4.4 General One Step Method . . . 20
4.5 Runge-Kutta method . . . 22
4.5.1 Second-order RK method . . . 23
4.5.2 Third-order RK method . . . 24
4.5.3 Fourth-order RK method . . . 24
4.5.4 Some important points . . . 24
4.6 Taylor's series method . . . 25
4.7 Stiff differential equations . . . 25
4.7.1 Stability of a numerical method . . . 26
4.8 Linear Multistep Methods . . . 26
4.8.1 Milne's Predictor Method . . . 27
4.8.2 Adams methods . . . 28

5 Matrix Algebra **31**
5.1 Preliminaries . . . 31
5.2 Solution of equations by iterative methods . . . 33

6 Boundary value problems of ODEs **35**
6.1 Introduction . . . 35

7 Finite difference methods of PDEs **37**
7.1 Preliminaries . . . 37
7.2 Consistency, Stability and Convergence . . . 38

7.2.1 Consistency 38
7.2.2 stability 38
7.2.3 Convergence 39
7.3 Vector and matrix norms 39
7.3.1 Vector norms 39
7.3.2 Matrix norms 40
7.3.3 Subordinate matrix norm 40
7.4 A necessary and sufficient condition for stability 41
7.5 Glabal rounding error 44
7.6 Finite difference approximation of parabolic equations in two space-dimensions 45
7.6.1 The θ-scheme 45
7.6.2 The alternating direction implicit (ADI) method LeVeque [2007] . . . 45
7.7 Hyperbolic PDEs 47
7.8 Numerical solution of first order hyperbolic equation 48
7.9 Upwind scheme 50
7.9.1 Geometric and physical interpretations of CFL condition 50
7.10 Lax-Wendroff methods 52
7.10.1 Lax-Wendroff explicit scheme 52
7.10.2 Lax-Wendroff implicit scheme 52

Chapter 1

Root findings

1.1 Problems

1. **(Steffensen's method)** Consider the iteration formula

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

   Where

   $$g(x) = \left[f(x + f(x)) - f(x)\right]/f(x)$$

   Show that this is quadratically convergent, under suitable hypotheses.

2. What is the purpose of the following iteration formula?

   $$x_{n+1} = 2x_n - x_n^2 y$$

   Identify it as the Newton iteration for a certain function.

3. Define $x_0 = 0$ and $x_{n+1} = x_n - [(\tan x_n - 1)/\sec^2 x_n]$. What is the $\lim_{n\to\infty} x_n$ in this example? Relate this to Newton's method.

4. Find the conditions on α to ensure that the iteration

   $$x_{n+1} = x_n - \alpha f(x_n)$$

   will converge linearly to a zero of f if started near the zero.

5. Perform four iterations of Newton's method for the polynomial

   $$p(x) = 4x^3 - 2x^2 + 3$$

   starting with $x_0 = -1$. Use a hand calculator.

6. Find the order of convergence of these sequences.

   (a) $x_n = (1/n)^{\frac{1}{2}}$

   (b) $x_n = \sqrt[n]{n}$

(c) $x_n = (1 + 1/n)^{\frac{1}{2}}$

(d) $x_n = \tan^{-1} x_n$

7. To find a zero of the function f, we can look for a fixed point of the function $F(x) = x - f(x)/f'(x)$. To find a fixed point of F, we can solve $F(x) - x = 0$ by Newton's method. When this is done, what is the formula for generating the sequence x_n?

8. Prove that the function F defined by $F(x) = 4x(1 - x)$ maps the interval $[0, 1]$ into itself and is not a contraction. Prove that it has a fixed point. Why does this not contradict the Contractive Mapping Theorem?

9. If the method of functional iteration is used on $f(x) = \frac{1}{2}(1 + x^2)^{-1}$ starting at $x_0 = 7$, will the resulting sequence converge? If so, what is the limit? Establish your answers rigorously.

10. A function F is called an **iterated contraction** if

$$|F(F(x)) - F(x)| \leq \lambda |F(x) - x| \qquad (\lambda < 1)$$

Show that every contraction is an iterated contraction. Show that an iterated contraction need not be a contraction nor continuous.

11. Consider a function of the form $F(x) = x - f(x)f'(x)$, where $f(r) = 0$ and $f'(r) \neq 0$. Find the precise conditions on the function f so that the method of functional iteration will converge at least **cubically** to r if started near r.

12. Show that the following method has third-order convergence for computing $\sqrt{R}$:

$$x_{n+1} = \frac{x_n(x_n^2 + 3R)}{3x_n^2 + R}.$$

13. Let $\frac{1}{2} \leq q \leq 1$, and define $F(x) = 2x - qx^2$. On what interval can it be guaranteed that the method of iteration using F will converge to a fixed point?

14. Write down two different fixed-point procedures for finding a zero of the function $f(x) = 2x^2 + 6e^{-x} - 4$.

15. On which of these intervals $[\frac{1}{2}, \infty], [\frac{1}{8}, 1], [\frac{1}{4}, 2], [0, 1], [\frac{1}{5}, \frac{3}{2}]$ is the function $f(x) = \sqrt{x}$ contractive?

Chapter 2

Polynomial interpolations

2.1 Problems

1. Write the Lagrange and Newton interpolating polynomials for these data:

x	2	0	3
$f(x)$	11	7	28

2. Find the Lagrange and Newton forms of the interpolating polynomial for these data: Write both polynomials in the form $a + bx + cx^2$ to verify that they are identical as

x	-2	0	1
$f(x)$	0	1	-1

functions.

x	$-\sqrt{\frac{3}{5}}$	0	$\sqrt{\frac{3}{5}}$
$f(x)$	$f(-\sqrt{\frac{3}{5}})$	$f(0)$	$f(\sqrt{\frac{3}{5}})$

3. What are the Newton interpolation polynomial and the Lagrange interpolation polynomial for the above data?

4. The formula for the leading coefficient in T_n is 2^{n-1}. What is the formula for the coefficient of x^{n-2}? What about x^{n-1}?

5. The equation $x - 9^{-x} = 0$ has a solution in $[0, 1]$. Find the interpolation polynomial on $x_0 = 0,\ x_1 = 0.5,\ x_2 = 1$ for the function on the left side of the equation. By setting the interpolation polynomial equal to 0 and solving the equation, find an approximate solution to the equation.

6. The functions $1/(1+x^2)$ and e^{-x^2} have a similar appearance. Do they behave similarly in the interpolation process for equally spaced nodes?

7. If we interpolate the function $f(x) = e^{x-1}$ with a polynomial p of degree 12 using 13 nodes in $[-1, 1]$, what is a good upper bound for $|f(x) - p(x)|$ on $[-1, 1]$?

8. Find the natural cubic spline function whose knots are —1, 0, and 1 and that takes the values $S(-1) = 13,\ S(0) = 7$, and $S(1) = 9$.

9. Can a and b be defined so that the function

$$S(x) = \begin{cases} (x-2)^3 + a(x-1)^2 & x \in (-\infty, 2], \\ (x-2)^3 - (x-3)^2 & x \in [2, 3], \\ (x-3)^3 + b(x-2)^2 & x \in [3, \infty) \end{cases}$$

is a natural cubic spline? Why or why not?

10. What value of (a, b, c, d) makes this a cubic spline?

$$f(x) = \begin{cases} x^3 & x \in [-1, 0], \\ a + bx + cx^2 + dx^3 & x \in [0, 1] \end{cases}$$

11. Determine the value of (a, b, c) that makes the function

$$f(x) = \begin{cases} x^3 & x \in [0, 1], \\ \frac{1}{2}(x-1)^3 + a(x-1)^2 + b(x-1) + c & x \in [1, 3] \end{cases}$$

a cubic spline. Is it a natural cubic spline?

12. Which properties of a natural cubic spline does the following function possess, and which properties does it not possess?

$$f(x) = \begin{cases} (x+1) + (x+1)^3 & x \in [-1, 0], \\ 4 + (x-1) + (x-1)^3 & x \in (0, 1] \end{cases}$$

13. Determine the values of a, b, and c so that this is a cubic spline having knots 0, 1, and 2:

$$f(x) = \begin{cases} 3 + x - 9x^2 & x \in [0, 1], \\ a + b(x-1) + c(x-1)^2 + d(x-1)^3 & x \in [1, 2] \end{cases}$$

Next, determine d so that $\int_0^2 [f''(x)]^2 dx$ is a minimum. Finally, find the value of d that makes $f''(2) = 0$ and explain why this value is different from the one previously determined.

Chapter 3

Numerical integrations or quadratures

Definition 3.0.1 (Inner product)**.** *For $f, g \in C[a,b]$, we consider the inner product*

$$\langle f,\, g\rangle := \int_a^b f(x)g(x)dx.$$

Definition 3.0.2 (Norm)**.**

$$\|f\|_2 = \sqrt{\langle f,\, f\rangle} = \sqrt{\int_a^b |f(x)|^2 dx}$$

defines a norm on $C[a,b]$.

Definition 3.0.3 (Orthogonality)**.** *Two polynomials p and q are said be orthogonal on $[a,b]$ if*

$$\langle p,\, q\rangle = \int_a^b p(x)q(x)dx = 0.$$

Definition 3.0.4 (Orthonomality)**.** *Additionally, if $\langle p,\, p\rangle = \langle q,\, q\rangle = 1$, then p and q are said to be orthonormal.*

Example 3.0.1. *The polynomials $1,\ x,\ x^2 - \dfrac{1}{3}$ are mutually orthogonal on $[-1,1]$.*

Definition 3.0.5 (Weighted inner product)**.** *If $w(x)$ is a positive function in $C[a,b]$ then*

$$\langle f,\, g\rangle_w := \int_a^b f(x)g(x)\mu(x)dx$$

is called a weighted inner product.

Definition 3.0.6. *A polynomial $p(x)$ is said to be orthogonal to $\mathcal{P}_n$ if $\langle p,\, q\rangle = 0$ for all $q \in \mathcal{P}_n$. In such a case, we write $p \perp \mathcal{P}_n$.*

Suppose that $I_n(f) = w_1 f(x_1) + \cdots + w_n f(x_n)$ is interpolatory. Then

$$I_n(f) = \int_a^b p_{n-1}(x)dx \ \text{ and } \ w_j = \int_a^b \ell_j(x)dx,\ j = 1:n,$$

where $p_{n-1}(x)$ is the interpolating polynomial of degree n that interpolates the data $(x_1, f(x_1))$ $, \ldots, (x_n, f(x_n))$. We now use orthogonal polynomials to determine the nodes. Recall that

$$f(x) = p_n(x) + \frac{f^{(n)}(\theta_x)}{n!}(x - x_1)\cdots(x - x_n).$$

Set $w(x) := (x - x_1)\cdots(x - x_n)$. Then

$$\int_a^b f(x)dx - \int_a^b p_{n-1}(x)dx = \int_a^b \frac{f^{(n)}(\theta_x)}{n!}w(x)dx = 0$$

when $f \in \mathcal{P}_{2n-1} \iff w(x) \perp \mathcal{P}_{n-1}$.

Consider the Gaussian quadrature

$$I_n(f) = w_1 f(x_1) + \cdots + w_n f(x_n) \approx \int_a^b f(x)dx.$$

Then the following hold:

- All the nodes x_j are real, distinct, and contained in (a, b).
- All the weights w_j are positive. Indeed, for $j = 1 : n$, we have

$$0 < \int_a^b \ell_j(x)^2 dx = \sum_{k=1}^n w_k \ell_j(x_k) = w_j.$$

- Let $f \in C[a, b]$. Let $E_n(f) := \|f - p_{2n-1}\|_\infty = \min\{\|f - p\|_\infty : p \in \mathcal{P}_{2n-1}\}$. Then

$$\left|\int_a^b f(x)dx - I_n(f)\right| \le 2(b - a)E_n(f) \longrightarrow 0 \text{ as } n \to \infty.$$

Proof.

$$\begin{aligned}
\left|\int_a^b f(x)\mathrm{d}x - I_n(f)\right| &\le \left|\int_a^b f(x)\mathrm{d}x - I_n(p_{2n-1})\right| + |I_n(p_{2n-1}) - I_n(f)| \\
&\le \left|\int_a^b (f(x) - p_{2n-1}(x)\mathrm{d}x\right| + \sum_{j=0}^n w_j |f(x_j) - p_{2n-1}(x_j)| \\
&\le (b - a)E_n(f) + E_n(f)\sum_{j=0}^n w_j = 2(b - a)E_n(f).
\end{aligned}$$

By Weierstrass theorem, $E_n(f) \to 0$ as $n \to \infty$. □

Consider $\mathcal{P}_n = \text{span}(1, x, \ldots, x^n)$. Let $\langle p, q\rangle$ be an inner product on $\mathcal{P}$. For example, $\langle p, q\rangle := \int_a^b p(x)q(x)\mathrm{d}x$ is an inner product on $\mathcal{P}$. Similarly,

$$\langle p, q\rangle := \int_0^\infty p(x)q(x)e^{-x}\mathrm{d}x \text{and} \langle p, q\rangle := \int_{-1}^1 p(x)q(x)(1 - x^2)^{-1/2}\mathrm{d}x$$

are inner products on $\mathcal{P}_n$.

Given an inner product on $\mathcal{P}_n$, there exist orthogonal polynomials $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$ of degree at most n such that

$$\text{span}(1, x, \ldots, x^j) = \text{span}(\phi_0(x), \phi_1(x), \ldots, \phi_j(x)), \quad j = 0 : n.$$

Gram-Schmidt process: Define $\phi_0, \ldots, \phi_n$ by $\phi_0(x) := 1$,

$$\begin{aligned}
\phi_1(x) &:= x - \frac{\langle x,\, \phi_0\rangle}{\langle \phi_0,\, \phi_0\rangle}\phi_0(x), \\
\phi_j(x) &:= x^j - \sum_{j=0}^{j-1} \frac{\langle x^j,\, \phi_j\rangle}{\langle \phi_j,\, \phi_j\rangle}\phi_j(x), \quad j = 2 : n
\end{aligned}$$

Consider the inner product $\langle p,\, q\rangle := \displaystyle\int_{-1}^{1} p(x)q(x)\mathrm{d}x$. Then

$$\begin{aligned}
\phi_0(x) &:= 1, \quad \phi_1(x) := x - \frac{\langle x,\, 1\rangle}{\langle 1,\, 1\rangle}\cdot 1 = x - \frac{\int_{-1}^{1} x\mathrm{d}x}{\int_{-1}^{1} 1\cdot \mathrm{d}x}\cdot 1 = x, \\
\phi_2(x) &= x^2 - \frac{\langle x^2,\, 1\rangle}{\langle 1,\, 1\rangle}\cdot 1 - \frac{\langle x^2,\, x\rangle}{\langle x,\, x\rangle}\cdot x = x^2 - \frac{\int_{-1}^{1} x^2\mathrm{d}x}{\int_{-1}^{1}\mathrm{d}x} - \frac{\int_{-1}^{1} x^3\mathrm{d}x}{\int_{-1}^{1} x^2\mathrm{d}x}\cdot x \\
&= x^2 - \frac{1}{3} - 0\cdot x = x^2 - \frac{1}{3}.
\end{aligned}$$

The polynomial $\phi_j(x)$ is called Legendre polynomial. The roots of $\phi_2(x) = x^2 - \dfrac{1}{3}$ are $\pm 1/\sqrt{3}$. This yields the two-point Gaussian quadrature rule

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx w_1 f\left(-\frac{1}{\sqrt{3}}\right) + w_2 f\left(\frac{1}{\sqrt{3}}\right).$$

Exactness at 1 and x yield $w_1 = w_2 = 1$.

Consider the inner product $\langle p,\, q\rangle := \displaystyle\int_{-1}^{1} p(x)q(x)(1-x^2)^{-1/2}\mathrm{d}x$. Then $\phi_0(x) = 1$ and

$$\begin{aligned}
\phi_1(x) &= x - \frac{\langle x,\, 1\rangle}{\langle 1,\, 1\rangle}\cdot 1 = x - \frac{\int_{-1}^{1} x(1-x^2)^{-1/2}dx}{\int_{-1}^{1}(1-x^2)^{-1/2}dx}\cdot 1 = x, \\
\phi_2(x) &= x^2 - \frac{\langle x^2,\, 1\rangle}{\langle 1,\, 1\rangle}\cdot 1 - \frac{\langle x^2,\, x\rangle}{\langle x,\, x\rangle}\cdot x \\
&= x^2 - \frac{\int_{-1}^{1} x^2(1-x^2)^{-1/2}\mathrm{d}x}{\int_{-1}^{1}(1-x^2)^{-1/2}\mathrm{d}x} - \frac{\int_{-1}^{1} x^3(1-x^2)^{-1/2}\mathrm{d}x}{\int_{-1}^{1} x^2(1-x^2)^{-1/2}\mathrm{d}x} \\
&= x^2 - \frac{\pi/2}{\pi} - 0\cdot x = x^2 - \frac{1}{2}.
\end{aligned}$$

Note that $\phi_2(x) = x^2 - 1/2$ is a Chebyshev polynomial with roots $\pm 1/\sqrt{2}$. This yields two-point Gaussian quadrature rule

$$\int_{-1}^{1} f(x)(1-x^2)^{-1/2}\mathrm{d}x \approx w_1 f\left(-\frac{1}{\sqrt{2}}\right) + w_2 f\left(\frac{1}{\sqrt{2}}\right).$$

Exactness at 1 and x yield $w_1 = w_2 = \pi/2$.

Recall that Chebyshev polynomials satisfy the recurrence relation

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ n = 1, 2, \ldots.$$

Definition 3.0.7 (Monic polynomial)**.** *A polynomial $p(x)$ is said to be monic if the coefficient of the highest power of x is* 1.

Theorem 3.0.1. *Let ϕ_{n+1} be monic polynomial of degree $n+1$ such that $\phi_{n+1} \perp \mathcal{P}_n$. Then for $n = 0, 1, \ldots,$ we have*

$$\phi_{n+1}(x) = (x - \alpha_n)\phi_n(x) - \beta_n\phi_{n-1}(x),$$

where $\phi_0(x) = 1, \phi_{-1}(x) = 0$ and

$$\alpha_n = \frac{\langle x\phi_n,\, \phi_n\rangle}{\langle \phi_n,\, \phi_n\rangle},\ \beta_n = \frac{\langle \phi_n,\, x\phi_{n-1}\rangle}{\langle \phi_{n-1},\, \phi_{n-1}\rangle} = \frac{\langle \phi_n,\, \phi_n\rangle}{\langle \phi_{n-1},\, \phi_{n-1}\rangle} > 0.$$

Proof. Note that $\phi_{n+1} - x\phi_n \in \mathcal{P}_n$, hence

$$\begin{aligned}\phi_{n+1} - x\phi_n &= \sum_{j=0}^{n} \frac{\langle \phi_{n+1} - x\phi_n,\, \phi_j\rangle}{\langle \phi_j,\, \phi_j\rangle}\phi_j = -\sum_{j=0}^{n} \frac{\langle x\phi_n,\, \phi_j\rangle}{\langle \phi_j,\, \phi_j\rangle}\phi_j \\ &= -\frac{\langle x\phi_n,\, \phi_n\rangle}{\langle \phi_n,\, \phi_n\rangle}\phi_n - \frac{\langle x\phi_n,\, \phi_{n-1}\rangle}{\langle \phi_{n-1},\, \phi_{n-1}\rangle}\phi_{n-1},\end{aligned}$$

since $\displaystyle\sum_{j=0}^{n-2} \frac{\langle x\phi_n,\, \phi_j\rangle}{\langle \phi_j,\, \phi_j\rangle}\phi_j = 0$ as $x\phi_j \in \mathcal{P}_{n-1}$ for $j = 0, \ldots, n-2$.

This shows that $\phi_{n+1}(x) = (x - \alpha_n)\phi_n(x) - \beta_n\phi_{n-1}(x)$, where

$$\alpha_n = \frac{\langle x\phi_n,\, \phi_n\rangle}{\langle \phi_n,\, \phi_n\rangle},\ \beta_n = \frac{\langle \phi_n,\, x\phi_{n-1}\rangle}{\langle \phi_{n-1},\, \phi_{n-1}\rangle}.$$

Since $x\phi_{n-1} = \phi_n +$ lower order terms, we have $\langle \phi_n,\, x\phi_{n-1}\rangle = \langle \phi_n,\, \phi_n\rangle$. Hence

$$\beta_n = \frac{\langle \phi_n,\, x\phi_{n-1}\rangle}{\langle \phi_{n-1},\, \phi_{n-1}\rangle} = \frac{\langle \phi_n,\, \phi_n\rangle}{\langle \phi_{n-1},\, \phi_{n-1}\rangle} > 0.$$

□

Theorem 3.0.2. *Let $\phi_0(x), \ldots, \phi_{n+1}(x)$ be orthogonal polynomials such that*

$$\phi_{n+1}(x) = (x - \alpha_n)\phi_n(x) - \beta_n \phi_{n-1}(x).$$

Define the Jacobi matrix

$$A = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & \\ & \sqrt{\beta_2} & \ddots & \ddots & \\ & & \ddots & \ddots & \sqrt{\beta_{n-1}} \\ & & & \sqrt{\beta_{n-1}} & \alpha_n \end{bmatrix}.$$

Then $\phi_{n+1}(x_j) = 0 \iff \det(A - x_j I) = 0$. Let $\mathbf{v_0}, \ldots, \mathbf{v_n}$ be orthonormal eigenvectors of A. Then $w_j = \langle 1, 1\rangle(\mathbf{e}_0^T \mathbf{v}_j)^2$, $j = 0 : n$, are the required weights for the Gaussian quadrature rule

$$\int_a^b f(x)\mu(x)dx \approx w_0 f(x_0) + \cdots + w_n f(x_n).$$

Proof. Define $\psi_j(x) := \phi_j(x)/\|\phi_j\|$. Dividing the recurrence relation by $\|\phi_n\|$ and noting that $\sqrt{\beta_n} = \|\phi_n\|//\|\phi_{n-1}\|$, we have

$$\sqrt{\beta_{n+1}}\psi_{n+1}(x) = (x - \alpha_n)\psi_n(x) - \sqrt{\beta_n}\psi_{n-1}$$

which gives

$$x\psi_n(x) = \sqrt{\beta_{n+1}}\psi_{n+1}(x) + \alpha_n\psi_n + \sqrt{\beta_n}\psi_{n-1}.$$

Set $\Psi(x) := [\psi_0(x), \ldots, \psi_n(x)]^\top$. Then

$$x\Psi(x) = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & \\ & \sqrt{\beta_2} & \ddots & \ddots & \\ & & \ddots & \ddots & \sqrt{\beta_{n-1}} \\ & & & \sqrt{\beta_{n-1}} & \alpha_n \end{bmatrix} \begin{bmatrix} \psi_0(x) \\ \psi_1(x) \\ \vdots \\ \vdots \\ \psi_n(x) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \sqrt{\beta_{n+1}}\psi_{n+1}(x) \end{bmatrix}.$$

This shows that

$$\phi_{n+1}(x_j) = 0 \iff \psi_{n+1}(x_j) = 0 \iff \det(A - x_j I) = 0.$$

□

3.1 Problems

1. Verify that the following formula is exact for polynomials of degree ≤ 4:

$$\int_0^1 f(x)\,dx \approx \frac{1}{90}\left[7f(0) + 32f\left(\frac{1}{4}\right) + 12f\left(\frac{1}{2}\right) + 32f\left(\frac{3}{4}\right) + 7f(1)\right].$$

2. (Continuation) From the formula in the preceding problem, obtain a formula for $\int_a^b f(x)\,dx$ that is exact for all polynomials of degree 4.

3. (Continuation) Calculate $\ln 2$ approximately by applying the formula in the preceding problem to

$$\int_0^1 \frac{dt}{t+1}$$

Compare your answer to the correct value and compute the error.

4. Calculate $\int_0^1 e^{x^2}\,dx$ to eight-decimal-place accuracy by use of the series in the text.

5. Use the Lagrange interpolation polynomial to derive the formula of the form

$$\int_0^1 f(x)\,dx \approx Af\left(\frac{1}{3}\right) + Bf\left(\frac{2}{3}\right)$$

Transform this formula to one for integration over $[a, b]$.

6. Determine values for A, B, and C that make the formula

$$\int_0^2 xf(x)\,dx \approx Af(0) + Bf(1) + Cf(2)$$

exact for all polynomials of degree as high as possible. What is the maximum degree?

7. Derive the Newton-Cotes formula for

$$\int_0^1 f(x)\,dx$$

based on the Lagrange interpolation polynomial at the nodes -2, -1, and 0. Apply this result to evaluate the integral when $f(x) = \sin \pi x$.

8. Using the polynomial of lowest order that interpolates $f(x)$ at x_1 and x_2, derive a numerical integration formula for

$$\int_{x_0}^{x_3} f(x)\,dx$$

Do not assume uniform spacing. Here $x_0 < x_1 < x_2 < x_3$.

9. There are two Newton-Cotes formulas for $n = 2$ and $[a, b] = [0, 1]$; namely,

$$\int_0^1 f(x)\,dx \approx af(0) + bf\left(\frac{1}{2}\right) + cf(1)$$

$$\int_0^1 f(x)\,dx \approx \alpha f\left(\frac{1}{4}\right) + \beta f\left(\frac{1}{2}\right) + \gamma f\left(\frac{3}{4}\right)$$

Which is better?

10. Derive the composite rule for $\int_a^b f(x)\,dx$ based on the midpoint rule

$$\int_{-1}^1 f(x)\,dx \approx 2f(0)$$

Give formulas for unequal spacing and equal spacing of nodes.

11. Is there a formula of the form

$$\int_0^1 f(x)\,dx \approx \alpha\,[f(x_0) + f(x_1)]$$

that correctly integrates all quadratic polynomials?

12. Determine the minimum number of subintervals needed to approximate

$$\int_1^2 (x + e^{-x^2})\,dx$$

to an accuracy of at least $\frac{1}{2} \times 10^{-7}$, using the trapezoid rule.

13. For what value of α is this formula exact on Π_3 ?

$$\int_0^2 f(x)\,dx \approx f(\alpha) + f(2 - \alpha)$$

14. (a) Determine appropriate values of A_i and x_i so that the quadrature formula

$$\int_{-1}^1 x^2 f(x)\,dx \approx \sum_{i=0}^n A_i f(x_i)$$

will be correct when f is any polynomial of degree 3. Use $n = 1$.

(b) Repeat when f is any polynomial of degree 5, using $n = 2$.

15. (a) Find a formula of the form

$$\int_0^1 xf(x)\,dx \approx \sum_{i=0}^n A_i f(x_i)$$

with $n = 1$, that is exact for all polynomials of degree 3.

(b) Repeat with $n = 2$, making the formula exact on Π_5 .

16. Prove that every quadrature formula of the type

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i)$$

is exact on some infinite-dimensional subspace of $C[a, b]$.

17. Which of these polynomials is orthogonal to Π_2 on the interval $[0, 1]$ with weight function $w(x) = 1$? $1 + x,\ x - \frac{1}{2},\ x^2 - 3x + 1,\ 35x^4 - 60x^2 + 32x - 3,\ x^3 - 3x^2 + x - 1$.

18. Consider a numerical integration rule of the form

$$\int_{-1}^{1} f(x)\,dx \approx Af\left(-\sqrt{\frac{3}{5}}\right) + Bf(0) + Cf\left(\sqrt{\frac{3}{5}}\right)$$

(a) What is the linear system that must be solved in the method of undetermined coefficients for finding $A,\ B$, and C? Solve for $A,\ B$, and C.

(b) What three integrals must be evaluated to determine $A,\ B$, and C in a Newton-Cotes formula? Solve for $A,\ B$, and C.

19. If the formula

$$\int_1^2 (x^4 - 1)f(x)\,dx = Af(x_0) + Bf(x_1) + Cf(x_2)$$

is correct for all f that are polynomials of degree ≤ 5, then $x_0,\ x_1$, and x_2 must be roots of a polynomial q having what properties?

20. Find a nonzero polynomial that is orthogonal to Π_2 on the interval $[-1, 1]$ with respect to the weight function $1 + x^2$.

21. Determine the coefficients $A_0,\ A_1$, and A_2 that make the formula

$$\int_0^2 f(x)\,dx \approx A_0 f(0) + A_1 f(1) + A_2 f(2)$$

exact for all polynomials of degree 3.

22. Given $f(0),\ f'(-1)$, and $f''(1)$, compute an approximation to $\int_{-1}^{1} x^2 f(x)\,dx$ by the method of undetermined coefficients. Your formula should give exact results for all f in Π_2.

23. Using the method of undetermined coefficients, find A, B, and C in the following rule, which should give exact results for polynomials of degree 2:

$$\int_{-3h}^{h} f(x)\,dx \approx h\left[Af(0) + Bf(-h) + Cf(-2h)\right]$$

Chapter 4

Numerical differentiation and initial value problems for ODEs

4.1 Introduction

Let $f : D \subset \mathbb{R}^2 \to \mathbb{R}$ is continuous. We are interested in solving an initial value problem (IVP) of the form

$$y' = f(x, y), \tag{4.1}$$
$$y(x_0) = y_0, \tag{4.2}$$

where (x_0, y_0) is a point in D.

4.2 Consistency of Euler Method

Consistency of a numerical method is related to the truncation error. If the truncation error tends to zero when the mesh parameter tends to zero, then the numerical scheme is called consistent. Recall IVP

$$y'(x) = f(x, y(x)), \;\; y(x_0) = y_0. \tag{4.3}$$

In operator notation, above equation can be written as

$$(T(y))(x) = f(x, y(x)), \;\; T(y) = y'. \tag{4.4}$$

In particular at $x = x_n$, we obtain

$$y'(x_n) = f(x_n, y(x_n)), \;\; y(x_0) = y_0. \tag{4.5}$$

Now, expanding $y(x_{n+1})$ about x_n, we have

$$\begin{aligned} y(x_{n+1}) := y(x_n + h) &= y(x_n) + hy'(x_n) + \frac{h^2}{2} y''(\xi_n), \;\; x_n \le \xi_n \le x_{n+1}, \\ &= y(x_n) + hy'(x_n) + \mathrm{O}(h^2). \end{aligned} \tag{4.6}$$

Above equation (4.6) yields

$$y'(x_n) = \frac{y(x_{n+1}) - y(x_n)}{h} + \mathrm{O}(h) \approx \frac{y(x_{n+1}) - y(x_n)}{h}. \tag{4.7}$$

Therefore

$$y'(x_n) \approx \frac{y(x_{n+1}) - y(x_n)}{h}. \tag{4.8}$$

Hence, at $x = x_n$, the equation in (4.3) can be approximated by following equation

$$\frac{y(x_{n+1}) - y(x_n)}{h} = f(x_n, y(x_n)). \tag{4.9}$$

It is important to note that above approximation leads to Euler scheme. In operator notation, above equation (4.9) can be written as

$$(T_h(y))(x_n) = f(x_n, y(x_n)), \quad (T_h(y))(x_n) = \frac{y(x_{n+1}) - y(x_n)}{h}. \tag{4.10}$$

Finally, at each grid point we have following error

$$((T - T_h)(y))(x_n) = (T(y))(x_n) - (T_h(y))(x_n) = y'(x_n) - \frac{y(x_{n+1}) - y(x_n)}{h}. \tag{4.11}$$

Above error is known as truncation error. Therefore, using (4.8), the truncation error for the Euler method is given by

$$((T - T_h)(y))(x_n) = \mathrm{O}(h), \tag{4.12}$$

which tends to 0 as $h \to 0$. Hence, Euler method is *consistent*.
Remark:

1. Above technique will be frequently used while discussing truncation errors for the approximations differential equations (ODE/ PDE). Thus, consistency means the convergence of the scheme that is $T - T_h$, which can be measured with respect to

   (a) the operator norm: $\|T - T_h\|$,

   (b) the norm in suitable function space (eg. $C[a, b]$, $C^1[a, b]$ etc.): $\|(T - T_h)(y)\|$,

   (c) the norm in $\mathbb{R}$: $\|((T - T_h)(y))(x)\| = |((T - T_h)(y))(x)|$,

2. Another issue which is related with the round of error. During computation, the input data is always influenced by round of error. So, a numerical scheme is called stable if the round of error does not grow exponentially. In laymen sense, you should have controlled over round of error. This will be discussed in a separate lecture.

4.3 Modified Euler Method

Integrate $y' = f(x, y(x))$ over $[x_n, x_{n+1}]$ to have

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(t, y(t))dt = y(x_n) + \int_{x_n}^{x_{n+1}} g(t)dt. \tag{4.13}$$

Then consider the Trapezoidal rule

$$\int_{x_n}^{x_{n+1}} g(t)dt \approx \frac{x_{n+1} - x_n}{2}\big(g(x_n) + g(x_{n+1})\big). \tag{4.14}$$

Applying (4.14) in the equation (4.13), we obtain

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + \int_{x_n}^{x_{n+1}} f(t, y(t))dt = y(x_n) + \int_{x_n}^{x_{n+1}} g(t)dt \\ &\approx y(x_n) + \frac{x_{n+1} - x_n}{2}\big(g(x_n) + g(x_{n+1})\big) \\ &= y(x_n) + \frac{h}{2}\big(f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))\big). \end{aligned} \tag{4.15}$$

Again from the Euler method, the quantity $y(x_{n+1})$ in the right hand side can be approximated as

$$y(x_{n+1}) \approx y_n + hf(x_n, y_n),$$

which together with (4.15), we arrive at

$$y(x_{n+1}) \approx y_n + \frac{h}{2}\big(f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))\big) = y_{n+1}. \tag{4.16}$$

Above approximation is known as Modified Euler method.

Remark: Modified Euler method can be interpreted as follows:

- First, we predict the approximation of $y(x_{n+1})$ by Euler method, then the resulting approximation is corrected by modified Euler method given by (4.16). Such method is known as **predictor and corrector method**.

- It is natural to expect that modified Euler method should perform better than classical Euler method in terms of order of convergence. This has been proved in Section 4.4.

- From (4.16), it is clear that evaluation of y_{n+1} depends on the y_n, so modified Euler method is a single step method. Some text book refer it as improved Euler method. But, we will not distinguish between modified and improved.

4.4 General One Step Method

Now, we turn our discussion to the convergence of general one step method. A general explicit one-step method may be written as in the form

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h). \tag{4.17}$$

For example,

- in the case Euler method
$$\Phi(x_n, y_n, h) = f(x_n, y_n),$$
- in the case of improved Euler method
$$\Phi(x_n, y_n; h) = \frac{1}{2}\big(f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n))\big).$$

In order to asses the accuracy of the numerical scheme (4.17), we define error e_n at each grid point by

$$e_n = y(x_n) - y_n.$$

Further, for our convenience, we define

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h) \;\text{ Or }\; y(x_{n+1}) = y(x_n) + h\Phi(x_n, y(x_n); h) + hT_n \tag{4.18}$$

Then next theorem provides a bound on the error e_n.

Theorem 4.4.1. *Consider the general one step method (4.17). We assume that Φ is continuous in a region $R \subset \mathbb{R}^2$ containing the initial point (x_0, y_0) and there exists a positive constant L_Φ such that*

$$|\Phi(x, y; h) - \Phi(x, z; h)| \le L_\Phi |y - z| \;\; \forall\, (x, y),\ (x, z) \in R. \tag{4.19}$$

Then we have following error bound

$$|e_k| \le \exp^{L_\Phi(x_k - x_0)}\, |e_0| + \Big[\frac{\exp^{L_\Phi(x_k - x_0)} - 1}{L_\Phi}\Big] T, n = 0,\ 1,\ 2, \ldots,\ n, \tag{4.20}$$

where $T = \max_{0 \le k \le n-1} |T_k|$.

Proof. Subtracting (4.17) from (4.18) we arrive at

$$e_{k+1} = e_k + h\big[\Phi(x_k, y(x_k); h) - \Phi(x_k, y_k; h)\big] + hT_k. \tag{4.21}$$

Then the condition (4.19) yields

$$\begin{aligned} |e_{k+1}| \le |e_k| + hL_\Phi|e_k| + hT_k &= (1 + hL_\Phi)|e_k| + h|T_k| \\ &\le (1 + hL_\Phi)|e_k| + h|T|. \end{aligned} \tag{4.22}$$

Recursively, we obtain

$$\begin{aligned} |e_{k+1}| &\leq (1+hL_\Phi)|e_k| + h|T| \\ &\leq (1+hL_\Phi)^k|e_0| + h\big[1+(1+hL_\Phi)+(1+hL_\Phi)^2+\ldots+(1+hL_\Phi)^{k-1}\big]|T| \\ &= (1+hL_\Phi)^k|e_0| + \big[(1+hL_\Phi)^k-1\big]T/L_\Phi. \end{aligned} \tag{4.23}$$

Finally, use the fact that $1+hL_\Phi \leq \exp^{hL_\Phi}$ to obtain the desire result.

Observations:

- The single step scheme (4.17) is consistent if and only if $\Phi(x,y;0) = f(x,y)$.

  Proof. Observe that the truncation error is given by

  $$\begin{aligned} ((T-T_h)(y))(x_n) &= y'(x_n) - f(x_n,y(x_n)) - \frac{y(x_{n+1})-y(x_n)}{h} + \Phi(x_n,y(x_n);h) \\ &= y'(x_n) - \frac{y(x_{n+1})-y(x_n)}{h} + \big[\Phi(x_n,y(x_n);h) - f(x_n,y(x_n))\big] \\ &= \mathrm{O}(h) + \big[\Phi(x_n,y(x_n);h) - f(x_n,y(x_n))\big]. \end{aligned}$$

  Clearly, the truncation error tends to zero provided

  $$\lim_{h\to 0}\Phi(x,y;h) = \Phi(x,y;0) = f(x,y).$$

- As an immediate consequence, we observe that the modified Euler scheme is consistent.

Remark:

- For the modified Euler method

  $$\Phi(x,y(x);h) = \frac{1}{2}\big(f(x,y(x)) + f(x+h, y(x)+hf(x,y(x)))\big)$$

  and so

  $$\begin{aligned} T_n &= \frac{y(x_{n+1})-y(x_n)}{h} - \Phi(x_n,y(x_n);h) \\ &= \frac{y(x_{n+1})-y(x_n)}{h} - \frac{1}{2}\big(f(x_n,y(x_n)) + f(x_{n+1},y(x_n)+hf(x_n,y(x_n)))\big) \\ &= \frac{y(x_{n+1})-y(x_n)}{h} - \frac{1}{2}\big(g(x_n)+g(x_{n+1})-g(x_{n+1}) + f(x_{n+1},y(x_n)+hf(x_n,y(x_n)))\big) \\ &= \frac{y(x_{n+1})-y(x_n)}{h} - \frac{g(x_n)+g(x_{n+1})}{2} \\ &\quad + \frac{1}{2}\big(g(x_{n+1}) - f(x_{n+1},y(x_n)+hf(x_n,y(x_n)))\big), \quad g(x) = f(x,y(x)) \\ &= \frac{y(x_{n+1})-y(x_n)}{h} - \frac{g(x_n)+g(x_{n+1})}{2} \\ &\quad + \frac{1}{2}\big(f(x_{n+1},y(x_{n+1})) - f(x_{n+1},y(x_n)+hf(x_n,y(x_n)))\big) \end{aligned}$$

Then try to establish following estimates

$$\begin{aligned}\frac{y(x_{n+1})-y(x_n)}{h} &= y'(x_{n+\frac{1}{2}})+\mathrm{O}(h^2)=f(x_{n+\frac{1}{2}},y(x_{n+\frac{1}{2}}))+\mathrm{O}(h^2)=g(x_{n+\frac{1}{2}}))+\mathrm{O}(h^2)\\ &= \frac{g(x_n)+g(x_{n+1})}{2}+\mathrm{O}(h^2)+\mathrm{O}(h^2).\end{aligned}$$

Then use MVT to have

$$\begin{aligned}f(x_{n+1},y(x_{n+1}))-f(x_{n+1},y(x_n)+hf(x_n,y(x_n)) &= \frac{\partial f}{\partial y}(y(x_{n+1})-y(x_n)-hf(x_n,y(x_n)))\\ &= \frac{\partial f}{\partial y}\times\mathrm{O}(h^2).\end{aligned}$$

Finally, we have $T_n=\mathrm{O}(h^2)$.

- For the grid points
$$x_0,x_1,x_2,\ldots,x_n=b,$$
along with Theorem 4.4.1, we obtain
$$\begin{aligned}|e_n| &\le \Big[\frac{\exp^{L_\Phi(x_n-x_0)}-1}{L_\Phi}\Big]\mathrm{O}(h^2),\ \text{ since }\ e_0=0\\ &= \Big[\frac{\exp^{L_\Phi(b-x_0)}-1}{L_\Phi}\Big]\mathrm{O}(h^2)\end{aligned}$$
Therefore, modified Euler method has second order convergence.

4.5 Runge-Kutta method

Consider the master equation

$$y_{j+1}=y_j+hF(*)\qquad j\ge 0, \tag{4.24}$$

where F is the generalized slope of the solution curve $y=y(x)$ of the IVP (4.1) – (4.2). For an n-order method, we assume

$$F=\sum_{j=1}^{n}a_jK_j, \tag{4.25}$$

where a_j's are constants and K_j's are slopes defined as

$$K_1=f(x_j,y_j), \tag{4.26}$$
$$K_2=f(x_j+p_1h,y_j+q_{11}K_1h), \tag{4.27}$$
$$K_3=f(x_j+p_2h,y_j+q_{21}K_1h+q_{22}K_2h), \tag{4.28}$$
$$\vdots \tag{4.29}$$
$$K_n=f(x_j+p_{n-1}h,y_j+q_{n-1,1}K_1h+q_{n-1,2}K_2h+\cdots+q_{n-1,n-1}K_{n-1}h), \tag{4.30}$$

where p_j's and q_{ij}'s are constants to be determined. Below we discuss the detailed derivation of second order Runge-Kutta (RK) method.

4.5.1 Second-order RK method

For a second-order RK method, the master equation (4.24) takes the form

$$y_{j+1} = y_j + h(a_1 K_1 + a_2 K_2), \tag{4.31}$$

where

$$K_1 = f(x_j, y_j), \tag{4.32}$$

$$K_2 = f(x_j + p_1 h, y_j + q_{11} K_1 h), \tag{4.33}$$

where $a_1,\ a_2,\ p_1$ and q_{11} are the unknown constants to be determined.

Using Taylor series expansion of $y(x_j + h)$ about the point x_j, we obtain

$$\begin{aligned} y_{j+1} &:= y(x_j + h) \\ &= y(x_j) + hy'(x_j) + \frac{h^2}{2!} y''(x_j) + \mathbf{O}(h^3) \\ &= y(x_j) + hy'(x_j) + \frac{h^2}{2!} [f_x(x_j, y_j) + f(x_j, y_j) f_y(x_j, y_j)] + \mathbf{O}(h^3). \end{aligned} \tag{4.34}$$

Taylor series of K_2 about the point (x_j, y_j) gives

$$\begin{aligned} K_2 &= f(x_j + p_1 h, y_j + q_{11} K_1 h) \\ &= f(x_j, y_j) + f_x(x_j, y_j) p_1 h + f_y(x_j, y_j) q_{11} K_1 h + \mathbf{O}(h^2). \end{aligned} \tag{4.35}$$

Using (4.34) and (4.35) in (4.31)–(4.33), we obtain

$$\begin{aligned} &y(x_j) + hf(x_j, y_j) + \frac{h^2}{2!} [f_x(x_j, y_j) + f(x_j, y_j) f_y(x_j, y_j)] + \mathbf{O}(h^3) \\ &\quad = y(x_j) + h(a_1 + a_2) f(x_j, y_j) + h^2 [p_1 a_2 f_x(x_j, y_j) + q_{11} a_2 f(x_j, y_j) f_y(x_j, y_j)], \end{aligned} \tag{4.36}$$

which yields

$$a_1 + a_2 = 1, \tag{4.37}$$

$$p_1 a_2 = \frac{1}{2!}, \tag{4.38}$$

$$q_{11} a_2 = \frac{1}{2!}. \tag{4.39}$$

The above under-determined system of equations has multiple solutions. We discuss two cases as follows.

Case 1: Huen's method

Choosing $a_1 = 1/2$ we obtain $a_2 = 1/2$, $p_1 = 1$ and $q_{11} = 1$. Therefore, the resultant second order RK method is

$$y_{j+1} = y_j + \frac{h}{2} [f(x_j, y_j) + f(x_j + h, y_j + hf(x_j, y_j))]. \tag{4.40}$$

Case 2: Mid-point rule

Choosing $a_2 = 1$ we obtain $a_1 = 0$, $p_1 = 1/2$ and $q_{11} = 1/2$. Therefore, the resultant second order RK method is

$$y_{j+1} = y_j + hf\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}hf(x_j, y_j)\right). \tag{4.41}$$

4.5.2 Third-order RK method

Third order RK method is given by

$$y_{j+1} = y_j + h\frac{1}{6}(K_1 + 4K_2 + K_3), \tag{4.42}$$
$$K_1 = f(x_j, y_j), \tag{4.43}$$
$$K_2 = f\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}K_1 h\right), \tag{4.44}$$
$$K_3 = f(x_j + h, y_j - K_1 h + 2K_2 h). \tag{4.45}$$

4.5.3 Fourth-order RK method

Third order RK method is given by

$$y_{j+1} = y_j + h\frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4), \tag{4.46}$$
$$K_1 = f(x_j, y_j), \tag{4.47}$$
$$K_2 = f\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}K_1 h\right), \tag{4.48}$$
$$K_3 = f\left(x_j + \frac{1}{2}h, y_j + \frac{1}{2}K_2 h\right), \tag{4.49}$$
$$K_4 = f(x_j + h, y_j + K_3 h). \tag{4.50}$$

4.5.4 Some important points

1. Explicit Euler method has local truncation error of $\mathbf{O}(h^2)$ and global truncation error of $\mathbf{O}(h)$.
2. Huen's method uses prediction and correction to improved the Euler method. It uses
   (a) the Euler method as a predictor, &
   (b) the trapezoidal rule as a corrector.
3. Midpoint rule improves the Euler method.
4. RK methods use linear combination of slopes.
5. Huen's method and midpoint rule belong to the class of second-order RK methods.
6. All the methods discussed can be easily extended to a system of equations.

4.6 Taylor's series method

Assuming that the solution curves of the IVP (4.1)–(4.2) are n times continuously differentiable, i.e., $y', ; y'',\ y''', \cdots y^{(n)}$ exist and are continuous, we can express second and higher order derivatives of y in terms of $f(x, y)$ and its partial derivatives of first and higher orders as follows:

$$y'' = \frac{\mathrm{d}}{\mathrm{d}x} f(x, y) \equiv f_x + f f_y, \tag{4.51}$$

$$y''' = \frac{\mathrm{d}^2}{\mathrm{d}x^2} f(x, y) \equiv f_{xx} + 2 f f_{yx} + f_x f_y + f^2 f_{yy} + f f_y^2, \tag{4.52}$$

$$\vdots$$

$$y^{(n)} = \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}} f(x, y). \tag{4.53}$$

Taylor series expansion of $y(x_h + h)$ around x_j gives

$$y(x_j+h) = y(x_j)+hy'(x_j)+\frac{h^2}{2!}y''(x_j)+\cdots+\frac{h^n}{n!}y^{(n)}(x_j)+\frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi_j), \quad x_j \le \xi_j \le x_{j+1}. \tag{4.54}$$

Writing derivatives of y in terms of f and its partial derivatives and neglecting the remainder term in (4.54), we obtain n-order Taylor series method. Local truncation error of n-order Taylor series method is $\mathbf{O}(h^{n+1})$. For example, the third-order Taylor series method reads as

$$y_{j+1} = y_j+hf(x_j, y_j)+\frac{h^2}{2!}(f_x+ff_y)_{(x_j,y_j)}+\frac{h^3}{3!}(f_{xx}+2ff_{yx}+f_xf_y+f^2f_{yy}+ff_y^2)_{(x_j,y_j)}. \tag{4.55}$$

4.7 Stiff differential equations

Consider the first-order linear ODE

$$\frac{\mathrm{d}y}{\mathrm{d}t} + P(t)y = Q(t). \tag{4.56}$$

This ODE can be solved exactly for continuous $P(t)$ and $Q(t)$. Assume $P(t) = a$, $Q(t) = b + ce^{\alpha t}$, where $a,\ b,\ c,\ \alpha \in \mathbb{R}$. The exact solution of this linear ODE is

$$y(y) = Ke^{-at} + \frac{b}{a} + c\frac{1}{a-\alpha}e^{-\alpha t}, \tag{4.57}$$

where K is the arbitrary constant. Using initial condition $y(0) = 0$, we obtain

$$K = -\left(\frac{b}{a} + c\frac{1}{a-\alpha}\right).$$

Therefore, the required solution of the IVP is

$$y(t) = -\left(\frac{b}{a} + c\frac{1}{a-\alpha}\right)e^{-at} + \frac{b}{a} + c\frac{1}{a-\alpha}e^{-\alpha t}. \tag{4.58}$$

4.7.1 Stability of a numerical method

The following linear ODE

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -ay, \qquad a > 0 \tag{4.59}$$

possesses an analytic solution $y = ce^{-at}$, where c is an arbitrary constant, that decays with t.

If a numerically approximated solution fails to capture this decay, it is called unstable. Here, we investigate stability of Euler method. Applying Euler method on (4.59), we obtain

$$\begin{aligned} y_{j+1} &= (1-ah)y_j, \quad j = 0, 1, \cdots, n-1. \\ \Rightarrow y_n &= (1-ah)y_{n-1} \\ &= (1-ah)^2 y_{n-2} \\ &\vdots \\ &= (1-ah)^n y_0. \end{aligned} \tag{4.60}$$

Therefore, the numerical approximation of the solution decays relative to the initial value y_0 provided $|1-ah| < 1$ – implying that the Euler method is stable if $h < 2/a$ and unstable if $h > 2/a$. Therefore, the Euler method is *conditionally stable*.

For implicit/backward Euler method,

$$y_n = \frac{1}{(1+ah)^n} y_0. \tag{4.61}$$

For $a > 0$, we have $1/(1+ah) < 1$ for every positive step size h leading to an *unconditional stability* of implicit/backward Euler method.

4.8 Linear Multistep Methods

The general form of the multistep method to be considered is

$$y_{n+1} = \sum_{j=0}^{p} a_j y_{n-j} + h \sum_{j=-1}^{p} b_j f(x_{n-j}, y_{n-j}), \qquad n \geq p, \tag{4.62}$$

where the coefficients $a_0, \ldots, a_p$, $b_{-1}, b_0, \ldots, b_p$ are constants, and $p \geq 0$. If either $a_p \neq 0$ or $b_p \neq 0$, then the method is called a $p+1$ step method, as the previous $p+1$ solution values are used to compute y_{n+1}. The values $y_1, \ldots, y_p$ must be obtained using other means, e.g., Taylor's method, RK method, etc. Euler's method is an example of one-step method with $p = 0$ and

$$a_0 = 1, \qquad b_{-1} = 0, \qquad b_0 = 1.$$

If $b_{-1} = 0$, then (4.62) represents *explicit methods*. If $b_{-1} \neq 0$, (4.62) represents *implicit* methods. The existence of the solution y_{n+1}, for all sufficiently small h, can be shown by using the fixed-point theory. Implicit methods are generally solved using iterations.

While Runge-Kutta methods give an improvement over Euler's method in terms of accuracy, this is achieved by investing additional computational effort; in fact, Runge-Kutta methods require more evaluations of $f(\cdot,\cdot)$ than would seem necessary. For example, the fourth-order method involves four function evaluations per step. For comparison, by considering three consecutive points $x_{n-1},\ x_n = x_{n-1} + h,\ x_{n+1} = x_{n-1} + 2h$, integrating the differential equation $y' = f(x, y)$ between x_{n-1} and x_{n+1}, yields

$$y(x_{n+1}) = y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(x, y(x))dx, \tag{4.63}$$

and applying Simpson's rule to approximate the integral on the right-hand side then leads to the method

$$y_{n+1} = y_{n-1} + \frac{h}{3}\Big[f(x_{n-1}, y_{n-1}) + 4f(x_n, y_n) + f(x_{n+1}, y_{n+1})\Big], \tag{4.64}$$

requiring only three function evaluations per step. In contrast with the one-step methods considered in the previous section where only a single value y_n was required to compute the next approximation y_{n+1}, here we need two preceding values, y_n and y_{n-1}, to be able to calculate y_{n+1}, and therefore (4.64) is not a one-step method. In this lecture we consider a class of methods of the type (4.64) for the numerical solution of the initial value problem

$$y' = f(x, y),\ \ y(x_0) = y_0, \tag{4.65}$$

called linear multistep methods. Further, due to implicit dependence on y_{n+1} the method is then called implicit. The method (4.64) is called linear because it involves only linear combinations of the y_{n+j} and the $f(x_{n+j}, y_{n+j})$. For the sake of notational simplicity, henceforth we shall often write f_n instead of $f(x_n, y_n)$.
Remark:

- The method (4.64) is known as classical Milne's corrector method. Like Improved Euler method, we predict y_{n+1} and then correct/improved it using Milne's corrector method (4.64). Therefore, we need Milne's predictor method.

4.8.1 Milne's Predictor Method

Integrating the differential equation $y' = f(x, y)$ between x_{n-3} and x_{n+1}, yields

$$y(x_{n+1}) = y(x_{n-3}) + \int_{x_{n-3}}^{x_{n+1}} f(x, y(x))dx, \tag{4.66}$$

Then use Lagrange polynomial approximation for $g(x) = f(x, y(x))$ based on four mesh points

$$(x_{n-3}, f_{n-3}),\ (x_{n-2}, f_{n-2}),\ (x_{n-1}, f_{n-1}),\ (x_n, f_n)$$

to have

$$y_{n+1} = y_{n-3} + \frac{4h}{3}\big(2f_{n-2} - f_{n-1} + 2f_n\big), \tag{4.67}$$

which is an explicit multistep method.

Example 4.8.1. *Consider* $y' = 1 + y^2,\ \ y(0) = 0$. *Find approximations at* 0.2, 0.4 *and* 0.6 *using RK-4 method. Then using Milne's method, evaluate the approximation at* 0.8 *and* 1.0.

Solution: Recall classical fourth-order method:

$$y_{n+1} = y_n + \frac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4),$$

where

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1), \\ k_3 &= f(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_2), \\ k_4 &= f(x_n + h, y_n + hk_3). \end{aligned}$$

We take $h = 0.2$, so that grid points are

$$x_0 = 0,\ x_1 = x_0 + h = 0.2,\ x_2 = 0.4,\ x_3 = 0.6,\ x_4 = 0.8,\ x_5 = 1.0,\ \ldots.$$

Then calculate $y_1,\ y_2$ and y_3. In fact

$$y_1 = 0.2027,\ y_2 = 0.4228,\ y_3 = 0.6841.$$

Using Milne's predictor method (4.67), we obtain

$$\begin{aligned} y_4 &= y_0 + \frac{4h}{3}\big(2f_1 - f_2 + 2f_3\big) \\ &= 0 + \frac{4 \times 0.2}{3}\Big(2f(x_1, y_1) - f(x_2, y_2) + 2f(x_3, y_3)\Big) = 1.0239. \end{aligned}$$

Then, to correct the value y_4, apply Milne's corrector method (4.64) so that

$$\begin{aligned} y_4 &= y_3 + \frac{h}{3}\Big[f(x_2, y_2) + 4f(x_3, y_3) + f(x_4, y_4)\Big] && (4.68) \\ &= 1.0294 && (4.69) \end{aligned}$$

4.8.2 Adams methods

Integrating (4.1) on (x_n, x_{n+1}), we obtain

$$\int_{x_n}^{x_{n+1}} \frac{\mathrm{d}y}{\mathrm{d}x}\mathrm{d}x = \int_{x_n}^{x_{n+1}} f(x, y(x))\mathrm{d}x \Rightarrow y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(x, y(x))\mathrm{d}x. \qquad (4.70)$$

The integral on the right-hand side will be approximated by a numerical quadrature scheme, and the resultant will be numerical method for solving the IVP (4.1)-(4.2).

Adams methods are based on the idea of approximating the integrand with a polynomial within the interval (x_n, x_{n+1}). Using a p th order polynomial results in a $(p+1)$th order method. There are two types of Adams methods, the explicit and the implicit types. The explicit type is called the Adams-Bashforth (AB) methods and the implicit type is called the Adams-Moulton (AM) methods. AB method is used as predictor method and AM method is used as corrector method.

Adams-Bashforth method

Suppose the resulting formula is of the form

$$y_{n+1} = y_n + b_0 f_n + b_1 f_{n-1} + b_2 f_{n-2} + b_3 f_{n-3}, \tag{4.71}$$

i.e., $a_1 = 0,\ a_j = 0,\ j \geq 2$, and $b_{-1} = 0,\ b_j = 0,\ j \geq 4$ in (4.62). Here, $f_n = f(x_n, y_n)$. An equation of this type is called an Adams-Bashforth formula. Here, we derive **Adams-Bashforth method of order 4** based on equally spaced points $x_j = x_0 + jh,\ 0 \leq j \leq n$.

To determine the coefficients $b_k,\ 0 \leq k \leq 3$, we approximate the integral in (4.70) as

$$\int_{x_n}^{x_{n+1}} f(x, y(x))\mathrm{d}x \approx h[B_0 f_n + B_1 f_{n-1} + B_2 f_{n-2} + B_3 f_{n-3}]. \tag{4.72}$$

The coefficients $B_0,\ B_1,\ B_2,\ B_3$ are determined by requiring that (4.77) be exact whenever the integrand is a polynomial of degree ≤ 3. Without any loss of generality, we can assume $x_0 = 0$ and $h = 1$ such that $x_{n-j} = -j,\ j = 1, 2, 3$. Substituting $f(x, y) = 1,\ x,\ x^2$, and x^3 in (4.77) we obtain

$$\begin{aligned} B_0 + B_1 + B_2 + B_3 &= 1, & (4.73)\\ -B_1 - 2B_2 - 3B_3 &= \frac{1}{2}, & (4.74)\\ B_1 + 4B_2 + 9B_3 &= \frac{1}{3}, & (4.75)\\ -B_1 - 8B_2 + 27B_3 &= \frac{1}{4}. & (4.76) \end{aligned}$$

Solving the above-system of linear algebraic equations, we obtain the desired coefficients and the **fourth-order Adams-Bashforth formula** is

$$y_{n+1} = y_n + \frac{h}{24}[55 f_n - 59 f_{n-1} + 37 f_{n-2} - 9 f_{n-3}]. \tag{4.77}$$

The procedure just illustrated is called the **method of undetermined coefficients**.

Adams-Multon method

Similarly, we can derive the **fourth-order Adams-Moulton formula**

$$y_{n+1} = y_n + \frac{h}{24}[9 f_{n+1} + 19 f_n - 5 f_{n-1} + f_{n-2}]. \tag{4.78}$$

Chapter 5

Matrix Algebra

5.1 Preliminaries

Theorem 5.1.1 (Equivalent systems)**.** *If one system of equations is obtained from another by a finite sequence of elementary operations, then the two systems are equivalent.*

Theorem 5.1.2 (Right inverse)**.** *A square matrix can possess at most one right inverse.*

Theorem 5.1.3 (Matrix inverse)**.** *If A and B are square matrices such that $AB = I$, then $BA = I$.*

Theorem 5.1.4 (Nonsingular matrix properties)**.** *For an $n \times n$ matrix A, the following properties are equivalent:*

1. *The inverse of A exists: that is, A is nonsingular.*
2. *The determinant of A is nonzero.*
3. *The rows of A form a basis for $\mathbb{R}^n$.*
4. *The columns of A for a basis for $\mathbb{R}^n$.*
5. *As a map from $\mathbb{R}^n$ to $\mathbb{R}^n$, A is injective (one to one).*
6. *As a map from $\mathbb{R}^n$ to $\mathbb{R}^n$, A is surjective (onto).*
7. *The equation $Ax = 0$ implies $x = 0$.*
8. *For each $b \in \mathbb{R}^n$, there is exactly one $x \in \mathbb{R}^n$ such that $Ax = b$.*
9. *A is a product of elementary matices.*
10. *0 is not an eigenvalue of A.*

Definition 5.1.1 (Positive definite matrix)**.** *A matrix A is positive definite if $x^{\mathrm{T}}Ax > 0$ for every nonzero x.*

Theorem 5.1.5 (LU-Decomposition)**.** *If all n leading principal minors of the $n \times n$ matrix A are nonsingular, then A has an LU-decomposition.*

Theorem 5.1.6 (Cholesky Theorem on LL^{T}-Factorization)**.** *If A is a real, symmetric, and positive definite matrix, then it has a unique factorization, $A = LL^{\mathrm{T}}$, in which L is lower triangular with a positive diagonal.*

Theorem 5.1.7. *Define a permutation matrix P whose elements are $P_{ij} = \delta p, j$. Define an upper triangular matrix U whose elements are $u_{ij} = a_{p,j}^{(n)}$ if $j \geq i$. Define a unit lower triangular matrix L whose elements are $l_{ij} = a_{p,j}^{(n)}$ if $j < i$. Then $PA = LU$.*

Theorem 5.1.8. *If the factorization $PA = LU$ is produced from the Gaussian algorithm with scaled row pivoting, then the solution of $Ax = b$ is obtained by first solving $Lz = Pb$ and then solving $Ux = z$. Similarly, the solution of $y^{\mathrm{T}}A = c^{\mathrm{T}}$ is obtained by solving $U^{\mathrm{T}}x = c$ and then $L^{\mathrm{T}}Py = z$.*

Theorem 5.1.9 (Theorem on Long Operations)**.** *If Gaussian elimination is used with scaled row pivoting, then the solution of the system $Ax = b$, with fixed A, and m different vectors b, involves approximately*

$$\frac{1}{3}n^3 + \left(\frac{1}{2} + m\right) n^2$$

long operations (multiplications and divisions).

Definition 5.1.2 (Diagonally dominant)**.** *A matrix A is called* diagonally dominant *if*

$$|a_{ij}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \qquad (1 \leq i \leq n).$$

Theorem 5.1.10 (Theorem on Preserving Diagonal Dominance)**.** *Gaussian elimination without pivoting preserves the diagonal dominance of a matrix.*

Corollary 5.1.10.1. *Every diagonally dominant matrix is non-singular and has an LU-factorization.*

Corollary 5.1.10.2. *If the scaled row pivoting version of Gaussian elimination recomputes the scale arrays after each major step and is applied to a diagonally dominant matrix, then the pivots will be the natural ones: $1, 2, \ldots, n$. Hence, the work of choosing the pivots can be omitted in this case.*

Theorem 5.1.11. *If $\|\cdot\|$ is any norm on $\mathbb{R}^n$, then the equation*

$$\|A| = \sup_{\|u\|=1} \{\|Au\| : u \in \mathbb{R}^n\}$$

defines a norm on the linear space of all $n \times n$ matrices.

Theorem 5.1.12. *If the vector norm $\|\cdot\|_\infty$ is defined by*

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|$$

then its subordinate matrix norm is given ny

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

If we solve a system of equations

$$Ax = b$$

numerically, we obtain not the exact solution x but an approximate solution $\tilde{x}$. One can test $\tilde{x}$ by forming $A\tilde{x}$ to see whether it is close to b. Thus, we obtain the *residual vector*

$$r = b - A\tilde{x}.$$

The difference between the exact solution x and the approximate solution $\tilde{x}$ is called the *error vector*

$$e = x - \tilde{x}.$$

The following relationship

$$Ae = r$$

between the error vector and the residual vector is of fundamental importance.

Theorem 5.1.13. *In solving system of equations $Ax = b$, the condition number $\kappa(A) := \|A\| \cdot \|A^{-1}\|$, the residual vector r, and the error vector e satisfy the following inequalities:*

$$\frac{1}{\kappa(A)} \frac{\|r\|}{\|b\|} \le \frac{\|e\|}{\|x\|} \le \kappa(A) \frac{\|r\|}{\|b\|}.$$

A matrix with a large condition number is said to be *ill conditioned*. For an ill-conditioned matrix A, there will be cases in which the solution of a system $Ax = b$ will be very sensitive to small changes in the vector. If the condition number of the matrix A is of moderate size, the matrix is said to be *well conditioned*.

5.2 Solution of equations by iterative methods

Consider solving the system of equations

$$Ax = b \tag{5.1}$$

using an iterative process. Assume that a matrix Q, called the *splitting matrix*, is prescribed. The original problem is rewritten in the equivalent form

$$Qx = (Q - A)x + b. \tag{5.2}$$

Equation (5.2) suggests an iterative process, defined by writing

$$Qx^{(k)} = (Q - A)x^{(k-1)} + b \qquad (k \ge 1). \tag{5.3}$$

The initial vector $x^{(0)}$ can be arbitrary. If a good guess for the solution is available, it should be used for $x^{(0)}$.

Theorem 5.2.1 (Theorem on Iterative Method Convergence)**.** *If $\|I - Q^{-1}A\| < 1$ for some subordinate matrix norm, then the sequence produced by equation* (5.3) *converges to the solution of $Ax = b$ for any initial vector $x^{(0)}$.*

Theorem 5.2.2 (Theorem on Convergence of Jacobi Method)**.** *If A is diagonally dominant, then the sequence produced by the Jacobi iteration converges to the solution of $Ax = b$ for any starting vector.*

Theorem 5.2.3 (Theorem on Similar Upper Triangular Matrices)**.** *Every square matrix is similar to an (possibly complex) upper triangular matrix whose off-diagonal elements are arbitrarily small.*

Theorem 5.2.4 (Theorem of Spectral Radius)**.** *The spectral radius function satisfies the equation*

$$\rho(A) = \inf_{\|\cdot\|} \|A\|$$

in which the infimum is taken over all subordinate matrix norms.

Theorem 5.2.5 (Theorem on Necessary and Sufficient Conditions for Iterative Method Convergence)**.** *In order that the iterative formula*

$$x^{(k)} = Gx^{(k-1)} + c$$

produce a sequence converging to $(I - G)^{-1}c$, for any starting vector $x^{(0)}$, it is necessary and sufficient that the spectral radius of G be less than 1*.*

Corollary 5.2.5.1. *The iterative formula* (5.3) *will produce a sequence converging to the solution of $Ax = bm$ for any $x^{(0)}$, if $\rho(I - Q^{-1}A) < 1$.*

Theorem 5.2.6 (Theorem on Gauss-Seidel Method Convergence)**.** *If A is diagonally dominant, then the Gauss-Seidel method converges for any starting vector.*

Chapter 6

Boundary value problems of ODEs

6.1 Introduction

Consider the following second order ODE

$$y'' = f(x, y, y'), \tag{6.1}$$

where f is sufficiently smooth function. Equation (6.1) associated with the following conditions

$$x(a) = \alpha \qquad x'(a) = \beta \tag{6.2}$$

form an initial value problem (IVP). On the other hand, equation (6.1) along with

$$x(a) = \alpha \qquad x(b) = \beta \tag{6.3}$$

conditions consist a *boundary value problem (BVP)*. The conditions (6.3) are called *Dirichlet boundary conditions*.

Theorem 6.1.1 (Existence Theorem, Boundary Value Theorem)**.** *The boundary value problem*

$$y'' = f(x, y) \tag{6.4}$$
$$y(0) = 0 \quad y(1) = 0 \tag{6.5}$$

has a unique solution if $\partial f/\partial x$ is continuous, nonnegative, and bounded in the infinite strip defined by the inequalities $0 \leq x \leq 1,\ -\infty < x < \infty$.

Theorem 6.1.2 (First Theorem on Two-Point Boundary-Value Problems)**.** *Consider these two-point boundary-value problems:*

$$y'' = f(x, y) \tag{6.6}$$
$$y(a) = \alpha \qquad y(b) = \beta \tag{6.7}$$
$$y'' = g(x, y) \tag{6.8}$$
$$y(0) = \alpha \qquad y(1) = \beta \tag{6.9}$$

in which

$$g(p,q) = (b-a)^2 f(a+(b-a)p, q).$$

If z us a solution of Problem (6.8)–(6.9), then the function y defined by

$$y(x) = z(x-a)/(b-a)$$

is a solution of Problem (6.6)–(6.7). Moreover, if y is a solution of Problem (6.6)–(6.7), then

$$z(x) = y(a+(b-a)x)$$

is a solution of Problem (6.8)–(6.9).

Proof. See pp 575 of [Cheney and Kincaid, 2012]. □

Theorem 6.1.3 (Second Theorem on Two-Point Boundary-Value Problems)**.** *Consider these two-point boundary-value problems:*

$$y'' = f(x,y) \tag{6.10}$$
$$y(0) = \alpha \qquad y(1) = \beta \tag{6.11}$$
$$y'' = g(x,y) \tag{6.12}$$
$$y(0) = 0 \qquad y(1) = 0 \tag{6.13}$$

in which

$$g(p,q) = f(p, q+\alpha+(\beta-\alpha)p).$$

If z us a solution of Problem (6.8)–(6.9), then the function y defined by

$$y(x) = z(x) + \alpha + (\beta-\alpha)x$$

is a solution of Problem (6.6)–(6.7). Moreover, if y is a solution of Problem (6.6)–(6.7), then

$$z(x) = y(x) - [\alpha + (\beta-\alpha)x]$$

is a solution of Problem (6.8)–(6.9).

Theorem 6.1.4. *Let f be a continuous function of (x,y), where $0 \le x \le 1$ and $-\infty < y < \infty$. Assume that on this domain*

$$|f(t,y_1) - f(t,y_2)| \le |y_1 - y_2|$$

Then the two-point BVP

$$y'' = f(x,y) \qquad y(0) = y(1) = 0$$

has a unique solution in $C[0,1]$.

Chapter 7

Finite difference methods of PDEs

7.1 Preliminaries

Definition 7.1.1 (Local Truncation Error (LTE)). *Let $\mathbb{F}(u) = 0$ represents the difference equation approximating the PDE $L(U) = 0$ at the (i, j)-th mesh points. Here, u is the exact solution of the difference equation and U corresponds to the exact solution of the PDE. If we replace u by U at the mesh point of the difference equation, the value of $F_{i,j}(U)$ is called the local truncation error (LTE) $T_{i,j}$ at the (i, j) mesh point.*

$F_{i,j}(U)$ clearly measures the amount by which the exact solution values of the PDE at the mesh points of the difference equation at the point $(i\Delta x, j\Delta t)$.

Example 7.1.1. *Consider the parabolic equation*

$$U_t = U_{xx} \qquad in\ \Omega. \tag{7.1}$$

Consider the Richardson explicit difference approximation

$$\frac{u_p^{q+1} - u_p^{q-1}}{2\Delta t} - \frac{u_{p+1}^q - 2u_p^q + u_{p-1}^q}{\Delta x^2} = 0, \quad p \in \mathbb{N},\ q \geq 0. \tag{7.2}$$

We calculate LTE:

$$T_{i,j} := F_{i,j}(U) = \frac{U_p^{q+1} - U_p^q}{2\Delta t} - \frac{U_{p+1}^q - 2U_p^q + U_{p-1}^q}{\Delta x^2}. \tag{7.3}$$

By Taylor's expansion, we obtain

$$\begin{aligned}
U_{p\pm1}^q &= U(x_p \pm \Delta x, t_q) \\
&= U_p^q \pm \Delta x \left(\frac{\partial U}{\partial x}\right)_{p,q} + \frac{\Delta x^2}{2!}\left(\frac{\partial^2 U}{\partial x^2}\right)_{p,q} \pm \frac{\Delta x^3}{3!}\left(\frac{\partial^3 U}{\partial x^3}\right)_{p,q} + \frac{\Delta x^4}{4!}\left(\frac{\partial^4 U}{\partial x^4}\right)_{p,q} \\
&\qquad \pm \frac{\Delta x^5}{5!}\left(\frac{\partial^5 U}{\partial x^5}\right)_{p,q} + \frac{\Delta x^6}{6!}\left(\frac{\partial^6 U}{\partial x^6}\right)_{p,q} + \cdots \\
U_p^{q\pm1} &= U(x_p, t_q \Delta t) \\
&= U_p^q \pm \Delta t \left(\frac{\partial U}{\partial t}\right)_{p,q} + \frac{\Delta t^2}{2!}\left(\frac{\partial^2 U}{\partial t^2}\right)_{p,q} \pm \frac{\Delta t^3}{3!}\left(\frac{\partial^3 U}{\partial t^3}\right)_{p,q} + \cdots
\end{aligned}$$

Therefore,

$$F_{i,j}(U) = \left[\frac{\Delta t^2}{3}\left(\frac{\partial^3 U}{\partial t^3}\right)_{p,q} - \frac{\Delta x^3}{12}\left(\frac{\partial^4 U}{\partial x^2}\right)_{p,q}\right] + \mathbf{O}(\Delta t^4, \Delta x^4).$$

This indicates that LTE $T_{i,j} \to 0$ as $(\Delta x, \Delta t) \to (0, 0)$. Therefore, the Richardson method (7.2) is consistent with the diffusion equation (7.1).

7.2 Consistency, Stability and Convergence

7.2.1 Consistency

A finite difference scheme $F_{i,j}(u) = 0$ is said to be consistent with a PDE (with two independent variables x and t) if the local trunction error (LTE) converges to zero as $(\Delta t, \Delta x) \to (0, 0)$.

7.2.2 stability

The essential idea defining stability is that this numerical process, applied exactly, should limit the amplification of all components of the initial conditions.

Stability by the Fourier series method (von Neumann's method) Smith [1985]

Assume we are concerned with the stability of a linear two time-level difference equation in $u(x, t)$ in the time interval $0 \le t \le T = J\Delta t$, T finite, as $\Delta x \to 0$ and $\Delta t \to 0$, i.e., as $J \to \infty$. The Fourier series or von Neumann method expresses the initial values at the mesh points along $t = 0$ in terms of a finite Fourier series, then considers the growth of a function that reduces to this series for $t = 0$ by a 'variables separable' method identical to that commonly used for solving partial differential equations.

The Fourier series can be formulated in terms of sines and cosines but the algebra is easier if the complex exponential form is used, i.e., with $\sum a_n \cos(n\pi x/l)$ or $\sum b_n \sin(n\pi x/l)$ replaced by the equivalent $\sum A_n e^{\mathrm{i}n\pi x/l}$, where $\mathrm{i} = \sqrt{-1}$ and l is the x-interval throughout which the function is defined. We defined $u_p^q := u(p\Delta x, q\Delta t)$. In terms of this notation,

$$A_n e^{\mathrm{i}n\pi x/l} = A_n e^{\mathrm{i}n\pi p\Delta x/(N\Delta x)} = A_n e^{\mathrm{i}\beta_n p\Delta x},$$

where $\beta_n = n\pi/(N\Delta x)$ and $N\Delta x = l$.

Denote the initial values at the pivotal points along $t = 0$ by $u(p\Delta x, 0) = u_{p,0}, \quad p = 0, 1, \ldots, N$. Then the $N + 1$ equations

$$u_{p,0} = \sum_0^N A_n e^{\mathrm{i}\beta_n p\Delta x}, \quad p = 0, 1, \ldots, N, \tag{7.4}$$

are sufficient to determine the $(N + 1)$ unknowns $A_0, A_1, \ldots, A_N$ uniquely, showing that the initial mesh values can be expressed in this complex exponential form. As we are considering only linear-difference equations we need investigate the propagation of only one initial value,

such as $e^{\mathrm{i}\beta p\Delta x}$, because separate solutions are additive. The coefficient An is a constant and can be neglected.

To investigate the propagation of this term as t increases, put

$$u_p^q = e^{\mathrm{i}\beta x}e^{\alpha t} = e^{\mathrm{i}\beta p\Delta x}e^{\alpha q\Delta t} = e^{\mathrm{i}\beta p\Delta x}\xi^q,$$

where $\xi = e^{\alpha\Delta t}$ and α n general, is a complex constant. ξ often called the amplification factor.

The finite-difference equations will be stable by the Lax-Richtmyer definition if $|u_p^q|$ remains bounded for all $q \leq J$ as $\Delta x \to 0$ and $\Delta t \to 0$, and for all values of β needed to satisfy the initial conditions.

If the exact solution of the difference equations does not increase exponentially with time, then a necessary and sufficient condition for stability is that

$$|\xi| \leq 1. \tag{7.5}$$

If, however, u_p^q does increase with t, then the necessary and sufficient condition for stability is,

$$|\xi| \leq 1 + K\Delta t = 1 + \mathbb{O}(\Delta t), \tag{7.6}$$

where the positive number K is independent of Δx, Δt and β.

It should be noted that this method applies only to linear equations with constant coefficients, and strictly speaking only to initial value problems with periodic initial data, of period l. For difference equations involving three or more time-levels or two or more dependent variables, the von Neumann conditions (7.5) and (7.6) are always necessary but may not be sufficient Richtmyer and Morton [1994]. In practice, the method often gives useful results even when its application is not fully justified.

7.2.3 Convergence

The numerical scheme is said to be convergent (the exact solution of the difference equations converges to the exact solution of the PDE) if it is consistent and stable.

7.3 Vector and matrix norms

7.3.1 Vector norms

The norm of a vector $\mathbf{x}$ is a positive real number giving a measure of the 'size' of the vector and is denoted by $\|\mathbf{x}\|$ satisfying the following properties

1. $\|\mathbf{x}\| \geq 0, \quad \forall \mathbf{x}, \qquad \|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0},$
2. $\|c\mathbf{x}\| = |c|\|\mathbf{x}\| \quad \forall \mathbf{x}, \ \ c \in \mathbb{F},$
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y}.$

For $\mathbf{x} \in \mathbb{F}^n$, the most commonly used norms are as follows:

- **1-norm:** $\|\mathbf{x}\|_1 := \sum_{j=1}^{N} |x_j|$.
- **2-norm:** $\|\mathbf{x}\|_2 := \left(\sum_{j=1}^{N} |x_j|^2 \right)^{1/2}$.
- **∞-norm:** $\|\mathbf{x}\|_\infty := \max_j |x_j|$.

7.3.2 Matrix norms

The norm of a matrix $\mathbf{A}$ is a positive real number giving a measure of the 'size' of the matrix and is denoted by $\|\mathbf{A}\|$ satisfying the following axioms

1. $\|\mathbf{A}\| \geq 0, \quad \forall \mathbf{A}, \qquad \|\mathbf{A}\| = 0 \iff \mathbf{A} = \mathbf{0}$,
2. $\|c\mathbf{A}\| = |c|\|\mathbf{A}\| \quad \forall \mathbf{A}, \;\; c \in \mathbb{F}$,
3. $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| \quad \forall \mathbf{A}, \mathbf{B}$,
4. $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$.

Compatible or consistent norms

Matrix and vector norms are said to be consistent or compatible if

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|, \quad \mathbf{x} \neq \mathbf{0}.$$

7.3.3 Subordinate matrix norm

Let $\mathbf{A}$ be an $n \times n$ matrix and $\mathbf{x}$ be a member of the set $S \subset \mathbb{R}^n$ whose norms are unity, i.e.,

$$S = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}.$$

In general, the norm of the vector $\mathbf{Ax}$ will vary as $\mathbf{x}$ varies, $\mathbf{x} \in S$. Let $\mathbf{x}_0$ be a member of S that makes $\|\mathbf{Ax}\|$ attain its maximum value. Then the norm of the matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\| = \|\mathbf{Ax}_0\| := \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|.$$

This matrix norm is said to be subordinate to the vector norm and automatically satisfies the compatibility condition, because, if $\mathbf{x}_1 \in S$ is any member of S,

$$\|\mathbf{Ax}_1\| \leq \|\mathbf{Ax}_0\| = \|\mathbf{A}\| = \|\mathbf{A}\|\|x_1\|,$$

since $\|\mathbf{x}_1\| = 1$.

The definitions of the 1-, 2- and $\infty-$norm with $\|\mathbf{x}\| = 1$ leads to the following matrix norms

- **1-norm:** $\|\mathbf{A}\|_1 := \max_j \sum_{i=1}^{N} |a_{ij}|$.
- **2-norm:** $\|\mathbf{A}\|_2 := \max_i \sum_{j=1}^{N} |a_{ij}|$.
- **∞-norm:** $\|\mathbf{A}\|_\infty := \sqrt{\varrho(\mathbf{A}^{\mathrm{H}}\mathbf{A})}$, where $\mathbf{A}^{\mathrm{H}}$ denotes the Hermitian of the matrix $\mathbf{A}$ and $\varrho(\mathbf{A})$ denotes the spectral radius of $\mathbf{A}$ defined as follows:

  $$\varrho(\mathbf{A}) := \max_s |\mu_s|$$

  where μ_s are the eigenvalues of the matrix $\mathbf{A}$.

A bound on spectral radius

Let λ be an eigenvalue of an $n \times n$ matrix $\mathbf{A}$ and $\mathbf{x}$ be the corresponding eigenvector. Hence,

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

and

$$|\lambda|\|\mathbf{x}\| = \|\lambda\mathbf{x}\| = \|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|.$$

Therefore,

$$|\lambda| \leq \|\mathbf{A}\|.$$

This inequality holds true for all the eigenvalues, λ, of the matrix $\mathbf{A}$. Hence,

$$\varrho(\mathbf{A}) = \|\mathbf{A}\|.$$

7.4 A necessary and sufficient condition for stability

Here, we discuss the necessary and sufficient condition for stability of a finite difference method with constant coefficients.

Theorem 7.4.1. *The necessary and sufficient condition for the stability for a two-level difference equations corresponding to the IBVP*

$$\begin{aligned} &U_t = U_{xx} \qquad 0 < x < 1,\ 0 < t \leq T \\ &U(0,t) = U_l(t), \quad U(1,t) = U_r(t) \qquad t \geq 0 \\ &U(x,0) = U_0 \qquad 0 \leq x \leq 1. \end{aligned}$$

is

$$\|\mathbf{A}\| \leq 1$$

when the solution of the partial differential equation does not increase as t increases.

Let the solution domain of the partial differential equation be the finite rectangle $0 \leq x \leq 1$, $0 \leq t \leq T$ is subdivided it into uniform rectangular meshes by the lines $x_i = i\Delta x$, $i = 0, 1, \ldots, N$, where $N\Delta x = 1$, and the lines $t_j = j\Delta t$, $j = 0, 1, \ldots, M$, where $M\Delta t = T$.

Proof. Assume that the finite-difference equation relating the mesh-points values at the jth and $(j+1)$th time levels is

$$c_{i-1}u_{i-1}^{j+1} + c_i u_i^{j+1} + c_{i+1}u_{i+1}^{j+1} = d_{i-1}u_{i-1}^{j} + d_i u_i^{j} + d_{i+1}u_{i+1}^{j}$$

for appropriate range of indices $i,\ j$, where the coefficients are constants. Since the boundary values at $i = 0,\ N\ \forall j > 0$, are known, these $N-1$ equations for $i = 1, 2, \ldots, N-1$ can be written in the matrix form

$$\mathbf{B}\mathbf{u}^{(j+1)} = \mathbf{C}\mathbf{u}^{(j)} + \mathbf{d}^{(j)} \qquad j \geq 0,$$

where $\mathbf{B}$ and $\mathbf{C}$ are tri-diagonal matrices having entries b_i's and c_i's, respectively. This may be expressed more conveniently as

$$\mathbf{u}^{(j+1)} = \mathbf{A}\mathbf{u}^{(j)} + \mathbf{f}^{(j)} \qquad j \geq 0,$$

where $\mathbf{A} = \mathbf{B}^{-1}\mathbf{C}$ and $\mathbf{f}^{(j)} = \mathbf{B}^{-1}\mathbf{d}^{(j)}$. Using the above relative recursively we obtain

$$\begin{aligned}
\mathbf{u}^{(j)} &= \mathbf{A}\mathbf{u}^{(j-1)} + \mathbf{f}^{(j-1)} = \mathbf{A}\left(\mathbf{A}\mathbf{u}^{(j-2)} + \mathbf{f}^{(j-2)}\right) + \mathbf{f}^{(j-1)} \\
&= \mathbf{A}^2\mathbf{u}^{(j-2)} + \mathbf{A}\mathbf{f}^{(j-2)} + \mathbf{f}^{(j-1)} \\
&= \mathbf{A}^3\mathbf{u}^{(j-3)} + \mathbf{A}^2\mathbf{f}^{(j-3)} + \mathbf{A}\mathbf{f}^{(j-2)} + \mathbf{f}^{(j-1)} \\
&= \cdots \\
&= \mathbf{A}^j\mathbf{u}^{(0)} + \mathbf{A}^{j-1}\mathbf{f}^{(0)} + \mathbf{A}^{(j-2)}\mathbf{f}^{(1)} + \ldots + \mathbf{f}^{(j-1)}, \qquad (7.7)
\end{aligned}$$

where $\mathbf{u}^{(0)}$ is the vector of initial values and $\mathbf{f}^{(0)},\ \mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(j-1)}$ are vectors of known boundary values. When we are concerned more about with a property of the equations, such as stability, than with a numerical solution, the constant vectors can be eliminated by investigating the propagation of a perturbation.

Assume that the perturbed initial values is denoted by $\mathbf{u}_*^{(0)}$. Corresponding to the perturbed initial values, the exact solution of the difference equations will be

$$\mathbf{u}_*^{(j)} = \mathbf{A}^j\mathbf{u}_*^{(0)} + \mathbf{A}^{j-1}\mathbf{f}^{(0)} + \mathbf{A}^{(j-2)}\mathbf{f}^{(1)} + \ldots + \mathbf{f}^{(j-1)}. \qquad (7.8)$$

Defining the error vector $\mathbf{e} = \mathbf{u}_* - \mathbf{u}$, it follows from (7.7) and (7.8) that

$$\mathbf{e}^{(j)} = \mathbf{A}^j\mathbf{e}^{(0)} \qquad j \geq 1.$$

This equation determines the propagation of an error $\mathbf{e}^{(0)}$.

Hence, for compatible matrix and vector norms,

$$\|\mathbf{e}^{(j)}\| \leq \|\mathbf{A}^j\|\|\mathbf{e}^{(0)}\|.$$

Lax and Richtmyer define the difference scheme to be stable when $\exists\ M > 0$ independent of $j, \Delta x, \Delta t$, such that

$$\|\mathbf{A}^j\| \leq M \qquad j \geq 1.$$

This clearly limits the amplification of any initial perturbation, and therefore for any initial rounding errors, because

$$\|\mathbf{e}^{(j)}\| \leq M\|\mathbf{e}^{(0)}\|.$$

Since

$$\|\mathbf{A}^j\| = \|\mathbf{A}\mathbf{A}^{(j-1)}\| \leq \|\mathbf{A}\|\|\mathbf{A}^{(j-1)}\| \leq \cdots \leq \|\mathbf{A}\|^j \qquad j \geq 1,$$

it follows that the Lax-Richtmyer defintion of stability is satisfied by

$$\|\mathbf{A}\| \leq 1.$$

□

Whenever this condition is satisfied it follows automatically that the spectral radius $\varrho(\mathbf{A}) \leq 1$ since $\varrho(\mathbf{A}) \leq \|\mathbf{A}\|$. However, the converse is not true.

Corollary 7.4.1.1 (Stability criteria for derivative boundary conditions)**.** *Consider the equation*

$$U_t = U_{xx}, \qquad 0 < x < 1$$

and the conditions,

$$\begin{aligned} U_x &= h_1(U - v_1) \quad at \quad x = 0, \quad t \geq 0, \\ U_x &= h_2(U - v_2) \quad at \quad x = 1, \quad t \geq 0, \end{aligned}$$

where h_1, h_2, v_1, v_2 are constants, $h_1 \geq 0$, $h_2 \geq 0$. For overall stability of FTCS (explicit scheme), we require

$$r \leq \min\left\{\frac{1}{2 + h_1\Delta x}, \frac{1}{2 + h_2\Delta x}\right\}.$$

Proof. See Smith [1985]. □

Theorem 7.4.2. *FTCS method is conditionally stable and the stability condition is*

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}.$$

Proof. The proof was discussed in the lecture using both the matrix stability and von-Neumann stability analyses.

Also, see Smith [1985]. □

Exercise 1. *Show that BTCS method is unconditionally stable.*

Exercise 2. *Show that Crank-Nicolson method is unconditionally stable.*

Exercise 3. *Discuss the stability of the Richardson method* (7.2)*.*

Theorem 7.4.3 (Gerschgorin's first theorem)**.** *The largest of the moduli of the eignevalues of the square matrix* $\mathbf{A}$ *cannot exceed the largest sum of the moduli of the elements along any row or any column. In other words,*

$$\varrho(\mathbf{A}) \leq \|\mathbf{A}\|_1 \quad or \quad \|\mathbf{A}\|_\infty.$$

Proof. See Smith [1985]. □

Theorem 7.4.4 (Gerschgorin's circl theorem or Brauer's theorem)**.** *Let P_s be the sum of moduli of the elements along the sth row excluding the diagonal element $a_{s,s}$. Then each eigenvalue of* **A** *lies inside or on the boundary of at least one of the circles $|\lambda - a_{s,s}| = P_s$.*

Proof. See Smith [1985]. □

Exercise 4. *Discuss the consistency and stability of the linear difference equation*

$$\frac{1}{\Delta t}\left(u_p^{q+1} - u_p^q\right) = \frac{a}{\Delta x^2}\left(u_{p+1}^q - 2u_p^q + u_{p-1}^q\right) + bu_p^q$$

approximating the parabolic equation

$$U_t = aU_{xx} + bU$$

at the point $(p\Delta x, q\Delta t)$, where a and b are positive constants.

Theorem 7.4.5 (Lax equivalence theorem)**.** *Given a well-posed linear IBVP and a linear finite-difference approximation to it that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence.*

7.5 Glabal rounding error

For simplicity, assume that all boundary values are zero (i.e., homogeneous Dirichlet boundary conditions) so that the finite difference equations approximating the IBVP in the solution domain $0 < x < 1$, $t > 0$, can be written as

$$\mathbf{u}^{(j)} = \mathbf{A}\mathbf{u}^{(j-1)} \quad j \geq 1$$

where $\mathbf{u}^{(0)} = \mathbf{U}^{(0)}$ is the vector of known initial values and let **A** is the square matrix of known elements of order $N - 1$.

In general, computer will not store the initial value $u_i^{(0)}$ exactly, but a numerical approximation $N_i^{(0)}$, so that

$$N_i^{(0)} = u_i^{(0)} - r_i^{(0)} \qquad \text{or} \qquad \mathbf{N}^{(0)} = \mathbf{u}^{(0)} - \mathbf{r}^{(0)},$$

where $\mathbf{r}^{(0)}$ is the vector of initial rounding errors. As rounding errors will be introduced at every stage of the calculations the numerical solution values calculated by computer at the first time-level will be

$$\mathbf{N}^{(1)} = \mathbf{A}\mathbf{N}^{(0)} - \mathbf{r}^{(0)} = \mathbf{A}\mathbf{u}^{(0)} - \mathbf{A}\mathbf{r}^{(0)} - \mathbf{r}^{(1)}.$$

Finally, at the jth time-level, the computed solution will be

$$\mathbf{N}^{(j)} = \mathbf{A}^j\mathbf{u}^{(0)} - \mathbf{A}^j\mathbf{r}^{(0)} - \mathbf{A}^{j-1}\mathbf{r}^{(1)} - \cdots - \mathbf{r}^{(j)}.$$

Without any rounding errors the exact solution of the difference equation would be

$$\mathbf{u}^{(j)} = \mathbf{A}^j \mathbf{u}^{(0)}.$$

Therefore, the difference between the exact solution and the computed solution, i.e., the global rounding error R_j, at the jth time-level is

$$\mathbf{u}^{(j)} - \mathbf{N}^{(j)} = \mathbf{A}^j \mathbf{r}^{(0)} + \mathbf{A}^{j-1}\mathbf{r}^{(1)} + \cdots + \mathbf{r}^{(j)}.$$

Thus, the local rounding error vector at each time-level propagates forward in the same way as the exact solution vector at that time-level. The effects of each local rounding error will diminish with increasing j if $\varrho(\mathbf{A}) < 1$, but the global rounding error cannot possibly tend to zero because of the terms $\mathbf{r}^{(j)}, \mathbf{A}\mathbf{r}^{(j-1)}, \ldots$.

7.6 Finite difference approximation of parabolic equations in two space-dimensions

7.6.1 The θ-scheme

By taking the convex combination of the explicit and implicit Euler schemes, with a parameter $\theta \in [0,1]$, with $\theta = 0$ corresponding to the explicit Euler scheme and $\theta = 1$ to the implicit Euler scheme, we obtain a one-parameter family of schemes, called the θ-scheme. It is defined as follows.

Let $\Delta x := (b-a)/N_x$, $\Delta y := (d-c)/N_y$, $\Delta t := T/M$, and, for $\theta \in [0,1]$, consider the finite difference scheme

$$\frac{u_{i,j}^{m+1} - u_{i,j}^{m}}{\Delta t} = (1-\theta)\left(\frac{\delta_x^2}{\Delta x^2} + \frac{\delta_y^2}{\Delta y^2}\right) u_{i,j}^{m} + \theta\left(\frac{\delta_x^2}{\Delta x^2} + \frac{\delta_y^2}{\Delta y^2}\right) u_{i,j}^{m+1},$$

for $i = 1, 2, \ldots, N_x - 1$, $j = 1, 2, \ldots, N_y - 1$, $m = 0, 1, \ldots, M-1$, subject to the initial condition

$$u_{i,j}^{0} := U_0(x_i, y_j), \qquad i = 1, 2, \ldots, N_x - 1, \;\; j = 1, 2, \ldots, N_y - 1,$$

and boundary conditions

$$u_{i,j}^{m+1} := B(x_i, y_j, t_{m+1}) \text{ at the boundary mesh poitns, for } \; m = 0, 1, \ldots, M-1.$$

7.6.2 The alternating direction implicit (ADI) method LeVeque [2007]

Except for $\theta = 0$ corresponding to the explicit Euler scheme, for all other values of $\theta \in (0,1]$ the θ-scheme is an implicit scheme, and its implementation therefore involves the solution of a large system of linear algebraic equations at each time level. This is true, in particular, in the case of the Crank–Nicolson scheme corresponding to $\theta = 1/2$. Our objective here is to propose a more economical scheme, which replaces the tedious task of solving such large systems of algebraic equations with the successive solution of smaller linear systems in the x

and y coordinate directions respectively, alternating between solves in the x and y coordinate directions. The resulting finite difference scheme is called the alternating direction (or ADI) scheme. We describe its construction starting from the Crank-Nicolson scheme, which has the form:

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2 - \frac{1}{2}\mu_y\delta_y^2\right) u_{i,j}^{m+1} = \left(1 + \frac{1}{2}\mu_x\delta_x^2 + \frac{1}{2}\mu_y\delta_y^2\right) u_{i,j}^{m}, \tag{7.9}$$

for $i = 1, 2, \ldots, N_x - 1$, $j = 1, 2, \ldots, N_y - 1$, $m = 1, 2, \ldots, M - 1$, subject to the initial condition

$$U_{i,j}^0 = u_0(x_i, y_j), \quad i = 1, 2, \ldots, N_x - 1, \;\; j = 1, 2, \ldots, N_y - 1,$$

and the boundary conditions

$$u_{i,j}^m := B(x_i, y_j, t_m), \quad \text{at the boundary mesh-points, for } m = 1, 2, \ldots, M.$$

Let us modify this scheme (subject to the same initial and boundary conditions) to:

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2\right)\left(1 - \frac{1}{2}\mu_y\delta_y^2\right) u_{i,j}^{m+1} = \left(1 + \frac{1}{2}\mu_x\delta_x^2\right)\left(1 + \frac{1}{2}\mu_y\delta_y^2\right) u_{i,j}^{m}. \tag{7.10}$$

As the expressions on the left-hand side and the right-hand side of (7.10) differ from those in the Crank-Nicolson scheme (7.9) above, the numerical solution computed from (7.10) will also differ from the one obtained from the Crank-Nicolson scheme (7.9). It can be shown however that the consistency error of (7.10) is still $\mathbf{O}(\Delta x^2 + \Delta y^2 + \Delta t^2)$ as in the case of the Crank-Nicolson scheme; there is therefore no significant loss of accuracy resulting from the replacement of (7.9) with (7.10). The benefits of replacing (7.9) with (7.10) will be made clear below.

By introducing the intermediate level $u^{m+1/2}$, we can rewrite the last equality in the following equivalent form

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2\right) u_{i,j}^{m+1/2} = \left(1 + \frac{1}{2}\mu_y\delta_y^2\right) u_{i,j}^{m}, \tag{7.11}$$

$$\left(1 - \frac{1}{2}\mu_y\delta_x^y\right) u_{i,j}^{m+1} = \left(1 + \frac{1}{2}\mu_x\delta_x^2\right) u_{i,j}^{m+1/2}. \tag{7.12}$$

The equivalence of the system (7.11), (7.12) to the scheme (7.10) is seen by applying the finite difference operator $\left(1 + \frac{1}{2}\mu_x\delta_x^2\right)$ to equation (7.11) and the finite difference operator $\left(1 - \frac{1}{2}\mu_x\delta_x^2\right)$ to equation (7.12), and noting that these two finite difference operators commute. Given $u_{i,j}^m$, the equation (7.11) amounts to solving, for each j, a one-dimensional problem in the x-direction to compute $u_{i,j}^{m+1/2}$; and then, by using the computed values $u_{i,j}^{m+1/2}$ one solves, for each i, a one-dimensional problem in the y-direction using (7.12) to determine $u_{i,j}^{m+1}$. Thus, by starting from the information at time level $m = 0$, where $u_{i,j}^0$ is specified by the initial datum, one proceeds from time level m to time level $m+1$, alternating successively between the x and y directions and advancing from time level m to time level $m + 1$ for $m = 0, 1, \ldots, M - 1$.

7.7 Hyperbolic PDEs

Consider the equation

$$a(x,y,U)\frac{\partial U}{\partial x}+b(x,y,U)\frac{\partial U}{\partial y}=c(x,y,U). \tag{7.13}$$

Such an equation is said to be *quasi-linear*. Assume we know the solution values U of (7.13) at every point on a curve C in the $x-y$ plane, where C does not coincide with the curve Γ on which initial values of U are specified. The question to be asked at this stage is: Can we determine values for U_x and U_y on C from the values of U on C so that they satisfy (7.13)?

If we can, then in directions tangential to C from points on C we shall automatically satisfy the differential relationship

$$\mathrm{d}U=\frac{\partial U}{\partial x}\mathrm{d}x+\frac{\partial U}{\partial y}\mathrm{d}y, \tag{7.14}$$

where $\mathrm{d}y/\mathrm{d}x$ is the slope of the tangent to C at $P(x,y)$ on C. Combining equations (7.13) and (7.14) we can write

$$\frac{\partial U}{\partial y}(a\mathrm{d}y-b\mathrm{d}x)+(c\mathrm{d}x-a\mathrm{d}U)=0. \tag{7.15}$$

This can be made independent of U_y by choosing the curve C so that its slope $\mathrm{d}y/\mathrm{d}x$ satisfies the equation

$$a\mathrm{d}y-b\mathrm{d}x=0 \tag{7.16}$$

representing a differential equations for the curve C – called the characteristic curves or characteristics. Thus, a differential equation for the solution values of U along C is

$$c\mathrm{d}x-a\mathrm{d}U=0. \tag{7.17}$$

Equations (7.16) and (7.17) can be combined to write

$$\frac{\mathrm{d}x}{a(x,y,U)}=\frac{\mathrm{d}y}{b(x,y,U)}=\frac{\mathrm{d}U}{c(x,y,U)}. \tag{7.18}$$

This also shows that U may be found from either the equation

$$\mathrm{d}U=\frac{c(x,y,U)}{a(x,y,U)}\mathrm{d}x$$

or the equation

$$\mathrm{d}U=\frac{c(x,y,U)}{b(x,y,U)}\mathrm{d}y.$$

Example 7.7.1. *Consider the equation*

$$y\frac{\partial U}{\partial x}+\frac{\partial U}{\partial y}=2 \tag{7.19}$$

where U is known along the initial segment Γ defined by $y = 0,\ \ 0 \leq x \leq 1$.

The family of characteristic curves is given by

$$\frac{\mathrm{d}x}{y} = \frac{\mathrm{d}y}{1} \tag{7.20}$$

that gives the following family $x = y^2/2 + A$, where the parameter A is a constant for each characteristic. For the characteristic through $(x_R, 0)$, $A = x_R$, so the equation of this particular characteristic is $y^2 = 2(x - x_R)$.

The solution along a characteristic curve is given by

$$\frac{\mathrm{d}y}{1} = \frac{\mathrm{d}U}{2} \tag{7.21}$$

which integrates to $U = 2y + B$, where B is constant along a particular characteristic. If $U = U_R$ at $(x_R, 0))$ then $B = U_R$ and the solution along the characteristic $y^2 = 2(x - x_R)$ is $U = 2y + U_R$.

Exercise 5. *Find the U that satisfies*

$$\sqrt{x}\frac{\partial U}{\partial x} + U\frac{\partial U}{\partial y} = -U^2 \tag{7.22}$$

and the condition $U = 1$ on $y = 0,\ \ 0 < x < \infty$.

7.8 Numerical solution of first order hyperbolic equation

Consider the linear advection equation

$$U_t + a_0 U_x = 0 \qquad x \in \mathbb{R},\ \ t > 0, \tag{7.23}$$

where a_0 is a non-zero constant, along with the initial data

$$U(x, 0) = f(x) \qquad x \in \mathbb{R}. \tag{7.24}$$

The function $U = f(x - a_0 t)$ is a solution to this Cauchy problem and represent a travelling wave. If $a_0 > 0$ the wave propagates in the positive x-direction, while the wave propagates in the negative x-direction if $a_0 < 0$. Equation (7.23) can be expressed as

$$U_t + F_x = 0 \qquad x \in \mathbb{R},\ \ t > 0, \tag{7.25}$$

for $F = a_0 U$.

We first try to solve problem (7.23) and (7.24) using forward time centered space finite difference scheme. This gives rise to the following difference equation

$$u_p^{q+1} = u_p^q - \frac{\mathcal{C}}{2}\left(u_{p+1}^q - u_{p-1}^q\right), \tag{7.26}$$

where $\mathcal{C} = a_0 \dfrac{\Delta t}{\Delta x}$ is called the Courant-Friedrich-Lewy (CFL) number Courant et al. [1928]. It is easy show that the method (7.26) is consistent to (7.23) and LTE $= \mathbf{O}(\Delta t, \Delta x^2)$. However, this method is unconditionally unstable for any value of the CFL number, as

$$|\xi| > 1, \quad \forall \mathcal{C}.$$

Therefore, this method is not viable and we seek alternative numerical scheme to solve hyperbolic equation.

We ask the question: **What causes the FTCS method to he unconditionally unstable?**

To address this question, we analyze the finite difference scheme

$$\frac{u_p^{q+1} - u_p^q}{2\Delta t} + a_0 \frac{u_{p+1}^q - u_{p-1}^q}{2\Delta t} = 0. \tag{7.27}$$

more closely. Taylor series expansion about $(p\Delta x, q\Delta t)$ points yields

$$\begin{aligned}
&\frac{1}{\Delta t}\left[\left\{u_p^q + \Delta t\left(\frac{\partial u}{\partial t}\right)_{p,q} + \frac{\Delta t^2}{2}\left(\frac{\partial^2 u}{\partial t^2}\right)_{p,q} + \cdots\right\} - u_p^q\right] \\
&+ \quad \frac{a_0}{2\Delta x}\left[\left\{u_p^q + \Delta x\left(\frac{\partial u}{\partial x}\right)_{p,q} + \frac{\Delta x^2}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{p,q} + \frac{\Delta x^2}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{p,q} + \frac{\Delta x^4}{24}\left(\frac{\partial^4 u}{\partial x^4}\right)_{p,q} + \cdots\right\}\right. \\
&\qquad \left. - \left\{u_p^q - \Delta x\left(\frac{\partial u}{\partial x}\right)_{p,q} + \frac{\Delta x^2}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{p,q} - \frac{\Delta x^3}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{p,q} + \frac{\Delta x^4}{24}\left(\frac{\partial^4 u}{\partial x^4}\right)_{p,q} + \cdots\right\}\right] = 0 \\
\Rightarrow \quad &\left(\frac{\partial u}{\partial t}\right)_{p,q} + a_0\left(\frac{\partial u}{\partial x}\right)_{p,q} = -\frac{\Delta t}{2}\left(\frac{\partial^2 u}{\partial t^2}\right)_{p,q} - a_0\frac{\Delta x^2}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{p,q} + \cdots
\end{aligned} \tag{7.28}$$

Using (7.23) we can write

$$\frac{\partial^2 U}{\partial t^2} = a_0^2 \frac{\partial^2 U}{\partial x^2}.$$

Thus, from equation (7.28) we obtain

$$\begin{aligned}
\left(\frac{\partial u}{\partial t}\right)_{p,q} + a_0\left(\frac{\partial u}{\partial x}\right)_{p,q} &= -\frac{\Delta t}{2}a_0^2\left(\frac{\partial^2 u}{\partial x^2}\right)_{p,q} - a_0\frac{\Delta x^2}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{p,q} + \cdots \\
&= -a_0\mathcal{C}\frac{\Delta x}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{p,q} - a_0\frac{\Delta x^2}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{p,q} + \cdots
\end{aligned} \tag{7.29}$$

This clearly shows that the finite difference scheme actually is unstable. In computational fluid mechanics the term $\nu_N = -(a_0\Delta x/2)\mathcal{C}$ is often called numerical viscosity in analogy to the physical viscosity that accompanies the diffusive term of advection-diffusion equations like $\dfrac{\partial U}{\partial t} + a_0\dfrac{\partial U}{\partial x} = \nu\dfrac{\partial^2 U}{\partial x^2}$. When using the scheme given in (7.27), this results in solving the advection-diffusion equation with a negative viscosity coefficient, which is certainly unstable. Equation (7.29) provides information on why oscillations decrease for decreasing $\mathcal{C}$. Equation (7.29) is called the modified equation for a given numerical scheme, since it shows which is the differential problem that the scheme is actually solving and it therefore provides a deep insight in the behavior that the numerical solution will exhibit.

7.9 Upwind scheme

Analysing equation (7.23) we observe that the initial data propagates along the x-axis at a speed a_0. Our next method is motivated by this fact and we forward (backward) difference approximation for the first order derivative U_x when $a_0 < 0$ ($a_0 > 0$). For $a_0 > 0$ the diffeernce schemes reads as

$$u_p^{q+1} = (1 - \mathcal{C})u_p^q + \mathcal{C}u_{p-1}^q. \tag{7.30}$$

Again, it is easy to prove that this scheme is consistent and LTE $= \mathbf{O}(\Delta t, \Delta x)$. von Neumann stability analysis yields that the method is conditionally stable and the condition of stability is $0 < \mathcal{C} \leq 1$. In general, we can define the CFL number as

$$\mathcal{C} := |a_0| \frac{\Delta t}{\Delta x}, \tag{7.31}$$

and the upwind method is stable for $0 < \mathcal{C} \leq 1$. This condition of stability is called *CFL condition*.

Similar to the FTCS scheme, we can derive the modified equation for the upwind scheme (7.30) that reads as

$$\left(\frac{\partial u}{\partial t}\right)_{p,q} + a_0 \left(\frac{\partial u}{\partial x}\right)_{p,q} = a_0(1 - \mathcal{C})\frac{\Delta x}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{p,q} + \cdots \tag{7.32}$$

In this case the numerical viscosity $\nu_N = a_0(1 - \mathcal{C})\dfrac{\Delta x}{2}$ is non-negative when the CFL condition is satisfied. This results a stability of the upwind scheme.

7.9.1 Geometric and physical interpretations of CFL condition

The family of characteristics of (7.23) is given by

$$\frac{\mathrm{d}t}{1} = \frac{\mathrm{d}x}{a_0}.$$

Therefore, the slope of the characteristic curves at any point in the $x - t$ plane is given by $\dfrac{\mathrm{d}t}{\mathrm{d}x} = \dfrac{1}{a_0}$. For stability of the upwind scheme we require the slope of $PD \geq$ slope of PA. This implies that $a_0 \dfrac{\Delta t}{\Delta x} \leq 1$, i.e., $\mathcal{C} \leq 1$.

The CFL condition can be also interpreted as follows. We know that a_0 is the velocity of propagation of the wave along the x-direction. On the other hand, $\dfrac{\Delta x}{\Delta t}$ corresponds to the numerical velocity of propagation of the wave along the x-axis. For the stability of the numerical scheme we require the numerical velocity of wave propagation is greater than or equal to the actual velocity of wave propagation. This yields, $a_0 \leq \dfrac{\Delta x}{\Delta t}$, i.e., $\mathcal{C} \leq 1$.

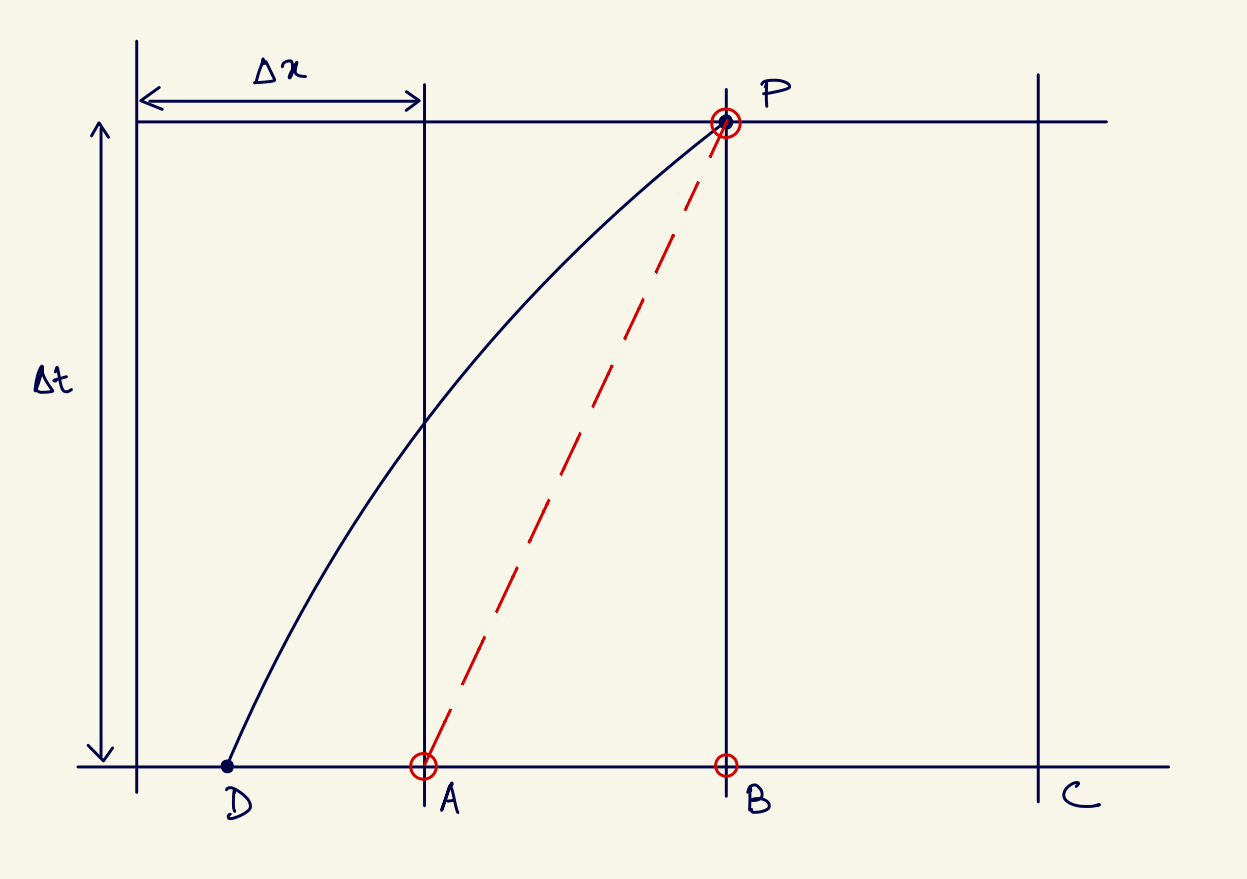

Figure 7.1: Geometric interpretation of CFL condition. PD represents a schematic representation of characteristic curves. The red circle points represents the points used in upwind scheme. Information at the point P at the $(j+1)$th time level depends on the information on AB at the jth time level. Whereas, the analytical solution at the point P at the $(j+1)$th time level depends on the point D too. This is is outside the region of influence of numerical solution.

7.10 Lax-Wendroff methods

These schemes were proposed in 1960 by P.D. Lax and B. Wendroff [Lax, 1973] for solving, approximately, systems of hyperbolic conservation laws on the generic form given in (7.25). A large class of numerical methods for solving (7.25) are the so-called conservative methods:

$$u_i^{j+1} = u_i^j + \frac{\Delta t}{\Delta x}\left(F_{i+1/2} - F_{i-1/2}\right). \tag{7.33}$$

where the form of the numerical fluxes $F_{i+1/2}$ and $F_{i-1/2}$ determines whether the scheme is explicit or implicit.

7.10.1 Lax-Wendroff explicit scheme

The Lax–Wendroff method belongs to the class of conservative schemes (7.25) and can be derived in various ways. For simplicity, we will derive the method by using a simple model equation for (7.25), namely the linear advection equation with $F(U) = a_0 U$ as in (7.23), where a_0 is a constant propagation velocity. The Lax-Wendroff outset is a Taylor approximation of u_i^{j+1}:

$$u_i^{j+1} = u_i^j + \Delta t \left(\frac{\partial u}{\partial t}\right)_{i,j} + \frac{\Delta t^2}{2}\left(\frac{\partial^2 u}{\partial t^2}\right)_{i,j} + \cdots \tag{7.34}$$

From the PDE (7.23) we have

$$\left(\frac{\partial}{\partial t}\right)_{i,j} = -a_0\left(\frac{\partial}{\partial x}\right)_{i,j} \qquad \left(\frac{\partial^2}{\partial t^2}\right)_{i,j} = a_0^2\left(\frac{\partial^2}{\partial x^2}\right)_{i,j}. \tag{7.35}$$

Using (7.35) and

$$\left(\frac{\partial u}{\partial x}\right)_{i,j} = \frac{u_{i+1}^j - u_{i-1}^j}{2\Delta x} \qquad \left(\frac{\partial^2 u}{\partial t^2}\right)_{i,j} = \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{\Delta x^2}$$

in (7.34) we obtain

$$u_i^{j+1} = \frac{1}{2}\mathcal{C}(1+\mathcal{C})u_{i-1}^1 + (1-\mathcal{C}^2)u_i^j - \frac{1}{2}\mathcal{C}(1-\mathcal{C})u_{i+1}^j. \tag{7.36}$$

This scheme is called Lax-Wendroff explicit scheme and the local trunction error of this method if

$$\text{LTE} = \frac{1}{6}\Delta t^2\left(\frac{\partial^3 U}{\partial t^3}\right)_{i,j} + \frac{1}{6}a_0\Delta x^2\left(\frac{\partial^3 U}{\partial x^3}\right)_{i,j} + \cdots \tag{7.37}$$

von Neumann stability analysis yields that this method is (conditionally) stable if $0 < \mathcal{C} \leq 1$.

7.10.2 Lax-Wendroff implicit scheme

Here, we give an outline for the derivation Lax-Wendroff implicit scheme for a linear hyperbolic equation of the form

$$a\frac{\partial U}{\partial t} + b\frac{\partial U}{\partial x} = c, \tag{7.38}$$

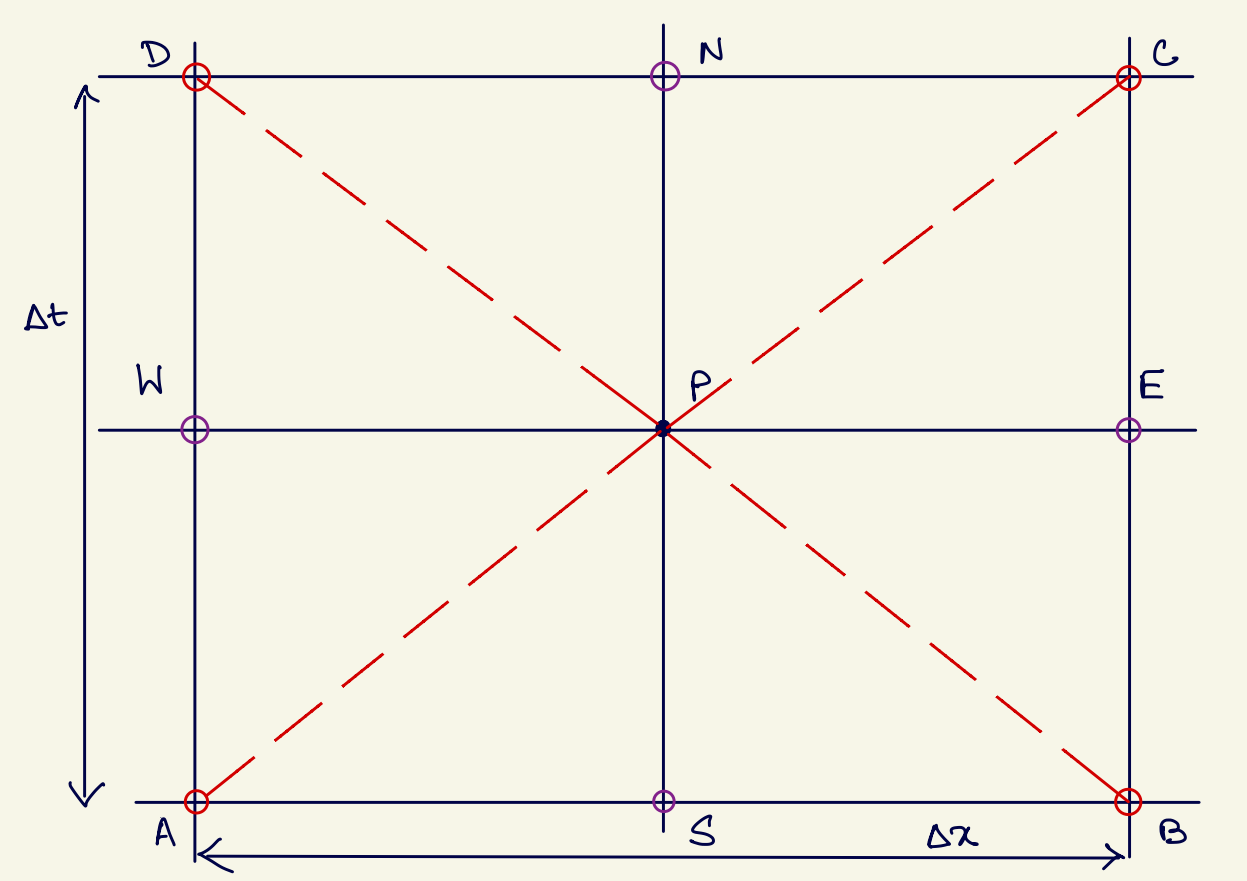

Figure 7.2: Schematic representation of approximating (7.38) at point P using adjacent points to obtain implicit Lax-Wendroff scheme.

where $a,\ b,\ c$ are constants such that a and b are non-zero.

We approximate (7.38) at the point P (see Fig. 7.2) as follows

$$\frac{a}{2}\left[\left(\frac{\partial U}{\partial t}\right)_E + \left(\frac{\partial U}{\partial t}\right)_W\right] + \frac{b}{2}\left[\left(\frac{\partial U}{\partial x}\right)_N + \left(\frac{\partial U}{\partial x}\right)_S\right] = c. \tag{7.39}$$

Further, approximate $\left(\frac{\partial U}{\partial t}\right)_E$, $\left(\frac{\partial U}{\partial t}\right)_W$, $\left(\frac{\partial U}{\partial x}\right)_N$ and $\left(\frac{\partial U}{\partial x}\right)_S$ using second-order central difference formulas to obtain an *unconditionally unstable* implicit scheme that has is second-order convergence both in space and time as $(\Delta t, \Delta x) \to (0, 0)$.

References

E Ward Cheney and David R Kincaid. *Numerical mathematics and computing*. Cengage Learning, 2012.

Richard Courant, Kurt Friedrichs, and Hans Lewy. Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische annalen*, 100(1):32–74, 1928.

Peter D Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*. SIAM, 1973.

Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

Robert D Richtmyer and Keith W Morton. Difference methods for initial-value problems. *Malabar*, 1994.

Gordon D Smith. *Numerical solution of partial differential equations: finite difference methods*. Oxford University Press, 1985.