

Numerical Analysis: Foundations and Numerical Partial Differential Equations

Jeongbin Jo

Department of Physics, Yonsei University, jeongbin033@yonsei.ac.kr

School of Mathematics and Computing (Computational Science and Engineering), Yonsei
University

September 21, 2026

Abstract

These notes follow the Numerical Analysis lectures based on Atkinson, covering preliminaries, rootfinding, interpolation, approximation, quadrature, ODE initial-value problems, linear algebra, linear systems, and eigenvalue computations. The second half develops finite-difference, finite-volume, finite-element, and discontinuous Galerkin methods for PDEs, using the lecture material as the analytic foundation.

Contents

I	Foundations	3
1	Preliminaries	3
1.1	Notation	4
1.2	Mathematical Preliminaries	4
1.3	Sources of Error	5
2	Rootfinding for Nonlinear Equations	5
2.1	The Bisection Method	5
2.2	Newton's Method	6
2.3	The Secant Method	6
2.4	Fixed-Point Iteration	6
2.5	Aitken Extrapolation	7
2.6	Systems of Nonlinear Equations	7
2.7	Unconstrained Optimization	7
3	Interpolation Theory	8
3.1	Polynomial Interpolation and the Vandermonde System	8
3.2	Newton Divided Differences	9
3.3	Hermite Interpolation	10
3.4	Piecewise Polynomials and Splines	10

4	Approximation of Functions	11
4.1	The Weierstrass Theorem and Taylor Expansion	11
4.2	The Least Squares Problem	12
4.3	Inner Products and Orthogonal Polynomials	12
4.4	Least Squares in an Orthogonal Basis	13
5	Numerical Integration	14
5.1	The Trapezoidal Rule	15
5.2	Simpson's Rule	15
5.3	Newton–Cotes Formulas	15
5.4	Gaussian Quadrature	16
5.5	Singular Integrals	16
II	Ordinary Differential Equations	16
6	Numerical Methods for Ordinary Differential Equations	16
6.1	Existence, Uniqueness, and Stability	17
6.2	Euler's Method	17
6.3	Linear Multistep Methods	18
6.4	One-Step Methods	19
6.5	Adams Methods and Comparison	19
6.6	Root Condition and Absolute Stability	19
III	Linear Algebra	20
7	Linear Algebra	20
7.1	Vector Spaces, Matrices, and Linear Systems	20
7.2	Eigenvalues and Canonical Forms	20
7.3	Vector and Matrix Norms	21
7.4	Convergence and Perturbation Theorems	21
8	Numerical Solution of Systems of Linear Equations	21
8.1	Gaussian Elimination and LU	22
8.2	Pivoting	22
8.3	Variants	22
8.4	Error Analysis	22
8.5	Residual Correction	22
8.6	Stationary Iterative Methods	23
8.7	Gradient and Conjugate Gradient Methods	23
8.8	Poisson's Equation	23
9	The Matrix Eigenvalue Problem	23
9.1	Location, Error, and Stability	24
9.2	The Power Method	24
9.3	Householder Transformations	24
9.4	The QR Method	25
9.5	Least Squares and the SVD	25

IV	Numerical Partial Differential Equations	25
10	Finite Differences and Elliptic Boundary Value Problems	25
10.1	Basic Difference Operators	25
10.2	Truncation Error	25
10.3	Undetermined Coefficients	26
10.4	Second Derivatives and the Laplacian	26
10.5	Heat Equation and Boundary Conditions	26
10.6	The Discrete Poisson Problem	26
10.7	Neumann Conditions	27
10.8	Uniqueness and Singularity	27
11	Diffusion Equations and Parabolic Problems	27
11.1	The Heat Equation	27
11.2	Crank–Nicolson	27
11.3	Von Neumann Stability	28
12	Advection Equations and Hyperbolic Systems	28
12.1	The Linear Advection Equation	28
12.2	Classical Finite Difference Schemes	28
13	Finite Volume Method and Conservation Laws	29
13.1	Integral Formulation and Numerical Fluxes	29
13.2	High-Resolution Schemes and Flux Limiters	29
14	The Finite Element Method	30
14.1	Weak Derivatives and Sobolev Spaces	30
14.2	Variational Formulation	30
14.3	Lax–Milgram and Energy Minimization	30
14.4	Galerkin Approximation	31
15	Discontinuous Galerkin Methods	31
15.1	Motivation	31
15.2	Formulation and Numerical Fluxes	31
15.3	Comparison of Discretizations	31

Part I

Foundations

1 Preliminaries

These notes follow the course lectures based on Atkinson [1], and later chapters add finite-difference, finite-volume, and finite-element methods for PDEs [10, 9].

1.1 Notation

Throughout, h denotes a spatial mesh size (or, in Chapter 6 only, the ODE time step when no spatial grid is present). When space and time appear together we write $h = \Delta x$ and $k = \Delta t$. Exact solutions are lowercase (u, y, f); grid functions are uppercase (U_j^n, Y_n).

1.2 Mathematical Preliminaries

Theorem 1.1 (Intermediate value). *Let f be continuous on the finite interval $[a, b]$, and set*

$$m = \inf_{a \leq x \leq b} f(x), \quad M = \sup_{a \leq x \leq b} f(x).$$

Then for every $c \in [m, M]$ there exists at least one $x_0 \in [a, b]$ with $f(x_0) = c$.

Theorem 1.2 (Mean value). *Let f be continuous on $[a, b]$ and differentiable on (a, b) . Then there exists $\xi \in (a, b)$ such that*

$$f(b) - f(a) = f'(\xi)(b - a).$$

Theorem 1.3 (Integral mean value). *Let w be nonnegative and integrable on $[a, b]$, and let f be continuous on $[a, b]$. Then*

$$\int_a^b w(x)f(x) dx = f(\xi) \int_a^b w(x) dx$$

for some $\xi \in [a, b]$.

Theorem 1.4 (Taylor). *Let $f \in C^{n+1}([a, b])$ and $x, x_0 \in [a, b]$. Then*

$$f(x) = \sum_{k=0}^n \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) + R(x), \quad R(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi)$$

for some ξ between x_0 and x .

Example 1.5. The expansion of e^x about $x_0 = 0$ is $e^x = \sum_{k=0}^n x^k/k! + e^\xi x^{n+1}/(n+1)!$.

Theorem 1.6 (Taylor in two variables). *Let $f \in C^{n+1}(D)$ with $D \subset \mathbb{R}^2$, and write $\xi = x - x_0$, $\eta = y - y_0$. Then*

$$f(x, y) = f(x_0, y_0) + \sum_{m=1}^n \frac{1}{m!} (\xi \partial_x + \eta \partial_y)^m f(x_0, y_0) + \frac{1}{(n+1)!} (\xi \partial_x + \eta \partial_y)^{n+1} f(x_0 + \theta \xi, y_0 + \theta \eta)$$

for some $\theta \in [0, 1]$.

Definition 1.7 (Norm). A map $\|\cdot\|$ is a norm if

- (1) $\|u\| \geq 0$, and $\|u\| = 0$ if and only if $u = 0$;
- (2) $\|\alpha u\| = |\alpha| \|u\|$ for all scalars α ;
- (3) $\|u + w\| \leq \|u\| + \|w\|$.

Standard examples include the vector norms $\|x\|_\infty = \max_i |x_i|$ and $\|x\|_2 = (\sum_i x_i^2)^{1/2}$, and the function norms $\|f\|_\infty = \max_{x \in [a, b]} |f(x)|$ and $\|f\|_{L^2(\Omega)} = (\int_\Omega |f|^2 dx)^{1/2}$.

1.3 Sources of Error

Write x for an exact value and x_A for an approximation. The absolute and relative errors are $x - x_A$ and $(x - x_A)/x$. Typical sources are

- modeling error (simplifications of the physical problem);
- blunders (arithmetic mistakes; programming errors);
- uncertainty in physical data;
- machine error (floating-point representation $fl(x)$, rounding or chopping, overflow/underflow);
- mathematical truncation error, which is the main object of later chapters.

2 Rootfinding for Nonlinear Equations

The chapter treats three related problems: finding roots of $f(x) = 0$, solving nonlinear systems, and unconstrained optimization. All methods are iterative: given x_0 , one produces $x_1, x_2, \dots$ [1].

Definition 2.1 (Order of convergence). A sequence $\{x_n\}$ converges to α with order $p \geq 1$ if

$$|\alpha - x_{n+1}| \leq c |\alpha - x_n|^p, \quad n \geq 0,$$

for some $c > 0$. If $p = 1$ the convergence is linear, and one requires $c < 1$; that constant is the rate of linear convergence.

2.1 The Bisection Method

Assume f is continuous on $[a, b]$ and $f(a)f(b) < 0$. The intermediate-value theorem guarantees at least one root in $[a, b]$.

1. Set $c = (a + b)/2$.
2. If $b - c \leq \varepsilon$, accept c as the root and stop.
3. If $\text{sign}(f(b)) \cdot \text{sign}(f(c)) < 0$, replace a by c ; otherwise replace b by c .
4. Return to step 1.

If c_n is the midpoint at step n and $\alpha = \lim c_n$, then

$$|\alpha - c_n| \leq 2^{-n}(b - a).$$

The method therefore converges linearly with rate $1/2$. It is slow, but it is guaranteed to converge and at every step it supplies an enclosure of the root.

2.2 Newton's Method

Assume $f \in C^2([a, b])$. Taylor expansion about x_n and the identity $f(\alpha) = 0$ give

$$\alpha = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{(\alpha - x_n)^2}{2} \frac{f''(\xi)}{f'(x_n)}.$$

Dropping the quadratic remainder yields Newton's method

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

and the error relation

$$\alpha - x_{n+1} = -(\alpha - x_n)^2 \frac{f''(\xi)}{2f'(x_n)}.$$

If $f'(\alpha) \neq 0$ and x_0 is sufficiently close to α , the iteration converges quadratically: $|\alpha - x_{n+1}| \leq c|\alpha - x_n|^2$. The method is fast when it converges, but it may diverge from a poor initial guess, and the update is undefined whenever $f'(x_n) = 0$. That is the usual motivation for a derivative-free substitute.

2.3 The Secant Method

Replace $f'(x_n)$ by the difference quotient $(f(x_n) - f(x_{n-1})) / (x_n - x_{n-1})$:

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

The error can be written with divided differences,

$$\alpha - x_{n+1} = -(\alpha - x_n)(\alpha - x_{n-1}) \frac{f[x_{n-1}, x_n, \alpha]}{f[x_{n-1}, x_n]}.$$

If f, f', f'' are continuous near α and $f'(\alpha) \neq 0$, nearby initial guesses produce convergence of order $p = (1 + \sqrt{5})/2 \approx 1.62$. The method needs only one new function evaluation per step (Newton needs f and f'), but it is slower than Newton and is not guaranteed to converge.

2.4 Fixed-Point Iteration

Rewrite the equation as $x = g(x)$ and iterate $x_{n+1} = g(x_n)$. Each solution of $x = g(x)$ is a fixed point of g . Convergence needs a contraction on a bounded invariant interval: g maps the interval into itself and the Lipschitz constant of g is strictly less than one. Under those hypotheses the error decays linearly.

Theorem 2.2. *Let g be continuous on $[a, b]$ with $g([a, b]) \subset [a, b]$, and suppose*

$$|g(x) - g(y)| \leq \lambda|x - y| \quad \text{for all } x, y \in [a, b]$$

with $0 < \lambda < 1$. Then $x = g(x)$ has a unique solution $\alpha \in [a, b]$. The iterates $x_n = g(x_{n-1})$ converge for every $x_0 \in [a, b]$, and

$$|\alpha - x_n| \leq \frac{\lambda^n}{1 - \lambda} |x_1 - x_0|.$$

Theorem 2.3. *If in addition $g \in C^1([a, b])$ and $\lambda = \max_{[a, b]} |g'(x)| < 1$, the same conclusions hold, and*

$$\lim_{n \rightarrow \infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = g'(\alpha).$$

2.5 Aitken Extrapolation

For a linearly convergent iteration with $0 < |g'(\alpha)| < 1$, the ratios $\lambda_n = (x_n - x_{n-1})/(x_{n-1} - x_{n-2})$ satisfy $\lambda_n \rightarrow g'(\alpha)$. Substituting the approximation $\alpha - x_n \approx \lambda_n(\alpha - x_{n-1})$ and rearranging produces Aitken's formula

$$\hat{x}_n = x_n - \frac{(x_n - x_{n-1})^2}{(x_n - x_{n-1}) - (x_{n-1} - x_{n-2})}.$$

The denominator must be monitored: if the sequence is already highly accurate the difference may underflow and pollute the correction.

2.6 Systems of Nonlinear Equations

A linear system $Ax = b$ can be solved by the methods of Chapter 8. A nonlinear problem is first linearized, so two errors appear: linearization error and, after a PDE discretization, discretization error.

For $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$ one may still use a fixed-point map $x = g(x)$. In two variables,

$$x = g_1(x, y), \quad y = g_2(x, y), \quad x_{n+1} = g(x_n).$$

The mean-value theorem on each component yields

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

where G_n is assembled from ∇g_1 and ∇g_2 evaluated between α and x_n . If $D \subset \mathbb{R}^m$ is closed, bounded, and convex, $g(D) \subset D$, and

$$\lambda = \max_{x \in D} \|G(x)\|_\infty < 1,$$

then $x = g(x)$ has a unique solution in D and $x_{n+1} = g(x_n)$ converges from any $x_0 \in D$. The condition $\lambda < 1$ is the multivariable analogue of $|g'(\alpha)| < 1$; the iteration remains linearly convergent.

Newton's method linearizes $f(\alpha) = 0$ about x_n . Dropping the second-order Taylor remainder gives

$$F(x_n)(\alpha - x_n) + f(x_n) \approx 0,$$

with Jacobian $F = Df$. In practice one solves

$$F(x_n)\delta_{n+1} = -f(x_n), \quad x_{n+1} = x_n + \delta_{n+1},$$

or $x_{n+1} = x_n - F(x_n)^{-1}f(x_n)$. Locally one has quadratic convergence, $\|\alpha - x_{n+1}\|_\infty \leq B\|\alpha - x_n\|_\infty^2$, provided the initial guess is close enough that the linearization is valid. A cheaper variant freezes the Jacobian for k steps. A damped update $x_{n+1} = x_n + sd_n$ with $d_n = -F(x_n)^{-1}f(x_n)$ and $s \in (0, 1]$ chosen to decrease $\|f\|_2$ improves global behavior from a poor initial guess.

2.7 Unconstrained Optimization

A local minimizer α of a smooth $f(x_1, \dots, x_m)$ satisfies $\nabla f(\alpha) = 0$. Newton's method on this system is

$$x_{n+1} = x_n - H(x_n)^{-1}\nabla f(x_n),$$

where $H_{ij} = \partial^2 f / \partial x_i \partial x_j$ is the Hessian. Descent methods pick a direction d_n along which f decreases and then minimize $\phi(s) = f(x_n + sd_n)$ over $s \geq 0$. Steepest descent uses $d_n = -\nabla f(x_n)$;

quasi-Newton methods (DFP, Broyden) approximate H or H^{-1} ; conjugate-gradient methods generate mutually H -conjugate search directions and are taken up again in Chapter 8.

A constrained problem such as $\min_{x>0} F(x)$ can be reduced to an unconstrained one by a barrier, for instance

$$\min_x \left(F(x) + \delta |\log x| \right),$$

which sends the objective to $+\infty$ as $x \rightarrow 0^+$. The resulting unconstrained problem is again solved by Newton or a descent method.

3 Interpolation Theory

Given distinct nodes $x_0, \dots, x_n$ and values $y_0, \dots, y_n$, one seeks a polynomial p with $p(x_i) = y_i$ [1]. Interpolation matches data exactly; approximation in Chapter 4 instead minimizes a global residual.

3.1 Polynomial Interpolation and the Vandermonde System

Writing $p(x) = \sum_{j=0}^n a_j x^j$ produces $Aa = y$ with Vandermonde entries $A_{ij} = x_i^j$. For three nodes this is

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

The monomials $\{1, x, x^2, \dots\}$ become nearly linearly dependent as the degree grows, so A is severely ill-conditioned: a small perturbation of the data can change the coefficients dramatically. That is why one does not form A in practice.

Theorem 3.1. *Given $n + 1$ distinct nodes $x_0, \dots, x_n$ and ordinates $y_0, \dots, y_n$, there is a unique polynomial p of degree at most n interpolating y_i at x_i .*

Existence follows from the Lagrange formula below. For uniqueness, if r_1 and r_2 both interpolate, then $R = r_1 - r_2$ has degree at most n and $n + 1$ roots, hence $R \equiv 0$.

Lagrange's formula is

$$p_n(x) = \sum_{i=0}^n y_i \ell_i(x), \quad \ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}.$$

For $n = 2$,

$$p_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} y_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} y_2.$$

The representation is convenient, but adding a node forces every ℓ_i to be rebuilt.

Theorem 3.2 (Lagrange error). *If $x_0, \dots, x_n, x \in I$ and $f \in C^{n+1}(I)$, then for some $\xi \in I$*

$$E(x) := f(x) - p_n(x) = \frac{(x - x_0) \cdots (x - x_n)}{(n + 1)!} f^{(n+1)}(\xi).$$

Proof. Fix x distinct from the nodes and set $\omega(t) = (t - x_0) \cdots (t - x_n)$. Define

$$\phi(t) = f(t) - p_n(t) - \frac{f(x) - p_n(x)}{\omega(x)} \omega(t).$$

Then ϕ vanishes at $x_0, \dots, x_n$ and at x , hence has $n + 2$ roots. Rolle's theorem (the mean-value theorem applied repeatedly) implies that $\phi^{(n+1)}$ has at least one root ξ . Since $\deg p_n \leq n$ one has $p_n^{(n+1)} \equiv 0$ and $\omega^{(n+1)} = (n + 1)!$, so

$$0 = \phi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{f(x) - p_n(x)}{\omega(x)} (n + 1)!,$$

which is the claimed formula. $\square$

Example 3.3. For $f(x) = \log_{10} x$ and linear interpolation on $[x_0, x_1]$ with $h = x_1 - x_0$ and $x_0 > 1$,

$$|E(x)| \leq \frac{h^2}{8} \cdot \frac{\log_{10} e}{x_0^2} \leq 0.0542 h^2.$$

3.2 Newton Divided Differences

To raise the degree without rebuilding the whole formula, write $p_n = p_{n-1} + C$ with $\deg C = n$. Then $C(x_i) = 0$ for $i = 0, \dots, n - 1$, so $C(x) = a_n(x - x_0) \cdots (x - x_{n-1})$ and

$$a_n = f[x_0, \dots, x_n] := \frac{f(x_n) - p_{n-1}(x_n)}{(x_n - x_0) \cdots (x_n - x_{n-1})}.$$

The table of divided differences is built from the recurrence

$$f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0},$$

which follows by comparing two degree- $(n - 1)$ Newton interpolants, one on $\{x_0, \dots, x_{n-1}\}$ and one on $\{x_1, \dots, x_n\}$, and matching the leading coefficient of their difference. Divided differences are symmetric in their arguments. Newton's formula is

$$p_n(x) = f(x_0) + \sum_{k=1}^n (x - x_0) \cdots (x - x_{k-1}) f[x_0, \dots, x_k].$$

Including an extra node t gives

$$f(t) - p_n(t) = (t - x_0) \cdots (t - x_n) f[x_0, \dots, x_n, t],$$

and comparison with the Lagrange remainder yields

$$f[x_0, \dots, x_n, t] = \frac{f^{(n+1)}(\xi)}{(n + 1)!}.$$

The same uniqueness argument as above shows that the Newton and Lagrange interpolants of degree n coincide: their difference would be a degree- n polynomial with $n + 1$ roots.

Uniform-node interpolation of $f(x) = 1/(1 + x^2)$ on $[-5, 5]$ diverges near the endpoints as $n \rightarrow \infty$ (Runge phenomenon). Better node families appear in Chapter 4.

3.3 Hermite Interpolation

The $2n$ conditions $p(x_i) = y_i$ and $p'(x_i) = y'_i$ determine a unique polynomial of degree at most $2n - 1$. With the Lagrange factors ℓ_i , the basis

$$h_i(x) = (1 - 2\ell'_i(x_i)(x - x_i))[\ell_i(x)]^2, \quad \tilde{h}_i(x) = (x - x_i)[\ell_i(x)]^2$$

gives

$$H_n(x) = \sum_{i=1}^n y_i h_i(x) + \sum_{i=1}^n y'_i \tilde{h}_i(x).$$

Differentiating the basis confirms $h_i(x_j) = \delta_{ij}$, $h'_i(x_j) = 0$, $\tilde{h}_i(x_j) = 0$, and $\tilde{h}'_i(x_j) = \delta_{ij}$:

$$\begin{aligned} h'_i(x) &= -2\ell'_i(x_i)[\ell_i(x)]^2 + 2(1 - 2\ell'_i(x_i)(x - x_i))\ell_i(x)\ell'_i(x), \\ \tilde{h}'_i(x) &= [\ell_i(x)]^2 + 2(x - x_i)\ell_i(x)\ell'_i(x). \end{aligned}$$

Uniqueness follows because the difference of two such polynomials has n double roots and degree at most $2n - 1$. On two nodes a, b the cubic Hermite interpolant is

$$\begin{aligned} H_2(x) &= \left(1 + 2\frac{x-a}{b-a}\right)\left(\frac{b-x}{b-a}\right)^2 f(a) + \left(1 + 2\frac{b-x}{b-a}\right)\left(\frac{x-a}{b-a}\right)^2 f(b) \\ &\quad + \frac{(x-a)(b-x)^2}{(b-a)^2} f'(a) - \frac{(x-a)^2(b-x)}{(b-a)^2} f'(b). \end{aligned}$$

3.4 Piecewise Polynomials and Splines

A function is piecewise polynomial of order r if its restriction to each subinterval $[x_i, x_{i+1}]$ has degree less than r . Piecewise Lagrange interpolation is local and cheap, but only C^0 . A spline of order $m \geq 1$ on $a = x_0 < \dots < x_n = b$ is piecewise of degree $< m$ and of class C^{m-2} . Cubic splines ($m = 4$) are the standard choice in animation and CAD: they are C^2 , have no Runge oscillation, and the second derivative (curvature) is continuous.

On each interval a cubic $s(x) = a_i + b_i x + c_i x^2 + d_i x^3$ has four coefficients, hence $4n$ unknowns. Interpolation $s(x_i) = y_i$ gives $n + 1$ conditions, and matching s, s', s'' at the $n - 1$ interior knots gives $3(n - 1)$ more, for a total of $4n - 2$. Two endpoint conditions remain.

Write $M_i = s''(x_i)$ and $h_i = x_{i+1} - x_i$. Since s'' is linear on $[x_i, x_{i+1}]$,

$$s''(x) = \frac{(x_{i+1} - x)M_i + (x - x_i)M_{i+1}}{h_i}.$$

Integrating twice,

$$s(x) = \frac{(x_{i+1} - x)^3 M_i + (x - x_i)^3 M_{i+1}}{6h_i} + C(x_{i+1} - x) + D(x - x_i).$$

The interpolation conditions $s(x_i) = y_i$ and $s(x_{i+1}) = y_{i+1}$ fix

$$C = \frac{y_i}{h_i} - \frac{h_i M_i}{6}, \quad D = \frac{y_{i+1}}{h_i} - \frac{h_i M_{i+1}}{6}.$$

Requiring s' to be continuous at interior knots produces the $n - 1$ interior equations

$$\frac{h_{i-1}}{6} M_{i-1} + \frac{h_{i-1} + h_i}{3} M_i + \frac{h_i}{6} M_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}.$$

The complete cubic spline imposes $s'(a) = y'_0$ and $s'(b) = y'_n$, which add

$$\frac{h_0}{3}M_0 + \frac{h_0}{6}M_1 = \frac{y_1 - y_0}{h_0} - y'_0, \quad \frac{h_{n-1}}{6}M_{n-1} + \frac{h_{n-1}}{3}M_n = y'_n - \frac{y_n - y_{n-1}}{h_{n-1}}.$$

The resulting system $AM = D$ is tridiagonal, symmetric, and positive definite, hence uniquely solvable.

Theorem 3.4. *If $f \in C^4([a, b])$ and s is the complete cubic spline interpolant on a partition τ_n of mesh $|\tau_n|$, then*

$$\max_{a \leq x \leq b} |f^{(j)}(x) - s^{(j)}(x)| \leq c_j |\tau_n|^{4-j} \max_{a \leq x \leq b} |f^{(4)}(x)|$$

for $j = 0, 1, 2$.

The truncated-power basis $s(x) = p_{m-1}(x) + \sum_j \beta_j (x - x_j)_+^{m-1}$ is theoretically complete but often ill-conditioned, so it is not used as a working representation. Cubic B -splines

$$B_i(x) = (x_{i+4} - x_i) f_x[x_i, \dots, x_{i+4}], \quad f_x(t) = (t - x)_+^3,$$

have support (x_i, x_{i+4}) , satisfy $0 \leq B_i \leq 1$ and $\sum_i B_i \equiv 1$ on $[x_0, x_n]$, and form a stable local basis for the cubic spline space.

4 Approximation of Functions

Interpolation matches data at prescribed nodes. Approximation instead seeks r minimizing a global residual, typically in an L^p norm [1]. On a domain Ω ,

$$\|g\|_p = \left(\int_{\Omega} |g|^p dx \right)^{1/p}, \quad \|g\|_{\infty} = \text{ess sup } |g|.$$

The L^2 case is computationally the most convenient because the objective is quadratic and the normal equations are linear.

4.1 The Weierstrass Theorem and Taylor Expansion

Theorem 4.1 (Weierstrass). *If f is continuous on $[a, b]$, then for every $\varepsilon > 0$ there is a polynomial p with $|f(x) - p(x)| \leq \varepsilon$ for all $x \in [a, b]$.*

Existence of some polynomial approximant is therefore guaranteed; the remaining question is how to construct a good one. Taylor polynomials are a convenient first choice but need not distribute error evenly. For $f(x) = e^x$ on $[-1, 1]$,

$$p_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}, \quad E(x) = \frac{x^4}{24} e^{\xi}, \quad |E(x)| \leq \frac{e}{24} x^4,$$

so the error grows with $|x|$ and is concentrated near the endpoints.

4.2 The Least Squares Problem

One minimizes $\|f - r\|_2$. For a linear trial function the objective is a quadratic in the coefficients, so critical points solve a linear system (the Hessian of F is the Gram matrix).

Example 4.2. For $f(x) = e^x$ on $[-1, 1]$ and $r_1(x) = a + bx$,

$$F(a, b) = \int_{-1}^1 (e^x - a - bx)^2 dx.$$

The normal equations $\partial F / \partial a = 0$ and $\partial F / \partial b = 0$ are

$$a = \frac{1}{2} \int_{-1}^1 e^x dx \approx 1.1752, \quad b = \frac{3}{2} \int_{-1}^1 x e^x dx \approx 1.1036,$$

and $\|e^x - r_1\|_\infty \approx 0.44$. The cubic least-squares fit reduces the uniform error to about 0.0112.

Let $w \geq 0$ be a weight with $\int_a^b |x|^n w(x) dx < \infty$ for all n , and such that $\int_a^b w g = 0$ for nonnegative continuous g forces $g \equiv 0$. Standard weights are $w \equiv 1$ on a finite interval, $w(x) = (1 - x^2)^{-1/2}$ on $[-1, 1]$, $w(x) = e^{-x}$ on $[0, \infty)$, and $w(x) = e^{-x^2}$ on $\mathbb{R}$. The weighted problem is

$$\min_{\deg r \leq n} \int_a^b w(x) (f(x) - r(x))^2 dx.$$

Expanding $r(x) = \sum_{k=0}^n a_k x^k$ and setting $\partial F / \partial a_i = 0$ produces

$$\sum_{k=0}^n a_k \int_a^b w(x) x^{i+k} dx = \int_a^b w(x) f(x) x^i dx,$$

or $Aa = b$ with the Hilbert-type matrix $A_{ik} = \int_a^b w x^{i+k} dx$. The monomials $\{x^k\}$ are nearly linearly dependent, so A is dense and extremely ill-conditioned: the same phenomenon already seen in the Vandermonde interpolant of Chapter 3. The remedy is to change basis.

4.3 Inner Products and Orthogonal Polynomials

Define

$$(f, g) = \int_a^b w(x) f(x) g(x) dx, \quad f, g \in C[a, b],$$

and $\|f\|_2 = \sqrt{(f, f)}$. This is an inner product:

- (1) $(u, u) \geq 0$, and $(u, u) = 0$ if and only if $u = 0$ (using the assumption on w);
- (2) $(\alpha u + \beta v, w) = \alpha(u, w) + \beta(v, w)$.

Cauchy–Schwarz $|(f, g)| \leq \|f\|_2 \|g\|_2$ follows at once.

Theorem 4.3 (Gram–Schmidt). *There is a sequence $\{\phi_n\}$ with $\deg \phi_n = n$ and $(\phi_n, \phi_m) = 0$ for $n \neq m$. The sequence may be normalized so that $\|\phi_n\| = 1$.*

Start with $\phi_0 = c$ chosen so that $(\phi_0, \phi_0) = c^2 \int w = 1$. Then $\psi_1 = x + c_{1,0}\phi_0$ with $c_{1,0} = -(x, \phi_0)$, and $\phi_1 = \psi_1/\|\psi_1\|_2$. Inductively,

$$\psi_n(x) = x^n + \sum_{j=0}^{n-1} c_{n,j}\phi_j(x), \quad c_{n,j} = -(x^n, \phi_j),$$

and $\phi_n = \psi_n/\|\psi_n\|_2$. On $[-1, 1]$ with $w \equiv 1$ this produces the orthonormal Legendre family

$$\phi_0 = \sqrt{1/2}, \quad \phi_1 = \sqrt{3/2}x, \quad \phi_2 = \frac{1}{2}\sqrt{5/2}(3x^2 - 1).$$

Classical (unnormalized) families are as follows.

Legendre. $w \equiv 1$ on $[-1, 1]$, $p_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n} (1 - x^2)^n$, with $p_0 = 1$, $p_1 = x$, $p_2 = \frac{1}{2}(3x^2 - 1)$, and $(p_n, p_n) = 2/(2n + 1)$. They solve Legendre's equation.

Chebyshev. $w = (1 - x^2)^{-1/2}$ on $[-1, 1]$, $T_n(x) = \cos(n \arccos x)$. Setting $x = \cos \theta$ gives $T_n = \cos(n\theta)$ and

$$T_{n+1} + T_{n-1} = 2 \cos(n\theta) \cos \theta = 2x T_n,$$

hence $T_{n+1} = 2xT_n - T_{n-1}$. The first members are $T_0 = 1$, $T_1 = x$, $T_2 = 2x^2 - 1$, $T_3 = 4x^3 - 3x$. The inner products are $(T_n, T_m) = \pi$ if $n = m = 0$, $\pi/2$ if $n = m > 0$, and 0 otherwise.

Laguerre. $w = e^{-x}$ on $[0, \infty)$, $L_n(x) = \frac{e^x}{n!} \frac{d^n}{dx^n} (x^n e^{-x})$. The first members are $L_0 = 1$, $L_1 = 1 - x$, $L_2 = \frac{1}{2}(x^2 - 4x + 2)$, and $\|L_n\|_2 = 1$.

Theorem 4.4. *If $\{\phi_n\}$ is orthogonal and f is a polynomial of degree m , then*

$$f = \sum_{n=0}^m \frac{(f, \phi_n)}{(\phi_n, \phi_n)} \phi_n.$$

In particular $(f, \phi_m) = 0$ whenever $\deg f \leq m - 1$.

Theorem 4.5. *Each ϕ_n has exactly n distinct real roots in the open interval (a, b) .*

Those roots are precisely the nodes of the Gaussian quadrature rule of Chapter 5, which is why orthogonal polynomials reappear in numerical integration.

The three-term recurrences are

$$\begin{aligned} p_{n+1}(x) &= \frac{2n+1}{n+1} x p_n(x) - \frac{n}{n+1} p_{n-1}(x), \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x), \\ L_{n+1}(x) &= \frac{2n+1-x}{n+1} L_n(x) - \frac{n}{n+1} L_{n-1}(x). \end{aligned}$$

4.4 Least Squares in an Orthogonal Basis

Let $\{\phi_j\}$ be orthogonal (not necessarily orthonormal) and write $r = \sum_{j=0}^n a_j \phi_j$. Expanding

$$G(a) = \left(f - \sum_j a_j \phi_j, f - \sum_i a_i \phi_i \right)$$

and setting $\partial G/\partial a_k = 0$ yields, by orthogonality,

$$a_k = \frac{(f, \phi_k)}{(\phi_k, \phi_k)} = \frac{\int_a^b w f \phi_k dx}{\int_a^b w \phi_k^2 dx}.$$

If the family is orthonormal this reduces to $a_k = (f, \phi_k)$, and

$$r_n = \sum_{j=0}^n (f, \phi_j) \phi_j, \quad \|f - r_n\|_2^2 = \|f\|_2^2 - \|r_n\|_2^2.$$

Bessel's inequality follows at once:

$$\|r_n\|_2^2 = \sum_{j=0}^n (f, \phi_j)^2 \leq \|f\|_2^2.$$

The sequence $\|r_n\|_2$ is increasing and bounded, hence convergent. The remainders then form a Cauchy sequence: by the triangle inequality

$$\|r_n - r_m\|_2 = \|r_n - f + f - r_m\|_2 \leq \|f - r_n\|_2 + \|f - r_m\|_2,$$

and $\|f - r_n\|_2^2 = \|f\|_2^2 - \|r_n\|_2^2$ tends to a limit. On a finite interval one obtains $\|f - r_n\|_2 \rightarrow 0$, which is Parseval's identity

$$\|f\|_2 = \left(\sum_{j=0}^{\infty} (f, \phi_j)^2 \right)^{1/2}.$$

In the orthogonal basis the Gram matrix is diagonal, so the ill-conditioned Hilbert system is avoided entirely. The same idea—orthogonal search directions—reappears in GMRES and conjugate gradient (Chapter 8).

5 Numerical Integration

The task is to approximate $I(f) = \int_a^b f(x) dx$ [1]. Most formulas come from integrating an interpolant of f and take the form

$$I_n(f) = \sum_{j=1}^n w_j f(x_j),$$

with weights w_j and nodes $x_j \in [a, b]$. The integral mean-value theorem (Theorem 1.3) is used repeatedly for error formulas.

Example 5.1. For $I = \int_0^1 (e^x - 1)/x dx$, the Taylor truncation $e^x - 1 = \sum_{j=1}^n x^j/j! + x^{n+1}e^\xi/(n+1)!$ yields

$$I_n = \sum_{j=1}^n \frac{1}{j \cdot j!}, \quad \frac{1}{(n+1)!(n+1)} \leq I - I_n \leq \frac{e}{(n+1)!(n+1)}.$$

5.1 The Trapezoidal Rule

The linear interpolant of f at a and b integrates to

$$I(f) \approx \frac{b-a}{2}(f(a) + f(b)).$$

The interpolation remainder $(x-a)(x-b)f[a, b, x]$ and Theorem 1.3 give

$$E(f) = -\frac{(b-a)^3}{12}f''(\eta).$$

On a uniform partition of n subintervals of width $h = (b-a)/n$ the composite rule is

$$I_n(f) = h\left(\frac{1}{2}f_0 + f_1 + \cdots + f_{n-1} + \frac{1}{2}f_n\right), \quad E_n(f) = -\frac{(b-a)h^2}{12}f''(\eta).$$

The asymptotic estimate

$$\tilde{E}_n(f) = -\frac{h^2}{12}(f'(b) - f'(a))$$

satisfies $\tilde{E}_n/E_n \rightarrow 1$. Adding it produces the corrected trapezoidal rule $CT_n = I_n + \tilde{E}_n$, which is $\mathcal{O}(h^4)$ for smooth f .

5.2 Simpson's Rule

Quadratic interpolation at a , $c = (a+b)/2$, and b gives

$$I_2(f) = \frac{h}{3}(f(a) + 4f(c) + f(b)), \quad h = \frac{b-a}{2}.$$

The cubic weight $(x-a)(x-c)(x-b)$ changes sign, so one integrates by parts against $w(x) = \int_a^x (t-a)(t-c)(t-b) dt$ to obtain

$$E_2(f) = -\frac{f^{(4)}(\eta)}{90}h^5.$$

Thus Simpson is exact for cubics (degree of precision 3), one order higher than the interpolant. The composite rule on an even number of panels is

$$I_n(f) = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{n-2} + 4f_{n-1} + f_n),$$

with $E_n(f) = -\frac{(b-a)h^4}{180}f^{(4)}(\eta) = \mathcal{O}(h^4)$. It matches the corrected trapezoidal rate without requiring derivative values.

5.3 Newton–Cotes Formulas

Trapezoidal and Simpson are the first two closed Newton–Cotes rules: if p_n interpolates f at the uniform nodes $x_j = a + jh$, then

$$I_n(f) = \int_a^b p_n(x) dx = \sum_{j=0}^n w_j f(x_j), \quad w_j = \int_a^b \ell_j(x) dx.$$

The $n = 3$ case is the 3/8 rule, $I_3(f) = \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3)$. High-order Newton–Cotes rules need not converge even for smooth integrands, because the nodes are forced to be equally spaced.

5.4 Gaussian Quadrature

To reach degree of precision $2n - 1$ one is free to choose both nodes and weights in

$$\int_a^b w(x)f(x)dx \approx \sum_{j=1}^n w_j f(x_j).$$

For $w \equiv 1$ on $[-1, 1]$, the $n = 1$ conditions $E(1) = E(x) = 0$ give $2f(0)$; the $n = 2$ conditions give $f(-\sqrt{3}/3) + f(\sqrt{3}/3)$.

Theorem 5.2. *Let $\{\varphi_n\}$ be the orthogonal polynomials for $w \geq 0$ on (a, b) , with $\varphi_n(x) = A_n \prod_{j=1}^n (x - x_j)$ and $\gamma_n = \int_a^b w \varphi_n^2$. There is a unique n -point formula of precision $2n - 1$. If $f \in C^{2n}([a, b])$,*

$$\int_a^b w f = \sum_{j=1}^n w_j f(x_j) + \frac{\gamma_n}{A_n^2 (2n)!} f^{(2n)}(\eta),$$

where the nodes are the zeros of φ_n and

$$w_j = -\frac{A_{n+1}\gamma_n}{A_n \varphi'_n(x_j) \varphi_{n+1}(x_j)}.$$

Gauss–Legendre uses Legendre zeros on $[-1, 1]$, with $w_j = 2/[(1 - x_j^2)(p'_n(x_j))^2]$. Gauss–Chebyshev uses $x_j = \cos((2j - 1)\pi/(2n))$ and equal weights $w_j = \pi/n$. These nodes are precisely the well-distributed grids that avoid the Runge phenomenon of Chapter 3.

5.5 Singular Integrals

Standard rules degrade when the integrand is singular or the interval is infinite. Three remedies are used.

Change of variables. $\int_0^b f(x)x^{-1/2}dx$ with $f \in C^\infty$ becomes $\int_0^{\sqrt{b}} 2f(u^2)du$ after $x = u^2$. An infinite integral $\int_1^\infty f(x)x^{-p}dx$ ($p > 1$) is mapped to a finite interval by $x = u^{-\alpha}$.

Gaussian rules with matching weights. Integrals such as $\int_0^\infty x^\alpha e^{-x} f(x)dx$, $\int_{-\infty}^\infty e^{-x^2} f(x)dx$, and $\int_0^1 \log(1/x) f(x)dx$ are treated by the orthogonal polynomials associated with those weights.

Analytic treatment of the singularity. Split $I = I_1 + I_2$ at a small ε , integrate the singular piece by a Taylor expansion, and apply a standard rule to the regular remainder. For $\int_0^1 \cos x \log x dx$,

$$\int_0^\varepsilon \cos x \log x dx = \varepsilon(\log \varepsilon - 1) - \frac{\varepsilon^3}{6} \left(\log \varepsilon - \frac{1}{3} \right) + \cdots.$$

Part II

Ordinary Differential Equations

6 Numerical Methods for Ordinary Differential Equations

The goal is to solve the initial-value problem

$$y' = f(x, y), \quad y(x_0) = y_0, \tag{1}$$

with f continuous on a domain D [1]. A function u is a solution on $[a, b]$ if $(x, u(x)) \in D$, $u(x_0) = y_0$, and $u'(x) = f(x, u(x))$. In this chapter the ODE step is written h ; when the same methods are used later for PDEs we write $k = \Delta t$.

Example 6.1 (Duhamel). The linear problem $y' = \lambda y + g(x)$, $y(0) = y_0$, has the variation-of-constants solution

$$y(x) = y_0 e^{\lambda x} + \int_0^x e^{\lambda(x-t)} g(t) dt.$$

For a linear system $u' = Au + g$ with constant A the same formula holds with the matrix exponential [6].

Example 6.2. The nonlinear equation $y' = -y^2$ has the trivial solution $y \equiv 0$ and the family $y(x) = 1/(x + c)$. Direction fields (e.g. $y' = -y$ and $y' = x^2 + y^2$) make uniqueness for a fixed initial value geometrically plausible.

6.1 Existence, Uniqueness, and Stability

Theorem 6.3 (Existence and uniqueness). *Let $D = \{(x, y) : |y - y_0| < \beta, |x - x_0| < \alpha\}$ and let f be continuous on D and Lipschitz in y ,*

$$|f(x, y_1) - f(x, y_2)| \leq K|y_1 - y_2|.$$

Then (1) has a unique solution on $[x_0 - \gamma, x_0 + \gamma]$, where $\gamma = \min\{\alpha, \beta/M\}$ and $M = \max_D |f|$.

If $\partial f / \partial y$ exists, one may take $K = \max |\partial f / \partial y|$. For $y' = 1 + \sin(xy)$ on $0 \leq x \leq 1$ this gives $K = 1$.

Theorem 6.4 (Stability of the IVP). *Under the same hypotheses, the perturbed problem $y' = f(x, y) + \delta(x)$, $y(x_0) = y_0 + \varepsilon$ has a unique solution $y(x; \delta, \varepsilon)$ for all sufficiently small ε and $\|\delta\|_\infty$, and*

$$\max_{|x-x_0| \leq \gamma} |y(x) - y(x; \delta, \varepsilon)| \leq \frac{1}{1 - \gamma K} (|\varepsilon| + \gamma \|\delta\|_\infty).$$

Thus the solution depends continuously on the data. Continuous dependence does not preclude ill-conditioning: $y' = 100y - 101e^{-x}$, $y(0) = 1$ has solution e^{-x} , but the perturbation $y(0) = 1 + \varepsilon$ produces an extra term εe^{100x} .

A first-order system $y' = f(x, y)$ is the natural setting. A scalar equation of order m is reduced by $y_1 = y$, $y_2 = y'$, $\dots$, $y_m = y^{(m-1)}$. In particular

$$y'' = a_1(x)y' + a_0(x)y + g(x)$$

becomes $Y' = AY + G$ with

$$A = \begin{pmatrix} 0 & 1 \\ a_0(x) & a_1(x) \end{pmatrix}.$$

6.2 Euler's Method

On the grid $x_n = x_0 + nh$ the forward Euler method is

$$y_{n+1} = y_n + hf(x_n, y_n).$$

It arises by dropping the remainder in $y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(\xi_n)$, or by a one-sided difference, or by a left-rectangle quadrature of y' .

Theorem 6.5 (Convergence of Euler). *If f is Lipschitz and y'' is bounded on $[x_0, b]$, and if $|y(x_0) - y_0| \leq c_1 h$, then $\max_{x_n \in [x_0, b]} |y(x_n) - y_n| \leq Ch$.*

The proof writes $e_n = y(x_n) - y_n$ and $\tau_n = \frac{h}{2} y''(\xi_n)$, obtains $|e_{n+1}| \leq (1 + hK)|e_n| + h|\tau_n|$, and sums the geometric series using $(1 + x)^n \leq e^{nx}$. The same recurrence with a perturbed right-hand side shows that Euler is a stable method.

6.3 Linear Multistep Methods

A general $(p + 1)$ -step method is

$$y_{n+1} = \sum_{j=0}^p a_j y_{n-j} + h \sum_{j=-1}^p b_j f(x_{n-j}, y_{n-j}), \quad n \geq p. \quad (2)$$

The values $y_1, \dots, y_p$ must be supplied by another method. The scheme is explicit if $b_{-1} = 0$ and implicit otherwise. Euler is the case $p = 0$, $a_0 = 1$, $b_0 = 1$.

The truncation error of a smooth y is

$$T_n(y) = y(x_{n+1}) - \left(\sum_{j=0}^p a_j y(x_{n-j}) + h \sum_{j=-1}^p b_j y'(x_{n-j}) \right), \quad \tau_n = T_n/h.$$

The method is consistent if $\max |\tau_n| \rightarrow 0$ as $h \rightarrow 0$.

Theorem 6.6 (Consistency order). *(2) is consistent if and only if*

$$\sum_{j=0}^p a_j = 1, \quad - \sum_{j=0}^p j a_j + \sum_{j=-1}^p b_j = 1.$$

It has order m for $y \in C^{m+1}$ if these hold together with

$$\sum_{j=0}^p (-j)^i a_j + i \sum_{j=-1}^p (-j)^{i-1} b_j = 1, \quad i = 2, \dots, m.$$

If the starting errors satisfy $\eta(h) = \max_{i \leq p} |y(x_i) - y_i| \rightarrow 0$, the method is consistent, and the a_j are nonnegative, then $\max |y(x_n) - y_n| \leq c_1 \eta(h) + c_2 \tau(h)$. Order- m starting data therefore produce an $\mathcal{O}(h^m)$ method.

Midpoint method. The centered difference $y'(x_n) = (y(x_{n+1}) - y(x_{n-1})) / (2h) - \frac{h^2}{6} y'''(\xi)$ yields the explicit two-step rule $y_{n+1} = y_{n-1} + 2hf(x_n, y_n)$ of order 2.

Trapezoidal method. Integrating y' on $[x_n, x_{n+1}]$ and applying the trapezoidal rule of Chapter 5 gives the implicit one-step scheme

$$y_{n+1} = y_n + \frac{h}{2} (f(x_n, y_n) + f(x_{n+1}, y_{n+1}))$$

of order 2. The nonlinear equation is solved by the fixed-point iteration of Chapter 2,

$$y_{n+1}^{(j+1)} = y_n + \frac{h}{2} (f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(j)})),$$

which converges when $hK/2 < 1$. On $y' = \lambda y + g$ the error w_n of a perturbed initial value satisfies

$$w_{n+1} = \frac{1 + h\lambda/2}{1 - h\lambda/2} w_n.$$

If $\operatorname{Re} \lambda < 0$ then $|r| < 1$ for every $h > 0$, so the method is **A-stable**.

6.4 One-Step Methods

- Forward Euler: $y_{n+1} = y_n + hf(x_n, y_n)$, explicit, order 1.
- Backward Euler: $y_{n+1} = y_n + hf(x_{n+1}, y_{n+1})$, implicit, order 1, unconditionally stable.
- Trapezoidal rule, as above.
- Taylor methods replace higher derivatives of y by differentiated f ; they are accurate but impractical.
- The second-order two-stage Runge–Kutta method evaluates f at the Euler midpoint: $y_{n+1} = y_n + hf(x_n + h/2, y_n + (h/2)f(x_n, y_n))$.
- Classical RK4 [10]:

$$\begin{aligned}
 Y_1 &= y_n, \\
 Y_2 &= y_n + \frac{h}{2}f(x_n, Y_1), \\
 Y_3 &= y_n + \frac{h}{2}f(x_n + \frac{h}{2}, Y_2), \\
 Y_4 &= y_n + hf(x_n + \frac{h}{2}, Y_3), \\
 y_{n+1} &= y_n + \frac{h}{6} [f(x_n, Y_1) + 2f(x_n + \frac{h}{2}, Y_2) + 2f(x_n + \frac{h}{2}, Y_3) + f(x_{n+1}, Y_4)].
 \end{aligned}$$

6.5 Adams Methods and Comparison

Coefficients may be obtained from the order conditions or by integrating an interpolant of f . Simpson quadrature on $[x_{n-1}, x_{n+1}]$ produces Milne’s method $y_{n+1} = y_{n-1} + \frac{h}{3}(f_{n-1} + 4f_n + f_{n+1})$.

Adams–Bashforth (explicit) and Adams–Moulton (implicit) formulas include

$$y_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1}) \quad (\text{AB2}),$$

$$y_{n+1} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}) \quad (\text{AB3}),$$

$$y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n) \quad (\text{AM1 / trapezoidal}),$$

$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}) \quad (\text{AM2}).$$

Adams–Moulton truncation errors are substantially smaller than those of Adams–Bashforth of the same step number; implicitness is also the source of better stability. Dahlquist’s equivalence theorem states that a linear multistep method converges if and only if it is consistent and zero-stable [6, 10].

One-step methods are self-starting and tolerate step-size changes or isolated singularities at grid points. Multistep methods reuse previous f -values (typically one new evaluation per step) but need a starter and lose accuracy if f is nonsmooth inside the stencil.

6.6 Root Condition and Absolute Stability

A method is *stable* if a perturbation of size ε in the starting values produces a solution change bounded by $c\varepsilon$ uniformly in h . It is *convergent* if $\max |Y(x_n) - y_n| \rightarrow 0$ whenever the starting errors vanish with h .

The first characteristic polynomial is $\rho(r) = r^{p+1} - \sum_{j=0}^p a_j r^{p-j}$. Consistency forces $\rho(1) = 0$. The **root condition** requires $|r_j| \leq 1$ for every root, and that every root on the unit circle be simple.

Theorem 6.7 (Dahlquist). *For a consistent multistep method, the following are equivalent: convergence; the root condition; stability.*

The explicit two-step rule $y_{n+1} = 3y_n - 2y_{n-1} + \frac{h}{2}(f_n - 3f_{n-1})$ has $\rho(r) = r^2 - 3r + 2 = (r-1)(r-2)$ and is unstable: on $y' = 0$ a starting perturbation ε grows like 2^n .

Applying (2) to $y' = \lambda y$ produces the characteristic polynomial $P(r) = (1 - h\lambda b_{-1})r^{p+1} - \sum_{j=0}^p (a_j + h\lambda b_j)r^{p-j}$. The region of absolute stability is the set of $z = h\lambda$ for which all roots of P satisfy $|r| \leq 1$ (strictly, with simplicity on the boundary). Forward Euler is stable for z in the disk $|1+z| \leq 1$; backward Euler and the trapezoidal rule contain the whole left half-plane (A-stability). The trapezoidal amplification tends to -1 as $|z| \rightarrow \infty$, so very large steps, while stable, can produce slowly decaying oscillations (lack of L-stability) [6].

Part III

Linear Algebra

7 Linear Algebra

This chapter collects the linear-algebraic language used for linear systems, eigenvalues, and later PDE discretizations [1].

7.1 Vector Spaces, Matrices, and Linear Systems

A vector space is a set closed under addition and scalar multiplication. Standard examples are $\mathbb{R}^n$, $\mathbb{C}^n$, the polynomials of degree at most n , and $C[a, b]$. Vectors $v_1, \dots, v_m$ are linearly independent if $\sum \alpha_i v_i = 0$ forces all $\alpha_i = 0$. A basis is a linearly independent spanning set; its cardinality is the dimension. $\{1, x, \dots, x^n\}$ is a basis of the polynomials of degree $\leq n$; the standard unit vectors form a basis of $\mathbb{R}^n$; $C[a, b]$ is infinite-dimensional.

An $m \times n$ matrix $A = (a_{ij})$ admits the usual operations $A + B$, αA , AB , the transpose A^t , and the conjugate transpose A^* . A is symmetric if $A = A^t$ and skew-symmetric if $A^t = -A$. The rank is the number of linearly independent rows (equivalently, columns).

Theorem 7.1. *For a square $n \times n$ matrix A the following are equivalent: $Ax = b$ has a unique solution for every b ; $Ax = 0$ implies $x = 0$; A^{-1} exists; $\det A \neq 0$; $\text{rank}(A) = n$.*

The Euclidean inner product is $(x, y) = \sum x_i y_i$ on $\mathbb{R}^n$ and $(x, y) = \sum x_i \bar{y}_i$ on $\mathbb{C}^n$, with $\|x\|_2 = \sqrt{(x, x)}$. Cauchy–Schwarz and the triangle inequality hold; $(Ax, y) = (x, A^*y)$; two vectors are orthogonal when $(x, y) = 0$ and orthonormal when they are also unit. A square U is unitary if $U^*U = UU^* = I$ (orthogonal if U is real); its columns form an orthonormal basis.

Theorem 7.2. *If $\{u_j\}$ is an orthogonal basis and $u = \sum \alpha_j u_j$, then $\alpha_j = (u, u_j)/\|u_j\|^2$. For an orthonormal basis, $\alpha_j = (u, u_j)$.*

7.2 Eigenvalues and Canonical Forms

λ is an eigenvalue of A with eigenvector $x \neq 0$ if $Ax = \lambda x$, equivalently $\det(A - \lambda I) = 0$. The characteristic polynomial $f_A(\lambda) = \det(A - \lambda I)$ has degree n , so A has between one and n eigenpairs counting multiplicity.

A is similar to B if $B = P^{-1}AP$ for some invertible P . Similar matrices share the characteristic polynomial and the eigenvalues; eigenvectors transform by $z = P^{-1}x$.

Theorem 7.3 (Schur). *For every $A \in \mathbb{C}^{n \times n}$ there is a unitary U such that $T = U^*AU$ is upper triangular. The diagonal of T is the spectrum of A , and $\det A = \prod \lambda_i$.*

Theorem 7.4 (Principal axes). *A Hermitian matrix $A^* = A$ has n real eigenvalues and an orthonormal basis of eigenvectors. Equivalently, $U^*AU = \text{diag}(\lambda_1, \dots, \lambda_n)$ for a unitary U .*

Theorem 7.5 (SVD). *If A is $n \times m$ there are unitary V ($n \times n$) and U ($m \times m$) such that $V^*AU = F$ is rectangular-diagonal. The positive diagonal entries $\mu_1 \geq \dots \geq \mu_r > 0$ are the singular values, and $r = \text{rank}(A)$.*

A Jordan block $J_k(\lambda)$ is $k \times k$ with λ on the diagonal and 1 on the superdiagonal. Every square matrix is similar to a direct sum of Jordan blocks. For Hermitian A every block is 1×1 .

7.3 Vector and Matrix Norms

A vector norm satisfies positivity, homogeneity, and the triangle inequality. The p -norms are $\|x\|_p = (\sum |x_j|^p)^{1/p}$ for $1 \leq p < \infty$ and $\|x\|_\infty = \max |x_j|$.

Theorem 7.6. *On a finite-dimensional space any two norms N, M are equivalent: $c_1 N(x) \leq M(x) \leq c_2 N(x)$.*

A matrix norm also satisfies $\|AB\| \leq \|A\|\|B\|$ and is consistent with a vector norm, $\|Ax\| \leq \|A\|\|x\|$. The operator norm induced by a vector norm is $\|A\| = \sup_{x \neq 0} \|Ax\|/\|x\|$. In particular

$$\|A\|_1 = \max_j \sum_i |a_{ij}|, \quad \|A\|_\infty = \max_i \sum_j |a_{ij}|, \quad \|A\|_2 = \sqrt{\rho(A^*A)},$$

and $\|A\|_2 = \rho(A)$ if A is Hermitian. The spectrum is $\sigma(A) = \{\text{eigenvalues of } A\}$ and the spectral radius is $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$. For every operator norm, $\rho(A) \leq \|A\|$.

7.4 Convergence and Perturbation Theorems

Theorem 7.7. *$A^m \rightarrow 0$ if and only if $\rho(A) < 1$.*

The proof reduces A to Jordan form $J = D + N$ and expands $(D + N)^m$; the nilpotent part N dies after n factors, and $\rho(A) < 1$ forces $D^m \rightarrow 0$.

Theorem 7.8 (Neumann series). *If $\rho(A) < 1$, then $(I - A)^{-1}$ exists and $(I - A)^{-1} = \sum_{m=0}^{\infty} A^m$. If some matrix norm satisfies $\|A\| < 1$, the same conclusion holds and $\|(I - A)^{-1}\| \leq 1/(1 - \|A\|)$.*

These facts are the basis of the residual-correction and iterative methods in Chapter 8.

8 Numerical Solution of Systems of Linear Equations

We solve $Ax = b$ by direct factorization and by iteration [1]. Sparse structured systems arising from PDEs are the main application; Poisson's equation is previewed at the end and treated in full in Chapter 10.

8.1 Gaussian Elimination and LU

Elimination reduces $Ax = b$ to an upper-triangular system $Ux = g$, which is solved by back-substitution.

Theorem 8.1 (LU). *If no pivot vanishes, $A = LU$ with unit-lower-triangular L (multipliers $m_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$ below the diagonal) and upper-triangular U equal to the final eliminated matrix. Then $\det A = \prod_i u_{ii}$, and the factorization costs $\mathcal{O}(n^3)$ operations.*

8.2 Pivoting

If $a_{kk}^{(k)} = 0$, or is merely small, one interchanges rows. Partial pivoting at stage k selects $c_k = \max_{k \leq i \leq n} |a_{ik}^{(k)}|$ and swaps that row into position k . All multipliers then satisfy $|m_{ik}| \leq 1$, which limits growth and loss of significance. The row exchanges are a permutation matrix P , and the factorization becomes $PA = LU$.

Example 8.2. The system $0.729x + 0.81y + 0.9z = 0.6867$, $x + y + z = 0.8338$, $1.331x + 1.21y + 1.1z = 1$ has exact solution $(0.2245, 0.2814, 0.3279)$. Elimination without pivoting produces errors 7–16 times larger than partial pivoting.

8.3 Variants

Gauss–Jordan continues elimination into the upper triangle and converts $[A|I]$ into $[I|A^{-1}]$. Compact schemes store L and U in the array of A : Doolittle takes $\text{diag}(L) = I$, Crout takes $\text{diag}(U) = I$. If A is symmetric positive definite, the Cholesky factorization $A = LL^t$ exists and uses only about half the arithmetic of LU.

8.4 Error Analysis

For the uniquely solvable system $Ax = b$, consider $A\tilde{x} = b + r$ and $e = x - \tilde{x}$, so $Ae = r$. Operator-norm bounds give

$$\frac{\|r\|}{\|A\|\|x\|} \leq \frac{\|e\|}{\|x\|} \leq \|A^{-1}\| \frac{\|r\|}{\|x\|}.$$

Since $\|b\| \leq \|A\|\|x\|$,

$$\frac{\|e\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}, \quad \kappa(A) := \|A\|\|A^{-1}\|.$$

The condition number $\kappa(A) \geq 1$ measures the amplification of relative residual into relative error. Equivalently, a perturbation $A + E$ with $\|A^{-1}E\| < 1$ changes the solution by a factor comparable to $\kappa(A)\|E\|/\|A\|$.

8.5 Residual Correction

If $\tilde{x}$ is an approximate solution with residual $r = b - A\tilde{x}$, the error $e = x - \tilde{x}$ solves $Ae = r$. Solving this correction (often in higher precision, or with the already-computed LU factors) and setting $x^{(1)} = \tilde{x} + e$ is iterative refinement. The Neumann series of Chapter 7 explains why the iteration converges when the effective iteration matrix has spectral radius less than one.

8.6 Stationary Iterative Methods

Split $A = M - N$ and iterate $Mx^{(k+1)} = Nx^{(k)} + b$, or $x^{(k+1)} = Gx^{(k)} + c$ with $G = M^{-1}N$. Convergence for every b holds if and only if $\rho(G) < 1$.

- Jacobi: $M = \text{diag}(A)$.
- Gauss–Seidel: M is the lower-triangular part of A (including the diagonal).
- SOR: a relaxation parameter ω is introduced in the Gauss–Seidel update. Kahan’s theorem states that SOR cannot converge for $\omega \notin (0, 2)$.

Sparsity of A is decisive: only nonzero entries are stored and multiplied.

Krylov methods build approximations in $\text{span}\{r, Ar, A^2r, \dots\}$. GMRES is designed for non-symmetric systems; the iterates satisfy a residual-minimization property and a bi-orthogonality relation with the Arnoldi basis.

8.7 Gradient and Conjugate Gradient Methods

For SPD A , solving $Ax = b$ is equivalent to minimizing $g(x) = \frac{1}{2}x^tAx - b^tx$. The gradient method takes $d_k = -\nabla g(x_k) = b - Ax_k$ and an exact line search. Conjugate gradient (CG) instead generates A -orthogonal directions p_k and

$$\begin{aligned}\alpha_{k+1} &= \frac{r_k^t r_k}{p_{k+1}^t A p_{k+1}}, & x_{k+1} &= x_k + \alpha_{k+1} p_{k+1}, \\ r_{k+1} &= r_k - \alpha_{k+1} A p_{k+1}, & \beta_{k+1} &= \frac{r_{k+1}^t r_{k+1}}{r_k^t r_k}, & p_{k+1} &= r_{k+1} + \beta_{k+1} p_k.\end{aligned}$$

In exact arithmetic CG terminates in at most n steps; in practice the error after k steps satisfies

$$\|x - x_k\|_A \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x - x_0\|_A, \quad \kappa = \kappa_2(A).$$

Preconditioned CG applies the same iteration to $M^{-1}A$ with an SPD preconditioner $M \approx A$ that is cheap to invert, reducing the effective condition number.

8.8 Poisson’s Equation

The model boundary-value problem $-u'' = f$ on $(0, 1)$, or $-\nabla^2 u = f$ in a rectangle, is the first PDE that produces a large sparse linear system. On a uniform mesh of size h the second-difference stencil of Chapter 1 (Taylor) gives

$$\frac{-U_{j-1} + 2U_j - U_{j+1}}{h^2} = f(x_j)$$

in one dimension, and the five-point stencil in two dimensions. The matrix is SPD and tridiagonal (or block-tridiagonal), so CG and sparse direct methods apply directly. The full discussion of boundary conditions, uniqueness, and accuracy is Chapter 10.

9 The Matrix Eigenvalue Problem

The task is to compute eigenvalues and eigenvectors of a square matrix A [1]. Location theorems and perturbation bounds come first; computational methods follow.

9.1 Location, Error, and Stability

From Chapter 7, $\max_{\lambda \in \sigma(A)} |\lambda| \leq \|A\|$ for any matrix norm. For $A = (a_{ij})$ write $r_i = \sum_{j \neq i} |a_{ij}|$ and let Z_i be the closed disk of center a_{ii} and radius r_i .

Theorem 9.1 (Gerschgorin). *Every eigenvalue of $A \in \mathbb{C}^{n \times n}$ lies in $\bigcup_i Z_i$. If m of the disks form a connected set S disjoint from the rest, then S contains exactly m eigenvalues, counted with multiplicity.*

The same statement holds with column sums, because A and A^t share eigenvalues.

Example 9.2. For $A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & -4 \end{pmatrix}$ the eigenvalues lie in $|\lambda - 4| \leq 1$, $|\lambda| \leq 2$, and $|\lambda + 4| \leq 2$.

Theorem 9.3 (Bauer–Fike). *Suppose $P^{-1}AP = D = \text{diag}(\lambda_i)$. If λ is an eigenvalue of $A + E$, then $\min_i |\lambda - \lambda_i| \leq \|P\| \|P^{-1}\| \|E\|$.*

If A is Hermitian one may take P unitary, so $\min_i |\lambda - \lambda_i| \leq \|E\|_2$: small perturbations of symmetric matrices move eigenvalues by at most the perturbation size. The Wielandt–Hoffman theorem sharpens this to $(\sum_i (\tilde{\lambda}_i - \lambda_i)^2)^{1/2} \leq \|E\|_F$ when both matrices are Hermitian and the eigenvalues are equally ordered.

For a computed eigenpair $(\tilde{\lambda}, \tilde{x})$ of a symmetric A , the residual $r = A\tilde{x} - \tilde{\lambda}\tilde{x}$ yields a computable bound on $|\tilde{\lambda} - \lambda|$. Nonsymmetric eigenvalues can be far more sensitive: a small departure from normality (a large $\|P\| \|P^{-1}\|$) makes the Bauer–Fike factor large even if $\|E\|$ is tiny.

9.2 The Power Method

To extract a dominant eigenpair, assume $|\lambda_1| > |\lambda_2| \geq \dots$ and iterate $w^{(m)} = Aw^{(m-1)}$, normalized at each step. Then

$$\lambda^{(m)} = \frac{u^T Aw^{(m)}}{u^T w^{(m)}} \rightarrow \lambda_1$$

with error $\mathcal{O}((\lambda_2/\lambda_1)^m)$ for a generic starting vector and a generic linear functional u .

Three accelerations are standard.

- **Translation.** Replace A by $A - \sigma I$ to change the ratio of the two largest eigenvalues.
- **Aitken.** The linearly convergent sequence $\lambda^{(m)}$ is extrapolated exactly as in Chapter 2.
- **Rayleigh–Ritz.** For a unit vector w , the Rayleigh quotient w^*Aw approximates λ_1 with error $\mathcal{O}((\lambda_2/\lambda_1)^{2m})$ on the power iterates of a Hermitian matrix.

Inverse iteration $(A - \sigma I)^{-1}$ applied to a shift near a desired eigenvalue produces the same theory for the eigenvalue closest to σ .

9.3 Householder Transformations

A Householder reflector $I - 2ww^t$ ($\|w\|_2 = 1$) is orthogonal and symmetric. Given a vector b , one can choose w so that the reflector maps b onto a multiple of e_1 . Repeating the construction on trailing submatrices reduces a general matrix to upper Hessenberg form (tridiagonal, if A is symmetric) in $\mathcal{O}(n^3)$ orthogonal similarity transformations, which is the standard prelude to the QR algorithm.

9.4 The QR Method

One step of the QR algorithm factors $A_k = Q_k R_k$ (orthogonal times upper triangular) and sets $A_{k+1} = R_k Q_k = Q_k^* A_k Q_k$. The matrices remain similar, hence isospectral. Under mild hypotheses A_k converges to upper triangular (or diagonal, if A is symmetric), revealing the eigenvalues. Shifts, as in the power method, accelerate convergence to individual eigenvalues. In practice QR is run on the Hessenberg reduction, so each step is $\mathcal{O}(n^2)$ rather than $\mathcal{O}(n^3)$.

9.5 Least Squares and the SVD

For a (possibly rectangular) system $Ax = b$ the least-squares problem is to minimize $\|Ax - b\|_2$. If $A = VFU^*$ is the SVD of Chapter 7, the minimizer is obtained by inverting the nonzero singular values and setting the components along the kernel to zero:

$$x = \sum_{\mu_i > 0} \frac{v_i^* b}{\mu_i} u_i.$$

This is the numerically stable form of the normal equations $A^* Ax = A^* b$, and it is the discrete counterpart of the L^2 projection already used in Chapter 4.

Part IV

Numerical Partial Differential Equations

10 Finite Differences and Elliptic Boundary Value Problems

Chapter 8 already produced the Poisson matrix. Here the same discretization is developed from Taylor's theorem (Chapter 1) and then equipped with boundary conditions [10].

10.1 Basic Difference Operators

The core idea of the finite difference method is to replace derivatives by values on a grid. Let u be smooth and let h be a uniform spacing. The first-order operators are

$$D_+ u(x_0) = \frac{u(x_0 + h) - u(x_0)}{h}, \quad (3)$$

$$D_- u(x_0) = \frac{u(x_0) - u(x_0 - h)}{h}, \quad (4)$$

$$D_0 u(x_0) = \frac{u(x_0 + h) - u(x_0 - h)}{2h} = \frac{1}{2}(D_+ u(x_0) + D_- u(x_0)). \quad (5)$$

As $h \rightarrow 0$ each operator converges to $u'(x_0)$.

10.2 Truncation Error

Taylor expansion (Theorem 1.4) of $u(x + h)$ gives

$$u'(x) = \frac{u(x + h) - u(x)}{h} - \frac{h}{2} u''(x) + \mathcal{O}(h^2),$$

so $D_{\pm}$ are first-order. The odd terms cancel in D_0 , and the centered difference is second-order: $u'(x) = \frac{1}{2}(D_+ u + D_- u) + \mathcal{O}(h^2)$.

10.3 Undetermined Coefficients

To approximate $u'(x)$ from the one-sided stencil $\{x, x-h, x-2h\}$, write $Du = au(x) + bu(x-h) + cu(x-2h)$ and match Taylor coefficients:

$$a + b + c = 0, \tag{6}$$

$$b + 2c = -1/h, \tag{7}$$

$$b + 4c = 0. \tag{8}$$

The unique solution $a = 3/(2h)$, $b = -2/h$, $c = 1/(2h)$ produces the second-order backward formula

$$Du(x) = \frac{1}{2h}(3u(x) - 4u(x-h) + u(x-2h)) + \mathcal{O}(h^2),$$

which is the standard tool for Neumann data at a boundary.

10.4 Second Derivatives and the Laplacian

Applying D_+D_- yields the centered second difference

$$D^2u(x) = \frac{1}{h^2}[u(x-h) - 2u(x) + u(x+h)],$$

which is $\mathcal{O}(h^2)$ and is the stencil of the one-dimensional Poisson and heat equations. In two dimensions the five-point Laplacian

$$\nabla_h^2 u(x, y) = \frac{u(x-h, y) + u(x+h, y) + u(x, y-h) + u(x, y+h) - 4u(x, y)}{h^2}$$

is likewise $\mathcal{O}(h^2)$.

10.5 Heat Equation and Boundary Conditions

The heat equation on a domain Ω is $\partial_t u = \nabla \cdot (\kappa \nabla u) + f$. If κ is constant this reduces to $\partial_t u = \kappa \nabla^2 u + f$. A well-posed problem needs

- an initial condition $u(x, 0) = u^0(x)$ on Ω ;
- Dirichlet data $u = g$ on (part of) $\partial\Omega$;
- or Neumann data $\partial u / \partial n = \sigma$ (insulation if $\sigma = 0$).

10.6 The Discrete Poisson Problem

At steady state one solves $-\kappa \nabla^2 u = f$. On $(0, 1)$ with $u(0) = \alpha$, $u(1) = \beta$ and $h = 1/(n+1)$, $x_j = jh$, the interior equations are

$$\frac{1}{h^2}(-U_{j-1} + 2U_j - U_{j+1}) = f(x_j).$$

The global system $AU = F$ has the SPD tridiagonal matrix

$$A = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 \end{bmatrix},$$

so the Cholesky and CG methods of Chapter 8 apply directly.

10.7 Neumann Conditions

Three standard treatments of $u'(0) = \sigma$ illustrate the tension between accuracy and matrix structure.

First-order one-sided difference. $(U_1 - U_0)/h = \sigma$ is easy but only $\mathcal{O}(h)$, and it reduces the global order to one.

Ghost node. A fictitious value U_{-1} is defined by $(U_1 - U_{-1})/(2h) = \sigma$, hence $U_{-1} = U_1 - 2h\sigma$. Substituting into $D^2U_0 = f(x_0)$ produces

$$\frac{1}{h}(-U_0 + U_1) = \sigma + \frac{h}{2}f(x_0),$$

which retains $\mathcal{O}(h^2)$ and the sparse symmetric pattern.

One-sided second-order formula. The three-point stencil of this chapter, $(-3U_0 + 4U_1 - U_2)/(2h) = \sigma$, is also $\mathcal{O}(h^2)$ but destroys symmetry of A .

10.8 Uniqueness and Singularity

A pure Neumann problem makes A singular: the constant vector lies in $\ker A$. If $\hat{u}$ is a solution then so is $\hat{u} + c$. The continuous problem is determined only up to a constant, and the discrete system needs an anchoring condition (for example, fixing U at one node) before the solvers of Chapter 8 can be applied.

11 Diffusion Equations and Parabolic Problems

The time-stepping theory of Chapter 6 and the spatial stencil of Chapter 10 combine to give schemes for the heat equation [10].

11.1 The Heat Equation

The one-dimensional heat equation

$$u_t = \kappa u_{xx}, \quad x \in (0, 1), \quad t > 0,$$

with $\kappa > 0$, is the model parabolic problem. On a grid of size $h = \Delta x$ and time step $k = \Delta t$ we write $U_j^n \approx u(x_j, t_n)$. Method-of-lines reduction produces the ODE $U' = (\kappa/h^2)AU$ in the interior, which may be integrated by any scheme of Chapter 6.

11.2 Crank–Nicolson

Forward Euler in time is limited by the parabolic CFL restriction $k \leq h^2/(2\kappa)$. Crank–Nicolson averages the explicit and implicit spatial operators,

$$\frac{U_j^{n+1} - U_j^n}{k} = \frac{\kappa}{2}(D_x^2 U_j^n + D_x^2 U_j^{n+1}),$$

where $D_x^2 U_j = (U_{j-1} - 2U_j + U_{j+1})/h^2$. With the diffusion number $r = \kappa k/(2h^2)$ this is the tridiagonal system

$$-rU_{j-1}^{n+1} + (1 + 2r)U_j^{n+1} - rU_{j+1}^{n+1} = rU_{j-1}^n + (1 - 2r)U_j^n + rU_{j+1}^n.$$

It is the trapezoidal rule of Chapter 6 applied to the method-of-lines ODE, and the linear solve at each step uses Chapter 8.

Taylor expansion about $(x_j, t_{n+1/2})$ produces the local truncation error

$$\tau_j^n = \left[u_t + \frac{k^2}{24} u_{ttt} + \mathcal{O}(k^4) \right] - \kappa \left[u_{xx} + \frac{h^2}{12} u_{xxx} + \mathcal{O}(h^4) \right].$$

Since $u_t = \kappa u_{xx}$, the scheme is $\mathcal{O}(k^2 + h^2)$.

11.3 Von Neumann Stability

A Fourier mode $U_j^n = g^n e^{i\xi jh}$ substituted into Crank–Nicolson yields the amplification factor [4, 10]

$$g(\xi) = \frac{1 - 4r \sin^2(\xi h/2)}{1 + 4r \sin^2(\xi h/2)}.$$

Thus $|g| \leq 1$ for every $r > 0$: the method is unconditionally stable, as predicted by the A-stability of the trapezoidal rule. As $r \rightarrow \infty$, however, $g \rightarrow -1$, so excessively large steps produce slowly decaying oscillations (lack of L-stability) [6]. The same Fourier test is an eigenvalue computation of the type treated in Chapter 9.

12 Advection Equations and Hyperbolic Systems

The linear advection equation is the model hyperbolic problem [9]. Unlike the heat equation, its domain of dependence is a finite characteristic, and stability is controlled by a CFL condition rather than a parabolic step restriction.

12.1 The Linear Advection Equation

$$u_t + au_x = 0, \quad u(x, 0) = \eta(x)$$

has the exact solution $u(x, t) = \eta(x - at)$. The equation is the standard test for numerical dissipation and phase error.

12.2 Classical Finite Difference Schemes

Write $\nu = ak/h$ for the Courant number. The CFL condition $|\nu| \leq 1$ requires that the numerical stencil contain the domain of dependence [4].

Upwind method. If $a > 0$, information travels to the right and

$$U_j^{n+1} = U_j^n - \nu(U_j^n - U_{j-1}^n)$$

is first-order, $\mathcal{O}(k + h)$. Modified-equation analysis shows that the scheme is consistent with $u_t + au_x = \frac{a^2 h}{2}(1 - \nu)u_{xx}$: the leading error is numerical diffusion [10].

Lax–Wendroff. The Taylor step $u(x, t + k) = u + ku_t + \frac{1}{2}k^2u_{tt}$ together with $u_t = -au_x$ and $u_{tt} = a^2u_{xx}$ produces

$$U_j^{n+1} = U_j^n - \frac{\nu}{2}(U_{j+1}^n - U_{j-1}^n) + \frac{\nu^2}{2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n),$$

which is $\mathcal{O}(k^2 + h^2)$. The modified equation now contains a third-derivative term, hence dispersive phase errors and oscillations near discontinuities.

Godunov’s theorem. No linear scheme of order two or higher can be monotonicity-preserving [9]. High-resolution nonlinear methods are developed in the next chapter.

13 Finite Volume Method and Conservation Laws

Finite differences approximate point values. The finite volume method evolves cell averages and is built so that the discrete scheme inherits the integral conservation law [9].

13.1 Integral Formulation and Numerical Fluxes

On the control volume $C_j = [x_{j-1/2}, x_{j+1/2}]$ the cell average is

$$Q_j^n \approx \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} q(x, t_n) dx.$$

Integrating $q_t + f(q)_x = 0$ over $C_j \times [t_n, t_{n+1}]$ gives the exact update

$$Q_j^{n+1} = Q_j^n - \frac{\Delta t}{\Delta x} (F_{j+1/2}^n - F_{j-1/2}^n).$$

The numerical flux $F_{j\pm 1/2}^n$ approximates the time-averaged physical flux at the interface and is typically obtained by solving a local Riemann problem.

13.2 High-Resolution Schemes and Flux Limiters

Godunov’s theorem (Chapter 12) forbids a linear scheme that is both second-order and monotone. High-resolution fluxes blend a monotone low-order flux F_L (upwind) with a higher-order flux F_H (Lax–Wendroff) [7]:

$$F_{j-1/2}^n = F_L + \Phi(\theta_{j-1/2})(F_H - F_L).$$

The smoothness sensor (for $a > 0$) is the consecutive-slope ratio

$$\theta_{j-1/2} = \frac{Q_j^n - Q_{j-1}^n}{Q_{j+1}^n - Q_j^n}.$$

A scheme is total-variation diminishing if $TV(Q^{n+1}) \leq TV(Q^n)$ [7]. This holds when Φ lies in the Sweby region [11]. Standard limiters are

$$\Phi(\theta) = \max(0, \min(1, \theta)) \quad (\text{minmod}), \quad (9)$$

$$\Phi(\theta) = \max(0, \min(1, 2\theta), \min(2, \theta)) \quad (\text{superbee}), \quad (10)$$

$$\Phi(\theta) = \max\left(0, \min\left(\frac{1+\theta}{2}, 2, 2\theta\right)\right) \quad (\text{MC}), \quad (11)$$

$$\Phi(\theta) = \frac{\theta + |\theta|}{1 + |\theta|} \quad (\text{van Leer}). \quad (12)$$

Minmod is the most dissipative; superbee is the sharpest at shocks; MC is a common compromise; van Leer is smooth and well-suited to steady problems.

14 The Finite Element Method

Chapters 7–8 treated finite-dimensional inner-product spaces and SPD linear systems. Finite elements apply the same ideas to Sobolev spaces, with element integrals evaluated by the quadrature of Chapter 5 [5, 2].

14.1 Weak Derivatives and Sobolev Spaces

A classical solution of a second-order PDE must lie in $C^2(\Omega)$. Weak derivatives relax this. If $u, v \in L^1_{\text{loc}}(\Omega)$, then $v = D^\alpha u$ provided

$$\int_{\Omega} u D^\alpha \phi \, dx = (-1)^{|\alpha|} \int_{\Omega} v \phi \, dx \quad \text{for all } \phi \in C_c^\infty(\Omega).$$

Differentiability is transferred from u to the test function.

The Sobolev space $W^{k,p}(\Omega)$ consists of L^p functions whose weak derivatives up to order k also lie in L^p . For $p = 2$ one writes $H^k(\Omega) = W^{k,2}(\Omega)$, a Hilbert space. The H^1 norm is

$$\|u\|_{H^1(\Omega)} = \left(\int_{\Omega} (|u|^2 + |\nabla u|^2) \, dx \right)^{1/2}.$$

The subspace $H_0^1(\Omega)$ consists of H^1 functions with vanishing trace on $\partial\Omega$ and is the natural space for homogeneous Dirichlet conditions. Galerkin orthogonality in this inner product is the infinite-dimensional analogue of the L^2 projection of Chapter 4.

14.2 Variational Formulation

For the elliptic problem

$$-\nabla \cdot (A \nabla u) + b \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega,$$

multiply by $v \in H_0^1(\Omega)$ and integrate by parts:

$$\int_{\Omega} (A \nabla u \cdot \nabla v + b \cdot \nabla u v + cuv) \, dx = \int_{\Omega} f v \, dx.$$

The variational problem is: find $u \in H_0^1(\Omega)$ such that $a(u, v) = \langle f, v \rangle$ for all $v \in H_0^1(\Omega)$, where a is a continuous bilinear form and $f \in H^{-1}(\Omega)$.

14.3 Lax–Milgram and Energy Minimization

Theorem 14.1 (Lax–Milgram). *If a is bounded, $|a(u, v)| \leq \alpha \|u\|_{H^1} \|v\|_{H^1}$, and coercive, $a(u, u) \geq \beta \|u\|_{H^1}^2$ with $\beta > 0$, then the variational problem has a unique solution.*

If a is symmetric, the solution is the unique minimizer of the energy $J(v) = \frac{1}{2}a(v, v) - \langle f, v \rangle$.

14.4 Galerkin Approximation

Restricting to a finite-element subspace $V_h \subset H_0^1(\Omega)$, Céa's lemma gives

$$\|u - u_h\|_{H^1} \leq \frac{\alpha}{\beta} \inf_{v_h \in V_h} \|u - v_h\|_{H^1}.$$

A basis of compactly supported piecewise polynomials produces $u_h = \sum_j U_j \phi_j$ and the sparse SPD system $KU = F$ with $K_{ij} = a(\phi_j, \phi_i)$ and $F_i = \langle f, \phi_i \rangle$. Assembly of K and F is a sum of element integrals, evaluated by the quadrature rules of Chapter 5; the solve uses Chapter 8.

15 Discontinuous Galerkin Methods

Continuous finite elements enforce C^0 continuity across faces and are poorly adapted to hyperbolic conservation laws with shocks. The discontinuous Galerkin method combines the local polynomial spaces of Chapter 14 with the numerical fluxes of Chapter 13 [8, 3].

15.1 Motivation

On each element D^k the solution is a polynomial of arbitrary degree, with no continuity required at x^k . The resulting method has a block-diagonal mass matrix, is locally conservative, and admits hp -adaptivity and explicit time stepping.

15.2 Formulation and Numerical Fluxes

A test function ϕ_i^k supported on $D^k = [x^k, x^{k+1}]$ produces the local weak form

$$\int_{D^k} \left(\partial_t u_h^k \phi_i^k - f_h^k \partial_x \phi_i^k - g \phi_i^k \right) dx = - [f_h^k \phi_i^k]_{x^k}^{x^{k+1}}.$$

The interface term is replaced by a numerical flux $F(u^-, u^+)$ as in the finite-volume update:

$$[f_h^k \phi_i^k]_{x^k}^{x^{k+1}} = F(u_h(x^{k+1,-}), u_h(x^{k+1,+})) \phi_i^k(x^{k+1}) - F(u_h(x^{k,-}), u_h(x^{k,+})) \phi_i^k(x^k).$$

The mass matrix inverts elementwise, so the semi-discrete system is an ODE of the type treated in Chapter 6.

15.3 Comparison of Discretizations

- **FDM** (Chapter 10): simple on Cartesian grids; awkward on complex geometry.
- **FVM** (Chapter 13): exactly conservative and natural for shocks; high-order accuracy on unstructured meshes is comparatively heavy.
- **FEM** (Chapter 14): strong for elliptic and parabolic problems and for unstructured geometry; less natural for strong advection.
- **DG-FEM**: high-order, locally conservative, geometry-flexible, and highly parallel; the cost is a larger number of degrees of freedom.

References

- [1] Kendall E. Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, New York, 2nd edition, 1989.
- [2] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New York, NY, 3rd edition, 2008.
- [3] Bernardo Cockburn, San-Yih Lin, and Chi-Wang Shu. Tvb runge-kutta local projection discontinuous galerkin finite element method for conservation laws. iii. one-dimensional systems. *Journal of Computational Physics*, 84(1):90–113, 1989. The critical paper that successfully coupled RK time-stepping with the spatial DG-FEM method.
- [4] Richard Courant, Kurt Friedrichs, and Hans Lewy. Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische Annalen*, 100(1):32–74, 1928. The original paper establishing the fundamental CFL condition for numerical stability.
- [5] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society (AMS), Providence, RI, 2nd edition, 2010. Mathematical foundation for Weak Derivatives, Sobolev Spaces, and the Lax-Milgram Lemma.
- [6] Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer-Verlag, Berlin, Heidelberg, 2nd edition, 1993.
- [7] Amiram Harten. High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357–393, 1983. A landmark paper that introduced the mathematical framework for High-Resolution TVD schemes.
- [8] Jan S. Hesthaven and Tim Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer, New York, NY, 2007.
- [9] Randall J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, Cambridge, UK, 2002.
- [10] Randall J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.
- [11] Peter K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM journal on numerical analysis*, 21(5):995–1011, 1984. The foundational paper defining the TVD region and standard flux limiters like Minmod and Superbee.