CHAPTER 1

FINITE DIFFERENCE

The finite difference method was one of the first numerical methods used to solve partial differential equations (PDEs). It replaces differential operators by finite differences and the PDE becomes a (finite) system of equations. Its simplicity and ease of computer implementation make it a popular choice for PDEs defined on regular geometries. One drawback is that it becomes awkward when the geometry is not regular or cannot be mapped to a regular geometry. Another disadvantage is that its error analysis is not as sharp as that of the other methods covered in this book (finite element and spectral methods). A recurring theme is that the analysis of, and properties of, discrete operators mimic those of the differential operators. Examples include integration by parts, maximum principle, energy method, Green's function and the Poincaré–Friedrichs inequality.

1.1 SECOND-ORDER APPROXIMATION FOR Δ

Consider the Poisson equation

$$-\Delta u = f \text{ on } \Omega, \qquad u = 0 \text{ on } \partial\Omega. \tag{1.1}$$

Numerical Analysis of Partial Differential Equations. By S. H. Lui

Here,

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

is the Laplacian in two dimensions. Throughout this chapter, except the sections on polar coordinates and curved boundaries, Ω is the unit square $(0,1)^2 = \{(x,y),\ 0 < x, y < 1\}$.

The Poisson equation is a fundamental equation which arises in elasticity, electromagnetism, fluid mechanics and many other branches of science and engineering. Because an explicit expression of the solution is available only in a few exceptional cases, we often rely on numerical methods to approximate the solution. The goal of the subject of numerical analysis of PDEs is to design a numerical method which approximates the solution accurately and efficiently. Roughly speaking, a method is accurate if the computed solution differs from the exact solution by an amount which goes to zero as h, a discretization parameter, goes to zero. A method is efficient if the amount of computation and storage requirement of the method do not grow quickly as a function of the size of the input data, f, in the case of the Poisson equation. The remaining pages of this book will be preoccupied with these issues and will culminate in algorithms (multigrid and domain decomposition) which are optimal in the sense that the amount of work to approximate the solution is no more than a linear function of the size of the input.

We take a uniform grid of size $h = 1/n$, where n is a positive integer. Let

$$\Omega_h = \{x_{ij} = (ih, jh),\ \ 1 \le i, j \le n-1\}$$

denote the set of interior grid points and

$$\partial\Omega_h = \{(0, jh), (1, jh), (jh, 0), (jh, 1),\ \ 1 \le j \le n-1\}$$

denote the boundary grid points. [Observe that the corner points $(0,0), (1,0), (0,1)$, $(1,1)$ are not in $\partial\Omega_h$.] Define $\overline{\Omega}_h = \Omega_h \cup \partial\Omega_h$, which consists of $(n+1)^2 - 4$ points.

The finite difference method seeks the solution of the PDE at the grid points in Ω. Specifically, the $(n-1)^2$ unknowns are $u_{ij} = u(ih, jh)$, $1 \le i, j \le n-1$. We obtain $(n-1)^2$ equations by approximating the differential equation by a finite difference approximation at each interior grid point. That is,

$$\frac{4u_{ij} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}}{h^2} = f_{ij} := f(x_{ij}).$$

(These equations will be derived later when we discuss the consistency of the scheme.) The boundary values are $u_{0j} = u_{nj} = u_{i0} = u_{in} = 0$, $1 \le i, j \le n-1$; they are known from the boundary condition. The system of linear equations is denoted by the discrete Poisson equation

$$-\Delta_h u_h = f_h.$$

Here, u_h is the vector of unknowns arranged in an order so that Δ_h is block tridiagonal:

$$u_h = [u_{11}, \ldots, u_{n-1,1}, u_{12}, \ldots, u_{n-1,2}, \ldots, u_{n-1,n-1}]^T,$$

$f_h = [f_{11}, \cdots, f_{n-1,n-1}]^T$ and

$$-\Delta_h = \frac{1}{h^2}\begin{bmatrix} T & -I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -I & T \end{bmatrix}, \quad T = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix}.$$

Both T and I above are $(n-1) \times (n-1)$ matrices with I the identity matrix. An alternative representation of Δ_h is given by the molecule

$$\frac{1}{h^2}\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}.$$

Before proceeding further, we define some norms which measure the size of vectors and matrices. First consider the infinity norm on $\mathbb{R}^N$. For any $x \in \mathbb{R}^N$, define

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

For any $N \times N$ matrix A, we claim that

$$|A|_\infty := \sup_{x \in \mathbb{R}^N \setminus 0} \frac{|Ax|_\infty}{|x|_\infty} = \max_{1 \le i \le N} \sum_{j=1}^{N} |a_{ij}|.$$

In other words, $|A|_\infty$ equals the maximum row sum of the matrix. To see this, for any $x \in \mathbb{R}^N \setminus 0$,

$$\frac{|Ax|_\infty}{|x|_\infty} = \frac{\left| \begin{bmatrix} \sum_{j=1}^N a_{1j}x_j \\ \vdots \\ \sum_{j=1}^N a_{Nj}x_j \end{bmatrix} \right|_\infty}{|x|_\infty} = \frac{\max_{1 \le i \le N} \left| \sum_{j=1}^N a_{ij}x_j \right|}{|x|_\infty} \le \max_{1 \le i \le N} \sum_{j=1}^{N} |a_{ij}|.$$

Hence $|A|_\infty \le \max_{1 \le i \le N} \sum_{j=1}^N |a_{ij}|$. To show equality, suppose $A \ne 0$ and the maximum row sum occurs on row i. Define $x_j = \text{sign}(a_{ij})$. Then $|x|_\infty = 1$ and $|Ax|_\infty = \sum_{j=1}^N |a_{ij}|$. As an application, we note that $|\Delta_h|_\infty = 8h^{-2}$.

Another useful norm is the Euclidean (or 2-) norm defined by $|x|_2^2 = x^T x$ and for matrix A,

$$|A|_2 = \sup_{x \ne 0} \frac{|Ax|_2}{|x|_2}.$$

One useful property is $|A|_2 = \sigma_1$, where σ_1^2 is the largest eigenvalue of $A^T A$. If A is symmetric, then $|A|_2 = |\lambda_1|$ while $|A^{-1}|_2 = |\lambda_N^{-1}|$, where λ_1 and λ_N are eigenvalues of A of largest and smallest magnitude, respectively.

Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ in a normed vector space X are said to be **equivalent** if there are positive constants c_i such that for every v in X, $c_1\|v\|_1 \le \|v\|_2 \le c_2\|v\|_1$. It is well known that any two norms on a finite-dimensional space are equivalent. This means that we can use $|\cdot|_2$ or $|\cdot|_\infty$, whichever is more convenient. Although c_1 and c_2 are independent of $v \in X$, they do depend on N, the dimension of the space.

Finally, we define $C^r(\overline{\Omega})$, for $0 \le r \le \infty$, as the space of r times continuously differentiable functions on $\overline{\Omega}$ with norm

$$\|v\|_{C^r(\overline{\Omega})} = \max_{0\le|\alpha|\le r} \sup_{x\in\overline{\Omega}} |D^\alpha v(x)|.$$

Here $D^\alpha v$ denotes any derivative of v up to rth order, where α is a multi-index. For instance, if $\alpha = (2,1)$, then $D^\alpha v$ denotes the 3rd derivative v_{xxy}. We abbreviate $C^0(\overline{\Omega})$ as $C(\overline{\Omega})$. If $v \in C^r(\overline{\Omega})$, then $D^\alpha v \in C(\overline{\Omega})$ for every α such that $|\alpha| := \alpha_1 + \alpha_2 \le r$. Define

$$\|v\|^*_{C^r(\overline{\Omega})} = \max_{|\alpha|=r} \sup_{x\in\overline{\Omega}} |D^\alpha v(x)|. \tag{1.2}$$

This is not a norm but it comes up frequently in the analysis of finite difference schemes. In the one-dimensional case,

$$\|v\|_{C^2([0,1])} = \max_{x\in[0,1]} (|v(x)|, |v'(x)|, |v''(x)|), \qquad \|v\|^*_{C^2([0,1])} = \max_{x\in[0,1]} |v''(x)|.$$

In this book, all functions are real unless otherwise specified. Also, c appears in many places and denotes a positive constant whose value may differ in different occurrences. The same remark applies to c_1, c_2, C, C_1, C_2, etc.

Next, we shall demonstrate several properties of $-\Delta_h$. These are crucial in estimating the error of the approximate solution.

Positive Definiteness

Let $(\cdot,\cdot)$ be the L^2 inner product defined by

$$(u,v) = \int_\Omega uv$$

for all square-integrable functions u and v defined on Ω. It is well known that $-\Delta$ is a self-adjoint and positive definite operator. This means that for all smooth functions u, v, w vanishing on $\partial\Omega$ with $w \not\equiv 0$,

$$(-\Delta u, v) = (u, -\Delta v) \text{ and } (-\Delta w, w) > 0.$$

(A rigorous justification of self-adjointness is non-trivial since it requires checking the domain of the operator.)

We now show that the discrete operator has analogous properties. By inspection, $-\Delta_h$ is symmetric. We now show that it is positive definite. This is shown in two

different ways. The first way uses summation by parts which is a discrete integration by parts. Let v be any smooth function which vanishes at $x = 0$ and $x = 1$. Then one form of integration by parts is

$$-\int_0^1 v(x)v''(x)dx = -v(x)v'(x)\Big|_0^1 + \int_0^1 (v'(x))^2 dx = \int_0^1 (v'(x))^2 dx.$$

Now let v_h be any non-zero vector defined on Ω_h with coordinates v_{ij}. Extend its definition to be zero on $\partial\Omega_h$. Then

$$\begin{aligned}
-h^2 v_h^T \Delta_h v_h &= \sum_{i,j=1}^{n-1} v_{ij}(4v_{ij} - v_{i+1,j} - v_{i-1,j} - v_{i,j+1} - v_{i,j-1}) \\
&= \sum_{i,j=1}^{n-1} v_{ij}(v_{ij} - v_{i+1,j}) + v_{ij}(v_{ij} - v_{i-1,j}) \\
&\quad + v_{ij}(v_{ij} - v_{i,j+1}) + v_{ij}(v_{ij} - v_{i,j-1}) \\
&= \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} v_{ij}(v_{ij} - v_{i+1,j}) + \sum_{i=0}^{n-2}\sum_{j=1}^{n-1} v_{i+1,j}(v_{i+1,j} - v_{ij}) \\
&\quad + \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} v_{ij}(v_{ij} - v_{i,j+1}) + \sum_{i=1}^{n-1}\sum_{j=0}^{n-2} v_{i,j+1}(v_{i,j+1} - v_{ij}) \\
&= \sum_{i=0}^{n-1}\sum_{j=1}^{n-1}(v_{ij} - v_{i+1,j})^2 + \sum_{i=1}^{n-1}\sum_{j=0}^{n-1}(v_{ij} - v_{i,j+1})^2.
\end{aligned}$$

Define the discrete forward difference operators

$$\delta_h^x v_{ij} = \frac{v_{i+1,j} - v_{ij}}{h}, \qquad \delta_h^y v_{ij} = \frac{v_{i,j+1} - v_{ij}}{h}$$

and

$$\delta_h^x v_h = [\delta_h^x v_{01}, \ldots, \delta_h^x v_{n-1,1}, \delta_h^x v_{02}, \ldots, \delta_h^x v_{n-1,2}, \ldots, \delta_h^x v_{0,n-1}, \ldots, \delta_h^x v_{n-1,n-1}]^T,$$

$$\delta_h^y v_h = [\delta_h^y v_{10}, \ldots, \delta_h^y v_{n-1,0}, \delta_h^y v_{11}, \ldots, \delta_h^y v_{n-1,1}, \ldots, \delta_h^y v_{1,n-1}, \ldots, \delta_h^y v_{n-1,n-1}]^T.$$

The above calculation can be compactly summarized by

$$-v_h^T \Delta_h v_h = |\delta_h^x v_h|_2^2 + |\delta_h^y v_h|_2^2 \tag{1.3}$$

for any vector v_h. This is **summation by parts**, which is clearly a discrete version of integration by parts. If v_h is not the zero vector then $-v_h^T \Delta_h v_h > 0$ and so $-\Delta_h$ is positive definite.

A second method to show positive definiteness is by **discrete Fourier analysis**. An explicit spectral decomposition for Δ_h is available:

$$-\Delta_h q^{(ij)} = \lambda_{ij} q^{(ij)}, \qquad 1 \le i, j \le n-1, \tag{1.4}$$

where

$$\lambda_{ij} = \frac{4}{h^2}\left[\sin^2\left(\frac{i\pi h}{2}\right) + \sin^2\left(\frac{j\pi h}{2}\right)\right], \qquad q^{(ij)} = 2h \sin i\pi x \, \sin j\pi y,$$

with $(x, y) \in \Omega_h$. Note that the eigenvectors $\{q^{(ij)}\}$ are orthonormal. Since all the eigenvalues are positive, $-\Delta_h$ is positive definite.

The proof of the eigenvalue relation (1.4) is by a straightforward calculation. The (r, s) component of the vector $-h/2\Delta_h q^{(ij)}$ is

$$\begin{aligned}
&2 \sin ir\pi h \, \sin js\pi h - \sin(r+1)i\pi h \, \sin js\pi h - \sin(r-1)i\pi h \, \sin js\pi h \\
&\qquad + \text{ finite difference of second } y \text{ derivative} \\
&= 2 \sin ir\pi h \, \sin js\pi h - 2 \sin ir\pi h \, \cos i\pi h \, \sin js\pi h + \cdots \\
&= 2 \sin ir\pi h \, \sin js\pi h \, (1 - \cos i\pi h) + \cdots \\
&= 4 \sin ir\pi h \, \sin js\pi h \left[\sin^2\left(\frac{i\pi h}{2}\right) + \sin^2\left(\frac{j\pi h}{2}\right)\right] \\
&= \frac{h}{2}\lambda_{ij} q_{rs}^{(ij)}.
\end{aligned}$$

Since $-\Delta_h$ is symmetric positive definite, $|\Delta_h|_2$ is equal to its largest eigenvalue while $|\Delta_h^{-1}|_2$ is equal to the inverse of its smallest eigenvalue. Hence we obtain $|\Delta_h|_2 = 8h^{-2}\sin^2((n-1)\pi h/2) \leq 8h^{-2}$ and

$$|\Delta_h^{-1}|_2 = \frac{h^2}{8\sin^2(\pi h/2)} \leq \frac{1}{8} \tag{1.5}$$

if $h \leq 1$. This is because the middle expression in (1.5) is an increasing function of h for $0 < h \leq 1$. The **condition number** of $-\Delta_h$, which is defined as the ratio of the largest eigenvalue to the smallest eigenvalue of $-\Delta_h$, is bounded above by a constant multiple of h^{-2}. The condition number is a fundamental quantity in numerical analysis and we shall encounter it many times in this book.

Consistency

Let r be a positive integer. The discretization $-\Delta_h$ is said to be **consistent of order r if**

$$|\Delta_h R_h v - R_h \Delta v|_\infty \leq c \, \|v\|^*_{C^{r+2}(\overline{\Omega})} h^r$$

holds for all $v \in C^{r+2}(\overline{\Omega})$ vanishing on $\partial\Omega$ and c is a constant independent of h and v. The restriction R_h is an operator from $C(\overline{\Omega})$ to the $(n-1)^2$-vector whose components are $v(x_{ij})$, where $x_{ij} \in \Omega_h$:

$$(R_h v)(P) = v(P), \qquad P \in \Omega_h.$$

In other words, R_h restricts a continuous function to the grid points. Consistency is, roughly speaking, a measure of how good the discrete operator approximates the continuous operator at the grid points.

Theorem 1.1 **Second-order consistency of Δ_h.** For any $v \in C^4(\overline{\Omega})$ vanishing on $\partial\Omega$, $|\Delta_h R_h v - R_h \Delta v|_\infty \le c\,\|v\|^*_{C^4(\overline{\Omega})} h^2$.

Proof: Let $v_{ij} = v(ih, jh)$. By Taylor's Theorem,

$$v_{i\pm1,j} = v_{ij} \pm \frac{\partial v(x_{ij})}{\partial x} h + \frac{\partial^2 v(x_{ij})}{\partial x^2} \frac{h^2}{2} \pm \frac{\partial^3 v(x_{ij})}{\partial x^3} \frac{h^3}{6} + \frac{\partial^4 v(x^{\pm}_{ij})}{\partial x^4} \frac{h^4}{24}$$

for some $x^{\pm}_{ij} \in \Omega$. Adding the above two equations and rearranging, we obtain

$$\frac{-2v_{ij} + v_{i+1,j} + v_{i-1,j}}{h^2} = \frac{\partial^2 v(x_{ij})}{\partial x^2} + E_1, \qquad \|E_1\|^*_{C^4(\overline{\Omega})} \le c\,\|v\|^*_{C^4(\overline{\Omega})} h^2.$$

Adding a similar equation for the second y derivative, we have

$$\Delta_h R_h v = R_h \Delta v + E, \qquad \|E\|^*_{C^4(\overline{\Omega})} \le C\,\|v\|^*_{C^4(\overline{\Omega})} h^2.$$

This is the desired result. □

The domain of Δ_h is $\mathbb{R}^N$, where $N = (n-1)^2$. It can also be thought of as the set of real-valued functions defined on Ω_h. For a vector v_h, we sometimes use the notation $v_h(P)$ to denote the component of v_h corresponding to the point $P \in \Omega_h$. Yet another equivalent description of the domain of Δ_h is the set of real-valued functions defined on $\overline{\Omega}_h$ vanishing on $\partial\Omega_h$. Sometimes it is convenient if the domain can be expanded to include vectors/functions which need not vanish at the boundary. For such a vector v_h, define

$$-(\overline{\Delta}_h v_h)_{ij} = \frac{4v_{ij} - v_{i+1,j} - v_{i-1,j} - v_{i,j+1} - v_{i,j-1}}{h^2}, \qquad 1 \le i,j \le n-1. \tag{1.6}$$

That is, the domain of $\overline{\Delta}_h$ is $\mathbb{R}^{(n+1)^2-4}$, or equivalently, the set of functions defined on $\overline{\Omega}_h$, and the range is the set of N-vectors corresponding to values of the discrete Laplacian applied to v_h at Ω_h. For instance,

$$-(\overline{\Delta}_h v_h)_{11} = \frac{4v_{11} - v_{21} - v_{01} - v_{12} - v_{10}}{h^2},$$

and so the boundary values v_{01}, v_{10} also contribute.

A more general definition of consistency does not require the function v to vanish on the boundary. The condition for second-order consistency in this case is

$$|\overline{\Delta}_h \overline{R}_h v - R_h \Delta v|_\infty \le c\,\|v\|^*_{C^4(\overline{\Omega})} h^2,$$

where $\overline{R}_h$ restricts v onto $\overline{\Omega}_h$. In case v vanishes on $\partial\Omega_h$, then $\Delta_h R_h v = \overline{\Delta}_h \overline{R}_h v$.

Stability

The discretization $-\Delta_h$ is said to be **stable** with respect to $|\cdot|_\infty$ if $|\Delta_h^{-1}|_\infty$ is bounded independently of h. Thus stability implies that for the scheme $-\Delta_h u_h = f_h := R_h f_h$, the ratio $|u_h|_\infty$ over $|f_h|_\infty$ is bounded independently of h and is certainly a very desirable property. It also means that a small change in the data leads to a small change in the solution for all small h. Suppose the data f_h is perturbed by δ. The new solution is $v_h = -\Delta_h^{-1}(f_h + \delta)$. Then $u_h - v_h = \Delta_h^{-1}\delta$ and consequently $|u_h - v_h|_\infty \leq |\Delta_h^{-1}|_\infty |\delta|_\infty$. Thus a stable scheme means that the change in the solution is bounded by a constant multiple of the size of the perturbation δ. We shall discuss a technique called the discrete maximum principle which can be used to show stability of a scheme.

Discrete Maximum Principle

The (weak) maximum principle states that if $u \in C^2(\Omega) \cap C(\overline{\Omega})$ and $\Delta u \geq 0$ on Ω, then the maximum value of u is achieved on the boundary. Similarly, if $\Delta u \leq 0$, then the minimum value of u is achieved on the boundary. For the Poisson equation (1.1), this says that if u is smooth and $f \leq 0$ on Ω, then $u \leq 0$ on Ω. This information is obtained without first solving the PDE. We shall show below that a similar statement holds for the discrete problem.

If w_h is a vector, by $w_h \geq 0$, we mean that each component of w_h is non-negative. Now we are ready for:

Theorem 1.2 Discrete Maximum Principle. If $\overline{\Delta}_h v_h \geq 0$ on Ω_h, then the maximum value of v_h is attained on $\partial\Omega_h$.

Proof: Suppose $\overline{\Delta}_h v_h \geq 0$. We emphasize that v_h does not vanish on $\partial\Omega_h$ in general. If the maximum value of v_h occurs at a boundary point, then there is nothing to prove. Otherwise, suppose the maximum value of v_h occurs at $x_{ij} \in \Omega_h$. Now

$$\frac{-4v_{ij} + v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1}}{h^2} \geq 0$$

implies that

$$v_{ij} \leq \frac{v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1}}{4} \leq \frac{v_{ij} + v_{ij} + v_{ij} + v_{ij}}{4} = v_{ij}$$

since v_{ij} is maximum. It can now be inferred that $v_{ij} = v_{i+1,j} = v_{i-1,j} = v_{i,j+1} = v_{i,j-1}$. Repeatedly applying this argument, we conclude that v_h is constant on $\overline{\Omega}_h$ and, in particular, also achieves the maximum on $\partial\Omega_h$. □

The theorem below states that our scheme $-\Delta_h u_h = f_h$ is indeed stable. Actually, this has already been proven for the 2-norm; see (1.5). Below, **1** stands for the vector of all ones.

Theorem 1.3 Stability of Δ_h. $|\Delta_h^{-1}|_\infty \leq 1/8$.

Proof: Given any f_h defined on Ω_h, define $u_h = -\Delta_h^{-1} f_h$. The goal is to show that

$$\frac{|\Delta_h^{-1} f_h|_\infty}{|f_h|_\infty} = \frac{|u_h|_\infty}{|f_h|_\infty} \le \frac{1}{8}.$$

Define $w(x,y) = [(x-\frac{1}{2})^2 + (y-\frac{1}{2})^2]/4$ and define the vector w_h on $\overline{\Omega}_h$ by

$$w_{ij} = w(ih, jh) = \left[\left(ih - \frac{1}{2}\right)^2 + \left(jh - \frac{1}{2}\right)^2\right] \Big/ 4. \tag{1.7}$$

We now show that $\overline{\Delta}_h w_h = \mathbf{1}$. It is easy to see that $\Delta w = \mathbf{1}$, $w_h = \overline{R}_h w$ and $\mathbf{1} - \overline{\Delta}_h w_h = R_h \Delta w - \overline{\Delta}_h \overline{R}_h w = 0$ since the consistency error involves fourth derivatives of w which all vanish. Hence $\overline{\Delta}_h w_h = \mathbf{1}$. Of course, one can also verify this equality by a direct calculation. (The terms $1/2$ in w give the smallest estimate of $|\Delta_h^{-1}|_\infty$.)

Note that w_h is non-negative everywhere and its maximum occurs along $\partial\Omega_h$ with $|w_h|_\infty = 1/8$. Now

$$\overline{\Delta}_h(|f_h|_\infty w_h + u_h) = |f_h|_\infty \mathbf{1} - f_h \ge 0.$$

By the discrete maximum principle, $|f_h|_\infty w_h + u_h$ achieves its maximum on $\partial\Omega_h$. For $(ih, jh) \in \Omega_h$, recalling that u_h vanishes on $\partial\Omega_h$,

$$u_{ij} \le |f_h|_\infty w_{ij} + u_{ij} \le |f_h|_\infty |w_h|_\infty = \frac{|f_h|_\infty}{8}.$$

Similarly,

$$\overline{\Delta}_h(|f_h|_\infty w_h - u_h) = |f_h|_\infty \mathbf{1} + f_h \ge 0,$$

which implies that

$$-u_{ij} \le |f_h|_\infty w_{ij} - u_{ij} \le |f_h|_\infty |w_h|_\infty = \frac{|f_h|_\infty}{8}.$$

These inequalities together mean $|u_h|_\infty \le |f_h|_\infty/8$ and hence the result. □

Convergence

The scheme $-\Delta_h u_h = f_h := R_h f$ is said to be **convergent of order r** if

$$|R_h u - u_h|_\infty \le c \|u\|^*_{C^{r+2}(\overline{\Omega})} h^r$$

holds for some constant c independent of h and u. Here u is the exact solution of the Poisson equation and is assumed to lie in $C^{r+2}(\overline{\Omega})$.

The following is one of the main results of this chapter–our scheme for the Poisson equation is convergent of order 2.

Theorem 1.4 **Second-order convergence of Δ_h.** Suppose the solution u of (1.1) is in $C^4(\overline{\Omega})$. Then $|u_h - R_h u|_\infty \leq c \|u\|^*_{C^4(\overline{\Omega})} h^2$.

Proof: Let $e_h = u_h - R_h u$. Then

$$\begin{aligned} \Delta_h e_h &= \Delta_h u_h - \Delta_h R_h u \\ &= -f_h - \Delta_h R_h u \\ &= -R_h f - \Delta_h R_h u \\ &= R_h \Delta u - \Delta_h R_h u, \end{aligned}$$

and thus $e_h = \Delta_h^{-1}(R_h \Delta u - \Delta_h R_h u)$. Since the scheme is stable and consistent,

$$|e_h|_\infty \leq \frac{c}{8} \|u\|^*_{C^4(\overline{\Omega})} h^2.$$

□

Let $u, v \in C(\Omega)$ and $u_h = R_h u$, $v_h = R_h v$. Define the **discrete L^2 inner product**

$$\langle u_h, v_h \rangle_h = h^2 \sum_{i,j=1}^{n-1} u_{ij} v_{ij} \tag{1.8}$$

and the **discrete L^2 norm** of v_h by

$$|v_h|_h = \langle v_h, v_h \rangle_h^{1/2} = h\,|v_h|_2. \tag{1.9}$$

As the name implies, this discrete norm approximates the L^2 norm of v. An analogous convergence result in the discrete L^2 norm is

$$|e_h|_h \leq c \|u\|^*_{C^4(\overline{\Omega})} h^2. \tag{1.10}$$

In the proof of Theorem 1.4, it can be seen that consistency and stability imply convergence. This is a general principle which occurs frequently in numerical analysis. It is of great practical importance since consistency is usually relatively easy to check by a Taylor's expansion while stability often follows from the analogous property of the differential operator. These two properties are enough to yield convergence, which is ultimately what we want to show and is non-trivial to show directly. An abstract version of this principle will be given at the end of the next chapter. See Exercise E1 for a converse of this theorem.

Note that a stable method need not be convergent. For instance, if the discrete Poisson equation is solved using the identity operator (i.e., define $u_h = I f_h = f_h$), then this method is stable but not convergent.

Discrete Energy Method

Thus far, we have discussed two methods to show stability: the discrete maximum principle and discrete Fourier analysis. In the latter case, stability is with respect to

$|\cdot|_2$; see (1.5). The discrete energy method is a third alternative. We first illustrate the energy method for the continuous problem $-\Delta u = f$ for a smooth function u which vanishes on $\partial\Omega$. Multiply the PDE by u and then integrate by parts to obtain

$$c\|u\|^2 \le \int_\Omega |\nabla u|_2^2 = \int_\Omega fu \le \|f\|\,\|u\|, \tag{1.11}$$

where

$$\|u\|^2 = \int_\Omega u^2$$

is the square of the L^2 norm of the function u. The first inequality of (1.11) is known as the **Poincaré–Friedrichs inequality** and will appear again in the next and subsequent chapters. The second inequality in (1.11) is the Cauchy–Schwarz inequality. From (1.11), we immediately obtain

$$\|u\| \le \frac{\|f\|}{c}.$$

Using the summation by parts formula (1.3) and the **discrete Poincaré–Friedrichs inequality**

$$2\,|v_h|_2^2 \le |\delta_h^x v_h|_2^2 + |\delta_h^y v_h|_2^2,$$

which will be shown later, we have

$$2\,|v_h|_2^2 \le |\delta_h^x v_h|_2^2 + |\delta_h^y v_h|_2^2 = -v_h^T \Delta_h v_h,$$

meaning that the smallest eigenvalue of $-\Delta_h$ is at least 2. This follows from the **variational characterization** of λ_m, the smallest eigenvalue of a symmetric matrix A:

$$\lambda_m = \min_{x \ne 0} \frac{x^T A x}{|x|_2^2}.$$

(An analogous characterization holds for the maximum eigenvalue where min is replaced by max.) Hence the stability result $|\Delta_h^{-1}|_2 \le 1/2$ is obtained.

To show the discrete Poincaré–Friedrichs inequality, notice that for $1 \le i, j \le n-1$,

$$|v_{ij}| = |v_{ij} - v_{0j}| \le \sum_{k=0}^{i-1} |v_{k+1,j} - v_{k,j}|.$$

Recall that $v_{0j} = v_{nj} = 0$ for all j. Hence

$$\begin{aligned} v_{ij}^2 &\leq \left(\sum_{k=0}^{n-1} |v_{k+1,j} - v_{kj}| \right)^2 \\ &= h^2 \left(\sum_{k=0}^{n-1} |\delta_h^x v_{kj}| \right)^2 \\ &\leq h^2 \sum_{k=0}^{n-1} (\delta_h^x v_{kj})^2 \sum_{k=0}^{n-1} 1^2 \\ &= h \sum_{k=0}^{n-1} (\delta_h^x v_{kj})^2. \end{aligned}$$

Thus

$$\sum_{j=1}^{n-1} \sum_{i=1}^{n-1} v_{ij}^2 \leq \sum_{j=1}^{n-1} \sum_{k=0}^{n-1} (\delta_h^x v_{kj})^2$$

or $|v_h|_2^2 \leq |\delta_h^x v_h|_2^2$. Adding this to a similar inequality for $\delta_h^y v_h$ results in the desired inequality.

Discrete Green's Function

The final technique to show stability is now introduced. Recall that the solution of

$$-\Delta u = f \text{ on } \Omega, \qquad u = 0 \text{ on } \partial\Omega$$

can be written as

$$u(P) = \int_\Omega G(P,Q) f(Q) dQ, \qquad P \in \overline{\Omega}, \tag{1.12}$$

where G is the **Green's function**, which is the solution of

$$\begin{aligned} -\Delta G(P,Q) &= \delta(P-Q), \quad Q \in \Omega, \\ G(P,Q) &= 0, \qquad\qquad Q \in \partial\Omega. \end{aligned}$$

In the above, $P \in \Omega$ is fixed, Δ is with respect to Q and δ is the delta distribution defined by $\int_\Omega \delta(P-Q) g(Q)\, dQ = g(P)$ for any continuous g. If $P \in \partial\Omega$, define $G(P,Q) = 0$ for all $Q \in \overline{\Omega}$.

We now discuss a discrete analog of the Green's function. For a fixed $P \in \Omega_h$, define $G_h(P,\cdot)$ as the solution of

$$-\Delta_h G_h(P,\cdot) = h^{-2} \delta_P, \tag{1.13}$$

where δ_P is the vector with each entry equal to zero except for the one corresponding to P, which takes on the value one. When $P \in \partial\Omega_h$, define $G_h(P,Q) = 0$ for all

$Q \in \overline{\Omega}_h$. By the discrete maximum principle, $G_h(P,\cdot) \geq 0$. Another property is $G_h(P,Q) = G_h(Q,P)$ for $P, Q \in \overline{\Omega}_h$. It is not difficult to see that the solution of $-\Delta_h u_h = f_h$ is

$$u_h(P) = u_h \cdot \delta_P = u_h \cdot (-h^2 \Delta_h G_h(P,\cdot)) = h^2 f_h \cdot G_h(P,\cdot) = \langle G_h(P,\cdot), f_h \rangle_h. \tag{1.14}$$

[See (1.8) for the definition of the discrete inner product $\langle \cdot, \cdot \rangle_h$.] Clearly, (1.14) is a discrete form of (1.12).

Let $\mathbf{1}$ be the vector of all ones. The following fact is useful.

Proposition 1.5 For any $P \in \overline{\Omega}_h$,

$$0 \leq G_h(P,\cdot), \qquad \langle G_h(P,\cdot), \mathbf{1} \rangle_h \leq \frac{1}{8}. \tag{1.15}$$

Proof: For $P \in \overline{\Omega}_h$, the inequality $0 \leq G_h(P,\cdot)$ follows from the discrete maximum principle. Define

$$v_h(P) = \langle G_h(P,\cdot), \mathbf{1} \rangle_h = h^2 \sum_{Q \in \Omega_h} G_h(P,Q).$$

Observe that

$$\overline{\Delta}_h v_h = -\mathbf{1}. \tag{1.16}$$

By the discrete maximum principle, $v_h \geq 0$. Recall that w_h defined in (1.7) satisfies $\overline{\Delta}_h w_h = \mathbf{1}$. Thus $\overline{\Delta}_h (v_h + w_h) = 0$. By the discrete maximum principle, the maximum of $v_h + w_h$ occurs at $\partial\Omega_h$. Since v_h vanishes at $\partial\Omega_h$ and the maximum of w_h is $1/8$,

$$0 \leq v_h(P) \leq (v_h + w_h)(P) \leq \frac{1}{8}.$$

□

Stability now follows easily. For $P \in \Omega_h$, use (1.14) and the above proposition to get

$$|u_h(P)| \leq \langle G_h(P,\cdot), \mathbf{1} \rangle_h \, |f_h|_\infty \leq \frac{|f_h|_\infty}{8},$$

recovering our earlier result.

We have discussed four techniques to show stability with respect to the norms $|\cdot|_\infty$ and $|\cdot|_2$ for the discrete Laplacian: discrete Fourier analysis, maximum principle, energy method and Green's function. For a more general (possibly nonlinear) differential operator with variable coefficients, typically only a small subset of these tools is applicable.

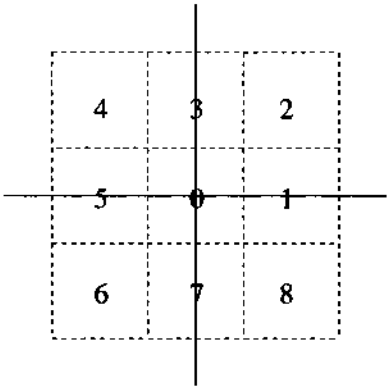

Figure 1.1 After a rotation by angle $\pi/4$ about the origin, cell i goes to $i+1$ for $1 \leq i \leq 7$ while cell 8 goes to cell 1. Of course, cell 0 retains its original position.

Rotationally Invariant Scheme

The Laplacian operator is rotationally invariant, meaning that Δ and any rotation operator commute. Fix any angle ϕ and define the rotation operator R_ϕ by $R_\phi v(r,\theta) = v(r,\theta+\phi)$. Here v is a function defined in polar coordinates. Hence the statement that Laplacian is rotationally invariant means that

$$R_\phi \Delta v = \Delta R_\phi v$$

holds for all ϕ and all twice differentiable functions v. In some applications, image processing to name one example, it is highly desirable to preserve this property as much as possible. For a uniform discretization in both directions, this means that a discrete Laplacian $\Delta_h^{(r)}$ should satisfy the criterion for a rotation angle ϕ equal to any integer multiple of $\pi/4$. In this section, we adopt the convention that u_{ij} is the approximation of the function u at the centre of the square cell with four corners $((i \pm 0.5)h, (j \pm 0.5)h)$. See Figure 1.1. Consider the infinite vectors u_h and v_h defined by

$$u_{ij} = \begin{cases} 1, & i > 0; \\ 0, & \text{otherwise}; \end{cases} \qquad v_{ij} = \begin{cases} 1, & i > j; \\ 0, & \text{otherwise}. \end{cases}$$

Here i, j range over the integers so that we need not be concerned about boundaries. Note that u_h is v_h rotated by an angle of $\pi/4$. It is simple to check that $(\Delta_h u_h)_{00} = h^{-2}$ while $(\Delta_h v_h)_{00} = 2h^{-2}$. Thus Δ_h does not respect grid rotational invariance.

Observe that $\bar{\Delta}_h$ defined by the molecule

$$\bar{\Delta}_h = \frac{1}{2h^2} \begin{bmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{bmatrix}$$

is also second-order consistent. Consider a new discrete Laplacian defined by

$$\Delta_h^{(r)} = \alpha\Delta_h + (1-\alpha)\bar{\Delta}_h, \qquad 0 \le \alpha \le 1.$$

Define α so that $(\Delta_h^{(r)}u_h)_{00} = (\Delta_h^{(r)}v_h)_{00}$ for the same vectors u_h, v_h defined above. A simple calculation shows that $\alpha = 1/3$. Consequently,

$$\Delta_h^{(r)} = \frac{1}{3h^2}\begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{1.17}$$

It is now apparent that $\Delta_h^{(r)}$ is grid-rotationally invariant.

We remark that a similar calculation leads to a second-order finite difference scheme for the first derivative which is grid-rotationally invariant. For instance,

$$\left.\frac{\partial u}{\partial x}\right|_{(ih,jh)} \approx \beta\frac{u_{i+1,j}-u_{i-1,j}}{2h} + \frac{1-\beta}{2}\left(\frac{u_{i+1,j+1}-u_{i-1,j+1}}{2h} + \frac{u_{i+1,j-1}-u_{i-1,j-1}}{2h}\right)$$

with $\beta = \sqrt{2}-1$.

The slight drawback with these schemes is the small additional complexity in both implementation and running time. One bonus feature enjoyed by these schemes is that they are less sensitive to noise. The reason is that the finite differences employ more grid points (nine versus five for the discrete Laplacian) and this has the effect of smoothing the data.

1.2 FOURTH-ORDER APPROXIMATION FOR Δ

In this section, we consider two fourth-order consistent schemes for the Poisson equation on the unit square. Provided that the solution is smooth, these schemes converge faster than second-order schemes and thus are more efficient. An obvious approach is to find a fourth-order consistent scheme for Δ. Unfortunately, this leads to problems at the boundary. To see this, simply consider the one-dimensional case. The stencil for a fourth-order consistent scheme for u'' spreads over five points. Applying the scheme centered at x_1 requires the knowledge of u_{-1} which is unknown. One simple fix is to use a second-order three-point scheme at x_1 and x_{n-1}. This new scheme appears to have an overall error of $O(h^2)$ due to the larger consistency errors at x_1 and at x_{n-1}. The surprise is that because there are so few points (namely, only two in the 1D case) with the larger consistency error, the global error is still $O(h^4)$. This can be shown using the discrete Green's function.

Define Ω_h^0 as the subset of Ω_h consisting of those points $P \in \Omega_h$ so that all four nearest neighbours of P lie in Ω_h. Thus points in $\Omega_h \setminus \Omega_h^0$ are adjacent to $\partial\Omega_h$.

Proposition 1.6 Let $P \in \Omega_h$ and G_h be the discrete Green's function defined in (1.13). Then

$$\sum_{Q \in \Omega_h \setminus \Omega_h^0} G_h(P,Q) \leq 1. \tag{1.18}$$

Proof: Define

$$v_h(P) = \begin{cases} 1, & P \in \Omega_h; \\ 0, & P \in \partial\Omega_h. \end{cases}$$

It can be checked that

$$(-\Delta_h v_h)(P) \begin{cases} = 0, & P \in \Omega_h^0; \\ \geq h^{-2}, & P \in \Omega_h \setminus \Omega_h^0. \end{cases}$$

Hence for $P \in \Omega_h$,

$$\begin{aligned} 1 &= v_h(P) \\ &= \langle -\Delta_h G_h(P,\cdot), v_h \rangle_h \\ &= \langle G_h(P,\cdot), -\Delta_h v_h \rangle_h \\ &= h^2 \sum_{Q \in \Omega_h \setminus \Omega_h^0} G_h(P,Q)(-\Delta_h v_h)(Q) \\ &\geq \sum_{Q \in \Omega_h \setminus \Omega_h^0} G_h(P,Q). \end{aligned}$$

□

Denote the fourth-order discrete Laplacian for points in Ω_h^0 by the molecule

$$\Delta_{h4} = \frac{1}{12h^2} \begin{bmatrix} & & -1 & & \\ & & 16 & & \\ -1 & 16 & -60 & 16 & -1 \\ & & 16 & & \\ & & -1 & & \end{bmatrix},$$

and denote by $\hat{\Delta}_h$ the discrete Laplacian which is fourth-order consistent in Ω_h^0 and second-order consistent in $\Omega_h \setminus \Omega_h^0$:

$$(\hat{\Delta}_h v_h)(Q) = \begin{cases} (\Delta_{h4} v_h)(Q), & Q \in \Omega_h^0; \\ (\Delta_h v_h)(Q), & Q \in \Omega_h \setminus \Omega_h^0. \end{cases} \tag{1.19}$$

For $P \in \Omega_h$, define another discrete Green's function $\hat{G}_h$ by

$$-\hat{\Delta}_h \hat{G}_h(P,\cdot) = \frac{\delta_P}{h^2}.$$

This discrete Green's function enjoys the same properties as G_h [see (1.15) and (1.18)]:

$$0 \le \hat{G}_h(P,\cdot), \qquad \langle \hat{G}_h(P,\cdot), \mathbf{1}\rangle_h \le \frac{1}{8} \tag{1.20}$$

and

$$\sum_{Q\in\Omega_h\setminus\Omega_h^0} \hat{G}_h(P,Q) \le 1. \tag{1.21}$$

The scheme $\hat{\Delta}_h u_h = R_h f$ is fourth-order convergent because in Ω_h^0, the consistency error is $O(h^4)$ while in $\Omega_h \setminus \Omega_h^0$, the error can be controlled by (1.21).

More precisely, define $e_h = u_h - R_h u$, where $\hat{\Delta}_h u_h = R_h f$. Using (1.19), (1.20) and (1.21), for any $P \in \Omega_h$,

$$\begin{aligned}
|e_h(P)| &= |\langle \hat{G}_h(P,\cdot), \hat{\Delta}_h e_h\rangle_h| \\
&\le h^2 \max_{Q\in\Omega_h^0} |(\Delta_{h4} e_h)(Q)| \sum_{Q\in\Omega_h^0} \hat{G}_h(P,Q) \\
&\quad + h^2 \max_{Q\in\Omega_h\setminus\Omega_h^0} |(\Delta_h e_h)(Q)| \sum_{Q\in\Omega_h\setminus\Omega_h^0} \hat{G}_h(P,Q) \\
&\le \frac{1}{8} \max_{Q\in\Omega_h^0} |(\Delta_{h4} e_h)(Q)| + h^2 \max_{Q\in\Omega_h\setminus\Omega_h^0} |(\Delta_h e_h)(Q)| \\
&\le \frac{1}{8} c_1 \|u\|^*_{C^6(\overline{\Omega})} h^4 + h^2 c_2 \|u\|^*_{C^4(\overline{\Omega})} h^2.
\end{aligned}$$

In the above, we used the fact that

$$\begin{aligned}
|\Delta_{h4} e_h(P)| &= |(\Delta_{h4} u_h)(P) - (\Delta_{h4} R_h u)(P)| \\
&= |(-R_h f - \Delta_{h4} R_h u)(P)| \\
&= |(R_h \Delta u - \Delta_{h4} R_h u)(P)| \\
&\le c_1 \|u\|^*_{C^6(\overline{\Omega})} h^4
\end{aligned}$$

for $P \in \Omega_h^0$ and similarly,

$$|\Delta_h e_h(P)| = |(\Delta_h u_h)(P) - (\Delta_h R_h u)(P)| \le c_2 \|u\|^*_{C^4(\overline{\Omega})} h^2$$

for $P \in \Omega_h \setminus \Omega_h^0$. These are the consistency errors on Ω_h. This completes the demonstration that, despite having an $O(h^2)$ consistency error near the boundary, the global error is still $O(h^4)$. Unfortunately, $\hat{\Delta}_h$ is non-symmetric.

Theorem 1.7 Fourth-order convergence of $\hat{\Delta}_h$. Suppose the solution u of (1.1) is in $C^6(\overline{\Omega})$ and $\hat{\Delta}_h u_h = R_h f$. Then $|R_h u - u_h|_\infty \le c\,(\|u\|^*_{C^4(\overline{\Omega})} + \|u\|^*_{C^6(\overline{\Omega})})\, h^4$.

A different approach to obtain a fourth-order scheme which is symmetric is to change both the scheme and the restriction operator. Define

$$\tilde{\Delta}_h = \frac{1}{6h^2}\begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix}.$$

Note that for this scheme, the discrete boundary $\partial\Omega_h$ includes the four corners of the square, namely, the points $(0,0),(1,0),(1,1),(0,1)$. Assume $v \in C^6(\overline{\Omega})$ and vanishes on $\partial\Omega$. By Taylor's expansion,

$$\tilde{\Delta}_h R_h v = R_h \Delta v + \frac{h^2}{12} R_h \Delta^2 v + \frac{h^4}{72}\left[\frac{1}{5}\left(\frac{\partial^6 v}{\partial x^6} + \frac{\partial^6 v}{\partial y^6}\right) + \frac{\partial^6 v}{\partial x^4 \partial y^2} + \frac{\partial^6 v}{\partial x^2 \partial y^4}\right], \tag{1.22}$$

where the derivatives in the h^4 term are evaluated at some point in Ω. Hence $\tilde{\Delta}_h$ is still an approximation of Δ with error $O(h^2)$. To obtain a fourth-order consistent scheme, we use a slightly different right-hand side. Define a new restriction operator by the molecule

$$\tilde{R}_h = \frac{1}{12}\begin{bmatrix} & 1 & \\ 1 & 8 & 1 \\ & 1 & \end{bmatrix} = \begin{bmatrix} & & \\ & 1 & \\ & & \end{bmatrix} + \frac{1}{12}\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}.$$

Note that $\tilde{R}_h v = R_h v + \frac{h^2}{12}\Delta_h R_h v$ and so by (1.6),

$$\begin{aligned} \tilde{R}_h \Delta v &= R_h \Delta v + \frac{h^2}{12}\overline{\Delta}_h \overline{R}_h \Delta v \\ &= R_h \Delta v + \frac{h^2}{12} R_h \Delta^2 v + E, \end{aligned}$$

where $|E|_\infty \le c\,\|v\|^*_{C^6(\overline{\Omega})} h^4$ since Δ_h is consistent of order 2. We need to use the more general definition (1.6) above since Δv may not vanish along $\partial\Omega_h$. Hence (1.22) now becomes

$$\tilde{\Delta}_h R_h v = \tilde{R}_h \Delta v - E + \frac{h^4}{72}\left[\frac{1}{5}\left(\frac{\partial^6 v}{\partial x^6} + \frac{\partial^6 v}{\partial y^6}\right) + \frac{\partial^6 v}{\partial x^4 \partial y^2} + \frac{\partial^6 v}{\partial x^2 \partial y^4}\right].$$

This means that the discretization $\tilde{\Delta}_h$ is a fourth-order consistent one. The scheme

$$-\tilde{\Delta}_h u_h = \tilde{f}_h := \tilde{R}_h f \tag{1.23}$$

can be shown (Exercise E5) to be fourth-order convergent:

Theorem 1.8 Fourth-order convergence of $\tilde{\Delta}_h$. Suppose the solution u of (1.1) is in $C^6(\overline{\Omega})$ and u_h satisfies (1.23). Then

$$|u_h - R_h u|_\infty \le c\,\|u\|^*_{C^6(\overline{\Omega})} h^4.$$

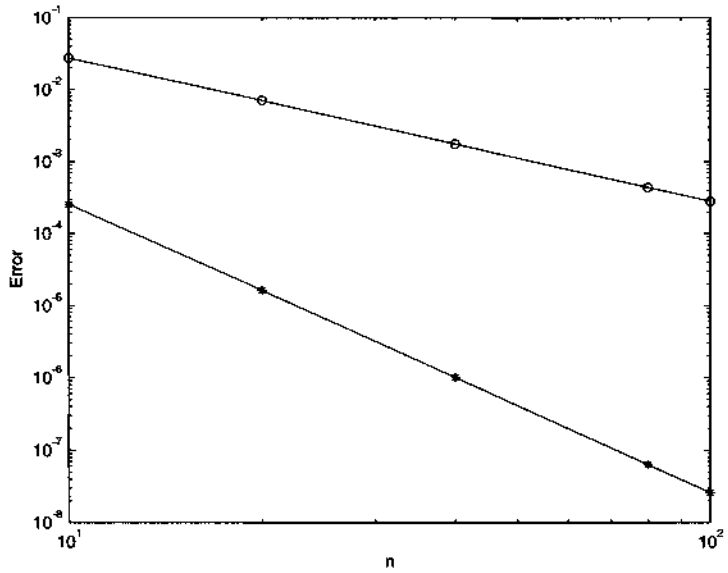

Figure 1.2 Errors in the infinity norm of schemes Δ_h ("$\circ$") and $\tilde{\Delta}_h$ ("$*$") as a function of $n = h^{-1}$. Here, $f(x, y) = 5\pi^2 \sin \pi x \sin 2\pi y$.

Richardson extrapolation is yet another approach to obtain a fourth-order scheme; see Exercise E7. Although slightly more complicated to implement, these fourth-order schemes are more efficient than the second-order scheme, requiring fewer grid points to obtain comparable accuracy. See Figure 1.2.

1.3 NEUMANN BOUNDARY CONDITION

In this section, replace the Dirichlet boundary condition $u = 0$ on $\partial\Omega$ by the Neumann boundary condition:

$$\nu \cdot \nabla u = \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega.$$

Here ν is the unit outward normal along $\partial\Omega$. Recall that the Neumann problem for $-\Delta u = f$ has a solution iff $\int_\Omega f = 0$ and all solutions differ by additive constants. As we shall see, the discrete cases have a similar property. Throughout this section $f_h = R_h f$. Recall that $\mathbf{1}$ is the vector whose entries are all ones.

There are three common approaches to tackle the Neumann boundary conditions. In the first two, the normal derivative is approximated by finite differences at $4(n-1)$ boundary points (excluding the corner points).

First-Order Difference

In the first approach, the Neumann boundary conditions are approximated by a first-order finite difference scheme. For instance, for $1 \leq j \leq n-1$,

$$0 = -\frac{\partial u}{\partial \nu}(x_{0j}) = \frac{\partial u}{\partial x}(x_{0j}) = \frac{u_{1j} - u_{0,j}}{h} + O(h). \tag{1.24}$$

The discrete Poisson equation at x_{1j} reads

$$\frac{4u_{1j} - u_{2j} - u_{0j} - u_{1,j+1} - u_{1,j-1}}{h^2} = f_{1j}.$$

Taking $u_{0j} = u_{1j}$, the above equation becomes

$$\frac{3u_{1j} - u_{2j} - u_{1,j+1} - u_{1,j-1}}{h^2} = f_{1j}$$

when $1 < j < n-1$ and

$$\frac{2u_{11} - u_{21} - u_{12}}{h^2} = f_{11}$$

when $j = 1$. This approximation is repeated for the other three sides. The discrete Laplacian operator becomes

$$-\Delta_h^{(1)} = \frac{1}{h^2}\begin{bmatrix} T-I & -I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -I & T-I \end{bmatrix},$$

where

$$T = \begin{bmatrix} 3 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 3 \end{bmatrix}.$$

Note that it is a symmetric, singular $(n-1)^2 \times (n-1)^2$ matrix.

Theorem 1.9 $-\Delta_h^{(1)} u_h = f_h$ is solvable iff $\mathbf{1} \cdot f_h = 0$. Any two solutions differ by a constant $c\mathbf{1}$, $c \in \mathbb{R}$.

Proof: By observation, $-\Delta_h^{(1)}\mathbf{1} = 0$. In fact, we show later that the dimension of the kernel of $-\Delta_h^{(1)}$ is one and hence any two solutions of the discrete Poisson equation $-\Delta_h^{(1)} u_h = f_h$ must differ by a constant times $\mathbf{1}$. By the Fredholm Alternative (see the appendix of Chapter 2), the discrete Poisson equation has a solution iff $\mathbf{1} \cdot f_h = 0$.

To show that $-\Delta_h^{(1)}$ has a one-dimensional null space, take $u_h(x_{11}) = 0$. Let L_h be $\Delta_h^{(1)}$ except that the row and column with the index corresponding to x_{11} are taken out so that L_h is a $[(n-1)^2 - 1] \times [(n-1)^2 - 1]$ matrix. It can be checked that L_h is irreducibly diagonally dominant (see the appendix at the end of this chapter) and so it is non-singular. Hence $\Delta_h^{(1)}$ has a one-dimensional null space. □

Consistency is defined separately for points in Ω_h and those in $\partial\Omega_h$. This is because the discretizations of the PDE and the boundary conditions are independent. For $v \in C^4(\overline{\Omega})$ and $x \in \Omega_h$,

$$(R_h \Delta v - \Delta_h^{(1)} R_h v)(x) = O(h^2).$$

For $x \in \partial\Omega_h$, the consistency error is $O(h)$ from (1.24). Hence the overall scheme is consistent of order 1. It is first-order convergent.

The fact that $\Delta_h^{(1)}$ is singular causes computational difficulties. Since the solution is defined only up to a constant, one simple solution is to prescribe a value, say, zero, to the discrete solution at some point. For the one-dimensional case, u_1 can be taken to be 0 and the resultant system can be written as

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}\begin{bmatrix} u_2 \\ u_3 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} f_2 \\ f_3 \\ \vdots \\ f_{n-2} \\ f_{n-1} \end{bmatrix}. \tag{1.25}$$

The new system is non-singular and thus standards methods can be used to solve it. By the discrete energy method, the new scheme is stable and first-order convergent. See Exercise E17.

Instead of setting one of the values of the unknown to be zero, another common approach is to demand that the solution have a zero average: $\mathbf{1} \cdot u_h = 0$. The extended system becomes

$$\begin{bmatrix} -\Delta_h^{(1)} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}\begin{bmatrix} u_h \\ \alpha \end{bmatrix} = \begin{bmatrix} f_h \\ 0 \end{bmatrix}.$$

Since the null space of $\Delta_h^{(1)}$ consists of constant multiples of $\mathbf{1}$, it is easy to show that the new matrix is non-singular. Assuming that the compatibility condition $\mathbf{1} \cdot f_h = 0$ is satisfied, α, the last component of the solution vector, must be zero. This scheme is symmetric and is also first-order convergent.

Second-Order Difference

In the second approach, the Neumann boundary conditions are approximated by a second-order finite difference. This requires the introduction of **fictitious points** which lie outside of $\overline{\Omega}_h$. For instance, for $1 \le j \le n-1$,

$$0 = \frac{\partial u}{\partial x}(x_{0j}) = \frac{u_{1j} - u_{-1,j}}{2h} + O(h^2).$$

Here, $x_{-1,j}$ is a fictitious point and we may take $u_{-1,j} = u_{1j}$. In contrast to previous cases, apply the discrete Poisson equation at the boundary points as well. For example, at x_{0j},

$$\frac{4u_{0j} - u_{-1,j} - u_{1j} - u_{0,j+1} - u_{0,j-1}}{h^2} = f_{0j}$$

for $0 < j < n$ becomes

$$\frac{4u_{0j} - 2u_{1j} - u_{0,j+1} - u_{0,j-1}}{h^2} = f_{0j}$$

and

$$\frac{4u_{00} - 2u_{10} - 2u_{01}}{h^2} = f_{00}.$$

This approximation is repeated for the other three sides. The discrete Laplacian operator becomes

$$-\Delta_h^{(2)} = \frac{1}{h^2}\begin{bmatrix} T & -2I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -2I & T \end{bmatrix},$$

where

$$T = \begin{bmatrix} 4 & -2 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -2 & 4 \end{bmatrix}.$$

Note that both T and I are $(n+1) \times (n+1)$ matrices and $\Delta_h^{(2)}$ is a singular, non-symmetric $(n+1)^2 \times (n+1)^2$ matrix. The unknowns include the values along the boundary grid points $\partial\Omega_h$ which includes the four corner points. This system is consistent of order 2:

$$(R_h \Delta v - \Delta_h^{(2)} \overline{R}_h v)(x) = O(h^2)$$

for $v \in C^4(\bar{\Omega})$ and $x \in \Omega_h$.

The system can be made symmetric by pre-multiplying by the diagonal matrix

$$D_h = \begin{bmatrix} D & & & & \\ & 2D & & & \\ & & \ddots & & \\ & & & 2D & \\ & & & & D \end{bmatrix}, \quad D = \begin{bmatrix} \frac{1}{4} & & & & \\ & \frac{1}{2} & & & \\ & & \ddots & & \\ & & & \frac{1}{2} & \\ & & & & \frac{1}{4} \end{bmatrix}.$$

The resultant system becomes

$$-D_h\Delta_h^{(2)}u_h = D_h f_h, \tag{1.26}$$

where u_h, f_h include points on $\partial\Omega_h$. It is a general principle of numerical analysis that *if the continuous problem has a special property such as symmetry, the discrete problem should preserve that property.*

Theorem 1.10 System (1.26) is solvable iff $1^T D_h f_h = 0$. Any two solutions differ by a constant multiple of 1.

In practice, a particular entry such as u_{00} can be set to zero, or the solution vector is required to have a zero average. The new system is stable and is second-order convergent. See Exercise E19.

Offset Grid

The final approach considers a grid which is offset or staggered by $h/2$ in both directions:

$$\tilde{\Omega}_h = \left\{ \left(\left(i + \frac{1}{2} \right) h, \left(j + \frac{1}{2} \right) h \right), \, 0 \le i, j < n \right\}.$$

Consequently, there are n^2 unknowns. The PDE is discretized at each of these points as before. A fictitious point and a second-order boundary condition are used to handle the points near the boundary. For instance, at $(\frac{1}{2}, j' = j + \frac{1}{2})$, $1 \le j < n-1$,

$$\begin{aligned} f_{\frac{1}{2},j'} &= \frac{4u_{\frac{1}{2},j'} - u_{-\frac{1}{2},j'} - u_{\frac{3}{2},j'} - u_{\frac{1}{2},j+\frac{3}{2}} - u_{\frac{1}{2},j-\frac{1}{2}}}{h^2} \\ &= \frac{3u_{\frac{1}{2},j'} - u_{\frac{3}{2},j'} - u_{\frac{1}{2},j+\frac{3}{2}} - u_{\frac{1}{2},j-\frac{1}{2}}}{h^2}, \end{aligned}$$

where the second-order boundary condition

$$\frac{u_{\frac{1}{2},j'} - u_{-\frac{1}{2},j'}}{h} = 0 + O(h^2)$$

has been used. Also,

$$f_{\frac{1}{2},\frac{1}{2}} = \frac{2u_{\frac{1}{2},\frac{1}{2}} - u_{\frac{3}{2},\frac{1}{2}} - u_{\frac{1}{2},\frac{3}{2}}}{h^2}.$$

The resultant discrete Laplacian operator is

$$-\Delta_h^{(3)} = \frac{1}{h^2} \begin{bmatrix} T-I & -I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -I & T-I \end{bmatrix},$$

where

$$T = \begin{bmatrix} 3 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 3 \end{bmatrix}.$$

This $n^2 \times n^2$ matrix is symmetric and singular. The theorem about solvability of this system is the same as before. Again, a stable and second-order convergent scheme is possible if some element such as $u_{\frac{1}{2},\frac{1}{2}}$ is assigned the value 0 or the solution vector is demanded to have a zero average.

1.4 POLAR COORDINATES

We now consider Ω as the unit (open) disk centered at the origin in $\mathbb{R}^2$. It is of course natural to work in polar coordinates (r, θ), in which the Laplacian operator can be written as

$$\Delta u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}.$$

Suppose the radial direction is divided into $n = 1/h$ grid points while the angular direction is divided into $m = 2\pi/k$ grid points. Let $u_{ij} = u(ih, jk)$, $1 \le i \le n-1$ and $1 \le j \le m$. The point at the origin must be given special treatment and this is discussed later. A second-order consistent discretization for Δ is

$$\frac{1}{r_i}\left[\frac{r_{i+\frac{1}{2}}\left(\frac{u_{i+1,j}-u_{ij}}{h}\right) - r_{i-\frac{1}{2}}\left(\frac{u_{ij}-u_{i-1,j}}{h}\right)}{h}\right] + \frac{1}{r_i^2}\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{k^2}, \tag{1.27}$$

where $r_i = ih$ and $r_{i\pm\frac{1}{2}} = \left(i \pm \frac{1}{2}\right)h$. The boundary conditions are $u_{nj} = 0$ and $u_{i0} = u_{im}$. The latter one is a periodic boundary condition for the θ variable.

The above discretization can be derived as follows. In the first step, approximate the outer r derivative by central differencing:

$$\frac{1}{r_i}\frac{(ru_r)_{i+\frac{1}{2}} - (ru_r)_{i-\frac{1}{2}}}{h}.$$

Secondly, approximate once again the remaining r derivative by central differencing. Note that we employed central differences of size $h/2$ in the above two steps. Had we taken differences of size h, the stencil would have a width of five mesh points (rather than three), which would have caused difficulties near the boundary. Note also that it would not be prudent to apply the central difference scheme to $r^{-1}u_r + u_{rr}$ (rather

than the above form $r^{-1}(ru_r)_r$) since the resultant scheme would be non-symmetric due to the term u_r.

The only remaining issue is the origin at which $r = 0$ and can give rise to problems if not dealt with properly. Note that the origin is a singularity of the coordinate system rather than that of the PDE. Let u_0 denote the value of u at the origin. Let $-\Delta u = f$, where f is smooth. For $\epsilon = h/2$,

$$\begin{aligned}
\int_0^{2\pi}\int_0^{\epsilon} fr\,dr\,d\theta &= -\int_0^{2\pi}\int_0^{\epsilon}\left[\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right)+\frac{1}{r}\frac{\partial^2 u}{\partial\theta^2}\right]dr\,d\theta \\
O(h^3)+2\pi f(0)\int_0^{\epsilon} r\,dr &= -\int_0^{2\pi} r\frac{\partial u}{\partial r}\bigg|_{r=0}^{r=\epsilon}d\theta-\int_0^{\epsilon}\frac{1}{r}\frac{\partial u}{\partial\theta}\bigg|_{\theta=0}^{\theta=2\pi}dr \\
O(h^3)+f(0)\pi\epsilon^2 &= -\frac{h}{2}\int_0^{2\pi}\frac{\partial u}{\partial r}\left(\frac{h}{2},\theta\right)d\theta \\
&= -\frac{hk}{2}\sum_{j=1}^{m}\frac{\partial u}{\partial r}\left(\frac{h}{2},jk\right)+O(hk^2) \\
&= -\frac{hk}{2}\sum_{j=1}^{m}\frac{u_{1j}-u_0}{h}+O(hk^2) \\
&= -\frac{k}{2}\left(\sum_{j=1}^{m}u_{1j}-mu_0\right)+O(hk^2).
\end{aligned}$$

In the fourth equality above, we approximate the integral using the trapezoidal rule with an error $O(k^2)$ for a twice continuously differentiable u. (If u happens to be infinitely differentiable, then it is known that the integration error is spectrally accurate, that is, it decreases faster than any polynomial function of m.) Rearranging, we obtain

$$u_0 = \frac{1}{m}\sum_{j=1}^{m}u_{1j}+\frac{f(0)h^2}{4}+O(h^3+hk^2). \tag{1.28}$$

Since our scheme in the interior is consistent of order two, only the first term of the right-hand side of the above formula is retained. In (1.27), replace u_{0j} by this approximation of u_0 for every j.

We have now a second-order accurate scheme but the resultant matrix is not symmetric. Since the original problem is self-adjoint, it is highly desirable for the discrete scheme to respect this property. A simple way to accomplish this is to solve the equivalent system $-r\Delta u = rf$. The new discretization becomes

$$-\frac{r_{i+\frac{1}{2}}(u_{i+1,j}-u_{ij})-r_{i-\frac{1}{2}}(u_{ij}-u_{i-1,j})}{h^2}-\frac{1}{r_i}\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{k^2}=r_i f_{ij}, \tag{1.29}$$

which is a symmetric system. The reason for the multiplication by r is that the inner product for functions and the corresponding discrete inner product are

$$\langle f, g \rangle = \int_\Omega f g \, r \, dr \, d\theta, \qquad \langle f, g \rangle_h = hk \sum_{i,j} f_{ij} g_{ij} r_i.$$

With respect to this discrete inner product, the negative discrete Laplacian is symmetric and positive definite.

While the above is a second-order accurate scheme, the singularity at the origin makes it a clumsy one. Note that if the boundary condition is of Neumann type rather than of Dirichlet type, then since the solution is defined only up to a constant, the value of u_0 can be set to zero instead of using (1.28). Using fictitious points to handle the Neumann boundary condition, this scheme becomes quite elegant.

Next, we derive another scheme which uses a simple trick to remove the coordinate singularity. The idea is to define the radial grid points so that the coefficient for u_0 becomes zero. Toward this end, define $s_i = (i - 0.5)\,h$, $i = 1, \ldots, n$ with $h = 2/(2n-1)$. Note that $s_n = 1$ and so the outermost grid point coincides with the boundary. Define $u_{ij} = u(s_i, jk)$, $1 \le i \le n-1$ and $1 \le j \le m$ for a total of $(n-1)m$ unknowns. (Since $u = 0$ on the boundary, $u_{nj} = 0$.) The angular direction is discretized as before with $m = 2\pi/k$ and the periodic boundary conditions $u_{i0} = u_{im}$ and $u_{i,m+1} = u_{i,1}$ must be imposed. The PDE $-r\Delta u = rf$ is now discretized as

$$-\frac{s_{i+\frac{1}{2}}(u_{i+1,j} - u_{ij}) - s_{i-\frac{1}{2}}(u_{ij} - u_{i-1,j})}{h^2} - \frac{1}{s_i}\frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{k^2} = s_i f_{ij}, \tag{1.30}$$

which is also a symmetric system. Here $s_{i-\frac{1}{2}} = (i-1)h$. The scheme when $i = 1$ is

$$-\frac{s_{\frac{3}{2}}u_{2j} - s_{\frac{3}{2}}u_{1j}}{h^2} - \frac{u_{1,j+1} - 2u_{1j} - u_{1,j-1}}{s_1 k^2} = s_1 f_{1j}.$$

The coefficient of u_0 is $s_{\frac{1}{2}}$, which is zero! This is a second-order scheme which gracefully avoids the point at the origin.

1.5 CURVED BOUNDARY

In this section, we give two different ways in which the finite difference method can be used to solve a PDE on an arbitrary domain. In the first method, a finite difference scheme is applied on the physical domain. The issue here is how to define the scheme for points near a curved boundary. In the second method, we perform a coordinate transformation so that in the new coordinate system, the domain becomes a disk or a rectangle. Usual finite difference schemes can then be applied to the PDE in the new coordinates.

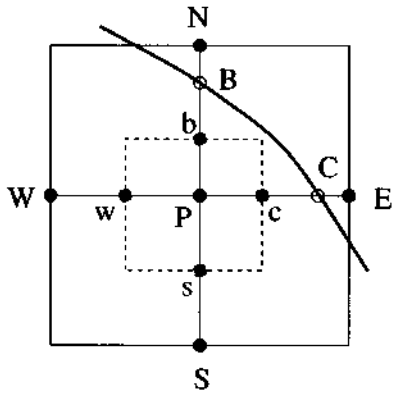

Figure 1.3 Example of a curved boundary (thick line) which does not pass through the grid points.

Finite Difference Scheme on Physical Domain

We define a second-order finite difference scheme for the self-adjoint PDE

$$-\nabla \cdot (a\nabla u) = f \text{ on } \Omega, \qquad u = g \text{ on } \partial\Omega, \tag{1.31}$$

where a is a positive differentiable function on $\overline{\Omega}$. Here Ω is no longer rectangular or circular. Consider the case as shown in Figure 1.3, where P is a grid point with its neighbors N, E, S, W. As before, a uniform grid of size h is imposed in both directions. Suppose the boundary intersects the grid lines at B and C. Let b be the midpoint between B and P. The points c, s, w are similarly defined. Suppose the lengths of the lines PB and PC are βh and γh, respectively. A difference scheme can be constructed as follows:

$$\frac{a(w)u_x(w) - a(c)u_x(c)}{\frac{1+\gamma}{2}\,h} + \frac{a(s)u_y(s) - a(b)u_y(b)}{\frac{1+\beta}{2}\,h} = f(P).$$

Using second-order central differences for the derivatives,

$$u_x(c) \approx \frac{u(C) - u(P)}{\gamma h} = \frac{g(C) - u(P)}{\gamma h}$$

and

$$u_y(b) \approx \frac{u(B) - u(P)}{\beta h} = \frac{g(B) - u(P)}{\beta h},$$

the difference scheme becomes

$$f(P) = \frac{\gamma a(w)(u(P)-u(W)) + a(c)u(P)}{\frac{1+\gamma}{2}\gamma h^2} + \frac{\beta a(s)(u(P)-u(S)) + a(b)u(P)}{\frac{1+\beta}{2}\beta h^2} - \frac{a(c)g(C)}{\frac{1+\gamma}{2}\gamma h^2} - \frac{a(b)g(B)}{\frac{1+\beta}{2}\beta h^2}. \tag{1.32}$$

Unfortunately, the scheme is no longer symmetric in general. Neumann boundary conditions can also be handled but it becomes awkward. The finite element method (to be presented in Chapter 3) gives a much more elegant solution.

The above example also illustrates that finite difference schemes for PDEs with non-homogeneous boundary conditions are really the same as those with homogeneous boundary conditions except for the addition of boundary terms on the right-hand side of the linear system. The matrices in both cases are the same.

The scheme is second-order consistent away from the boundary. Unfortunately, for nodes adjacent to the boundary, such as P in Figure 1.3, the scheme is only first-order consistent. Thus overall scheme is first-order consistent, stable and thus convergent. At first, it appears that the scheme is only first-order convergent. However, using the discrete Green's function (1.13), it can be shown that the scheme is actually second-order convergent. The argument is similar to the one used to show that the scheme $\hat{\Delta}_h$ (1.19) is fourth-order convergent despite a second-order consistency error near the boundary.

For simplicity, assume $a = 1$ and $g = 0$ in (1.31). A proof of stability of the scheme with respect to $|\cdot|_\infty$ is the subject of Exercise E25.

Theorem 1.11 Stability of Δ_h for general domain. Suppose Ω is bounded by a disk of diameter d. Let Δ_h be the discrete Laplacian with the above discretization near $\partial\Omega_h$. Then

$$|\Delta_h^{-1}|_\infty \le \frac{d^2}{16}. \tag{1.33}$$

Let $e_h = u_h - R_h u$, where R_h is the restriction to Ω_h. The following is a generalization of (1.15) and (1.18) for a domain Ω which is not necessarily a rectangle:

Proposition 1.12 Assume the hypotheses of the above theorem. For any $P \in \Omega_h$,

$$0 \le G_h(P,\cdot), \qquad \langle G_h(P,\cdot), 1\rangle_h \le \frac{d^2}{16} \tag{1.34}$$

and

$$\sum_{Q\in\Omega_h\setminus\Omega_h^0} G_h(P,Q) \le 1. \tag{1.35}$$

A proof of this proposition is the subject of Exercise E27.

We are now ready for a convergence analysis. For any $P \in \Omega_h$, use (1.34) and (1.35) to obtain

$$\begin{aligned} |e_h(P)| &= |(G_h(P,\cdot), \Delta_h e_h)_h| \\ &\le h^2 \max_{Q\in\Omega_h^0} |(\Delta_h e_h)(Q)| \sum_{Q\in\Omega_h^0} G_h(P,Q) \\ &\quad + h^2 \max_{Q\in\Omega_h\setminus\Omega_h^0} |(\Delta_h e_h)(Q)| \sum_{Q\in\Omega_h\setminus\Omega_h^0} G_h(P,Q) \\ &\le \frac{d^2}{16} \max_{Q\in\Omega_h^0} |(\Delta_h e_h)(Q)| + h^2 \max_{Q\in\Omega_h\setminus\Omega_h^0} |(\Delta_h e_h)(Q)| \\ &\le \frac{d^2}{16} c_1 \|u\|^*_{C^4(\overline{\Omega})} h^2 + h^2 c_2 \|u\|^*_{C^3(\overline{\Omega})} h. \end{aligned}$$

Thus, as claimed above, convergence rate is $O(h^2)$, despite a $O(h)$ consistency error near the boundary.

Theorem 1.13 Second-order convergence of Δ_h on general domain. Suppose $u \in C^4(\overline{\Omega})$ is the solution of (1.1) with Ω bounded by a disk of diameter d and $-\Delta_h u_h = R_h f$. Then $|e_h|_\infty \le c \|u\|^*_{C^4(\overline{\Omega})} h^2$.

An analogous result in the discrete L^2 norm is

$$|e_h|_h \le c \|u\|^*_{C^4(\overline{\Omega})} h^2. \tag{1.36}$$

We next introduce a superior scheme, known as the **ghost fluid method**, which yields a symmetric second-order method. It is sufficient to examine the one-dimensional case. Consider the ordinary differential equation (ODE) $-(au')' = f$ on $(0, x_e)$ with $0 < x_e < 1$ satisfying the boundary conditions $u(0) = g_0$, $u(x_e) = g_e$. Define a uniform grid with nodes $x_j = jh$, where $h = 1/n$ and $x_{n-1} < x_e < x_n$. The usual second-order discretization yields

$$\frac{a_{j-\frac{1}{2}}(u_j - u_{j-1}) - a_{j+\frac{1}{2}}(u_{j+1} - u_j)}{h^2} = f_j, \qquad 1 \le j < n-1,$$

where $a_{j-\frac{1}{2}} = a(x_{j-\frac{1}{2}})$. Suppose the solution can be extended smoothly to the fictitious or ghost point x_n which is exterior to the domain so that the above scheme holds for $j = n$. The value of u_n is defined by a simple linear extrapolation:

$$\frac{u_n - u_{n-1}}{h} = \frac{u_e - u_{n-1}}{\gamma h},$$

where $x_e - x_{n-1} = \gamma h$ for some γ satisfying $0 < \gamma < 1$. The scheme based at the point x_{n-1} now reads

$$\frac{a_{n-\frac{3}{2}}(u_{n-1} - u_{n-2}) + a_{n-\frac{1}{2}} u_{n-1}/\gamma}{h^2} = f_{n-1} + \frac{a_{n-\frac{1}{2}} g_e}{\gamma h^2}.$$

This scheme is symmetric and can be shown to be second-order convergent. The extension to the two-dimensional case is straightforward.

Finite Difference Scheme in New Coordinates

Suppose Ω is a simply connected domain, that is, a domain with no holes in it. Consider the same PDE (1.31). Suppose there is a smooth transformation taking $(x, y) \to (\xi, \eta)$ so that the domain becomes either a disk or rectangle in the new coordinates (ξ, η). For a simple domain such as the interior of an ellipse, the transformation in question is available analytically. In general, it must be approximated numerically by level set methods or grid generators. Let

$$x = x(\xi, \eta), \qquad y = y(\xi, \eta)$$

and $u(x, y) = u(x(\xi, \eta), y(\xi, \eta)) = U(\xi, \eta)$. It is easy to see that

$$u_x = \xi_x U_\xi + \eta_x U_\eta, \qquad u_y = \xi_y U_\xi + \eta_y U_\eta.$$

Since

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{bmatrix} \begin{bmatrix} d\xi \\ d\eta \end{bmatrix}, \qquad \begin{bmatrix} d\xi \\ d\eta \end{bmatrix} = \begin{bmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix},$$

we have

$$\begin{bmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{bmatrix} = \begin{bmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{bmatrix}^{-1} = \frac{1}{J} \begin{bmatrix} y_\eta & -x_\eta \\ -y_\xi & x_\xi \end{bmatrix},$$

where $J = x_\xi y_\eta - x_\eta y_\xi$ which is assumed to be non-zero for all (ξ, η). Hence

$$u_x = \frac{y_\eta U_\xi - y_\xi U_\eta}{J}, \qquad u_y = \frac{-x_\eta U_\xi + x_\xi U_\eta}{J}.$$

It should now be clear how to write the PDE in the new coordinates (ξ, η). Suppose $\partial\Omega$ is transformed to $\xi = \xi_1$. Then the boundary condition in the new coordinates becomes $U(\xi_1, \eta) = g(x(\xi_1, \eta), y(\xi_1, \eta))$. Usual finite difference schemes can be applied to the PDE in these new coordinates. In general, the schemes are not symmetric and have many more terms than the schemes in the original variables.

For special two-dimensional domains, techniques from complex analysis (conformal mapping) offer an elegant way to transform between different domains. See Exercise E37.

1.6 DIFFERENCE APPROXIMATION FOR Δ^2

The **biharmonic** operator $\Delta^2 = \partial^4_{xxxx} + \partial^4_{yyyy} + 2\partial^4_{xxyy}$ occurs in, for example, the bending of a plate with its sides clamped down and subject to a vertical force f:

$$\begin{aligned} \Delta^2 u &= f \text{ on } \Omega, \\ u = \frac{\partial u}{\partial \nu} &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{1.37}$$

The biharmonic operator is a self-adjoint, positive definite operator. However, it does not satisfy a maximum principle. That is, $\Delta^2 v \geq 0$ does not imply, in general, that the maximum of v is achieved on the boundary. To see this, consider the one-dimensional example $v(x) = -x^2$ on $(-1, 1)$. Note that $v'''' \equiv 0$ but v achieves its maximum at the origin and not at the boundary. A second-order consistent discretization of the biharmonic operator is given by

$$\frac{1}{h^4}\begin{bmatrix} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{bmatrix} \tag{1.38}$$

away from the boundary. As in the case of $\tilde{\Delta}_h$, the discrete boundary includes the four corners of the square. The boundary conditions are that u_h and its discrete normal derivative vanish on $\partial\Omega_h$. The normal derivative along $\partial\Omega_h$ is approximated by second-order finite difference with a fictitious point. As before, the value at each fictitious point is equated to the value of the corresponding interior point and the total number of unknowns is maintained at $(n-1)^2$. Let L_h denote the above discrete biharmonic operator. Note that it is symmetric. For the biharmonic operator, the discrete L^2 norm [see (1.9)] is more convenient for showing the stability and convergence of the scheme.

Theorem 1.14 **Second-order convergence of L_h.** For $u \in C^6(\overline{\Omega})$ in (1.37), the scheme $L_h u_h = f_h := R_h f$ is stable with respect to $|\cdot|_2$ and consistent of order 2. It is convergent of order 2 in the discrete L^2 norm.

Proof: Observe (after some work!) that L_h differs from Δ_h^2 only at $G := \Omega_h \setminus \Omega_h^0$ consisting of all grid points which are adjacent to the boundary $\partial\Omega_h$. The reason for the difference is that Δ_h assumes that the discrete function vanishes at the boundary, while for L_h, both the function and its normal derivative must be zero at the boundary. Let G_2 denote the set of four corner points of G and $G_1 = G \setminus G_2$. Now

$$\begin{aligned} u_h^T L_h u_h &= u_h^T \Delta_h^2 u_h + \frac{2}{h^4}\sum_{x\in G_1} u_h^2(x) + \frac{4}{h^4}\sum_{x\in G_2} u_h^2(x) \\ &= |\Delta_h u_h|_2^2 + \frac{2}{h^4}\sum_{x\in G_1} u_h^2(x) + \frac{4}{h^4}\sum_{x\in G_2} u_h^2(x) \\ &\geq c\,|u_h|_2^2 \end{aligned}$$

by the stability of Δ_h in the 2-norm (1.5). Thus L_h is symmetric positive definite with its smallest eigenvalue bounded below by a constant and so $|L_h^{-1}|_2 \leq c$. Therefore the scheme for the biharmonic operator is stable. The scheme at an interior point can be shown to be second-order consistent by a Taylor's series expansion. Since the boundary condition has also been discretized by a second-order scheme, the overall scheme is second-order consistent.

Let $e_h = u_h - R_h u$. Then

$$\begin{aligned} L_h e_h &= L_h u_h - L_h R_h u \\ &= R_h f - L_h R_h u \\ &= R_h \Delta^2 u - L_h R_h u \\ &=: \eta_h, \end{aligned}$$

where η_h is the consistency error. Using the inequality

$$|z|_2 \le \sqrt{N}\, |z|_\infty, \tag{1.39}$$

which holds for any N-vector z, we have

$$|\eta_h|_2 \le \frac{|\eta_h|_\infty}{h} \le \frac{ch^2 \|u\|^*_{C^6(\overline{\Omega})}}{h} = ch\|u\|^*_{C^6(\overline{\Omega})}.$$

Since the scheme is stable, $|e_h|_2 \le |L_h^{-1}|_2\, |\eta_h|_2 \le ch\|u\|_{C^6(\overline{\Omega})}$ and in the discrete L^2 norm,

$$|e_h|_h \le ch^2 \|u\|^*_{C^6(\overline{\Omega})}.$$

□

See Exercise E21 for a simple but sub-optimal error estimate in the infinity norm.

Note that $|L_h|_\infty = O(h^{-4})$. This is in contrast to $|\Delta_h|_\infty = O(h^{-2})$. Hence the condition number of L_h is $O(h^{-4})$ versus $O(h^{-2})$ for Δ_h. The consequence is that the biharmonic system is subject to far greater roundoff errors in the Gaussian elimination process or an iterative solution converges far slower. Higher-order PDEs are, in general, much more difficult to solve than second-order ones.

We conclude by mentioning that other boundary conditions are possible. One other example is $u = \Delta u = 0$ on $\partial\Omega$. This problem can be reduced to two Poisson equations:

$$\begin{aligned} -\Delta v &= f \text{ on } \Omega, \qquad v = 0 \text{ on } \partial\Omega; \\ -\Delta u &= v \text{ on } \Omega, \qquad u = 0 \text{ on } \partial\Omega. \end{aligned}$$

See Exercise E20. It is important to realize that the matrix of a discrete scheme takes into account the boundary conditions. For instance, a scheme for the above two Poisson equations reads $\Delta_h^2 u_h = f_h$. Observe that Δ_h^2 is different from L_h even though both discretize the biharmonic operator. The difference lies in the boundary conditions in the two cases.

1.7 A CONVECTION-DIFFUSION EQUATION

So far, we have encountered only self-adjoint operators. Let ϵ be a positive constant. The following convection-diffusion operator

$$Lu := -\epsilon u'' + u', \qquad u(0) = u(1) = 0, \tag{1.40}$$

is not self-adjoint because of the presence of the first derivative term. In the convection-dominated case, i.e., $0 < \epsilon \ll 1$, the problem becomes a singular perturbation one with a boundary layer of thickness ϵ near $x = 1$. A naive central difference scheme can lead to a certain instability. (Note that in the literature, some authors use the term advection in place of convection.)

First, by a direct computation, the eigenpairs of L are

$$\lambda_j = \frac{1}{4\epsilon} + j^2\pi^2\epsilon, \qquad \phi_j(x) = e^{\frac{x}{2\epsilon}} \sin j\pi x, \qquad j = 1, 2, 3 \ldots,$$

where $L\phi_j = \lambda_j\phi_j$.

Suppose we discretize L by the usual second-order finite difference scheme to obtain the discrete operator

$$L_h = \frac{\epsilon}{h^2}\begin{bmatrix} 2 & \mathrm{Pe}-1 & & & & \\ -\mathrm{Pe}-1 & 2 & \mathrm{Pe}-1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -\mathrm{Pe}-1 & 2 & \mathrm{Pe}-1 & \\ & & & -\mathrm{Pe}-1 & 2 & \end{bmatrix},$$

where

$$\mathrm{Pe} = \frac{h}{2\epsilon}$$

is the (mesh) **Péclet number**. Note that L_h is not a symmetric matrix.

The eigenvalues of L_h are

$$\frac{2\epsilon}{h^2} - \frac{2\epsilon}{h^2}\sqrt{1 - \mathrm{Pe}^2}\cos j\pi h, \qquad j = 1, \ldots, n-1.$$

Here we use the fact that the $(n-1) \times (n-1)$ tridiagonal matrix tridiag$[a,\ b,\ c]$ with constant subdiagonal a, diagonal b, and superdiagonal c has eigenvalues

$$b - 2\sqrt{ac}\cos\frac{j\pi}{n}, \qquad j = 1, \ldots, n-1.$$

Hence, if $\mathrm{Pe} < 1$, then all the eigenvalues of L_h are positive. On the other hand, if $\mathrm{Pe} > 1$, then L_h has complex eigenvalues, which is in contrast to the situation for L. In this case, spurious oscillations appear in the computed solution, which grow in amplitude as ϵ decreases (with h fixed). This is due to the instability of L_h whenever $\mathrm{Pe} > 1$. The oscillations are also apparent from the explicit expression of the analytical solution. See Exercise E32, where it is also revealed that a discrete maximum principle holds when $\mathrm{Pe} < 1$.

For a given ϵ, one can always satisfy the condition $\mathrm{Pe} < 1$ by making h sufficiently small. However, this means that the discrete problem is larger, which is often not feasible in 3D problems. Also it is a waste of resources when the solution is smooth in most of the domain and a very fine grid there is unnecessary.

Before we show how to overcome this stability problem, we digress to explain about the hyperbolic PDE $u_t + au_x = 0$ with initial condition $u(t = 0) = u_0(x)$. Here a is a given positive constant and the spatial domain is the entire real line. This PDE can be solved by the method of characteristics. A **characteristic** is defined as the solution of the ODE

$$\frac{dx(t)}{dt} = a, \qquad x(0) = \xi \in \mathbb{R},$$

which is $x(t) = at + \xi$. Along this characteristic, the total derivative

$$\frac{d}{dt}u(x(t),t) = u_x(x(t),t)\,x'(t) + u_t(x(t),t) = u_x(x(t),t)\,a + u_t(x(t),t) = 0$$

by the PDE. Hence the solution is constant along this characteristic:

$$u(x(t),t) = u(x(0),0) = u_0(\xi).$$

Since $\xi = x(t) - at$, the solution at any point (x,t) is $u_0(x - at)$. Observe that information travels to the right along a characteristic. This means that the solution at (x,t) for $t > 0$ only depends on the initial data at $x - at$.

One discretization of (1.40) which does not lead to spurious oscillations is an **upwind difference** scheme for the first derivative:

$$u'(x_i) \approx \frac{u_i - u_{i-1}}{h}.$$

Note that the PDE associated with the convection-diffusion operator (1.40) is

$$u_t - \epsilon u_{xx} + u_x = 0$$

when $\epsilon > 0$. The fact that information propagates in the direction of travel of the right-going characteristics of the associated hyperbolic operator $\partial_t + \partial_x$ (which dominates the diffusion operator $-\epsilon\partial_{xx}$ away from the boundary) explains the choice of finite difference of $u'(x_i)$–only take the point u_{i-1} which is the value to the left of x_i. The problem with the central approximation to $u'(x_i)$ is that it uses both u_{i-1} and u_{i+1} with the latter point to the right of x_i and thus violating the physical principle that information travels to the right. In case $\epsilon < 0$, then information propagates to the left and an appropriate upwind scheme is

$$u'(x_i) \approx \frac{u_{i+1} - u_i}{h}.$$

Using upwind differencing for the first derivative and the usual second-order difference for the second derivative, a discrete maximum principle holds and the scheme is stable (Exercise E33). Unfortunately, the order of convergence reduces to one. The consistency error of the scheme is, for some $\xi, \eta \in (0,1)$,

$$\epsilon\frac{u''''(\xi)h^2}{24} + \frac{u''(\eta)h}{2} = O(h),$$

which is sometimes called **numerical viscosity**. Roughly speaking, the viscosity term increases the boundary layer to a thickness which is comparable to the grid size and thus representable on the discrete domain.

There are several higher-order stable schemes which are free of oscillations for all Pe. See Exercise E35 for one example. Alternatively, some sort of adaptive (non-uniform) grid can be used, with more points inside the boundary layer. However, a non-uniform grid means that second-order accuracy of the discrete Laplacian operator is lost. Also, for nonlinear PDEs, the location of the boundary layer is unknown. In the multi-dimensional case, upwind differencing can be performed in the direction of convection; see Exercise E36. Convection-dominated differential equations are extremely difficult to solve in general and they are still under active investigation. For instance, the simple first-order upwind scheme fails for certain linear ODEs; see Exercise P11. For nonlinear equations, interior layers (regions where the solution changes rapidly occurring in the interior of the domain) may appear and their locations are usually unknown. They pose an even greater challenge.

1.8 APPENDIX: ANALYSIS OF DISCRETE OPERATORS

In this Appendix, we discuss some tools useful for analyzing discrete operators. First consider an $n \times n$ matrix A. We say that (distinct) indices i_0 and i_k are **connected** if there are indices $i_1, \ldots, i_{k-1}$ such that $a_{i_{j-1},i_j} \neq 0$, $j = 1, \ldots, k$. The index i is connected to itself if $a_{ii} \neq 0$. The matrix A is said to be **irreducible** if every index i is connected to every index j.

The matrix A is defined to be **diagonally dominant** if for every index i, $\sum_{j \neq i} |a_{ij}| < |a_{ii}|$. It is **irreducibly diagonally dominant** if

1. it is irreducible, and

2. for every index i, $\sum_{j \neq i} |a_{ij}| \leq |a_{ii}|$ and strict inequality holds for some index q.

Example 1.1 Let

$$A = \begin{bmatrix} 2 & -1 & \\ -1 & 2 & -1 \\ & -1 & 2 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 1 & \\ & & 1 \\ & & 2 \end{bmatrix}.$$

The graphs corresponding to these matrices are shown in Figure 1.4. By inspection, A is irreducible while B is not. While A is not diagonally dominant, it is irreducibly diagonally dominant. ◇

Before proving the main result, we need the following result.

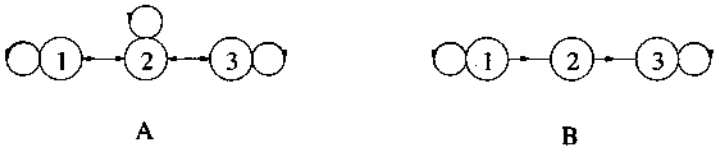

Figure 1.4 Graphs corresponding to matrices A and B.

Theorem 1.15 Gerschgorin's Theorem. All the eigenvalues of an irreducible matrix A lie in the set

$$\bigcup_i D_{r_i}(a_{ii}) \bigcup \left(\bigcap_j \partial D_{r_j}(a_{jj}) \right),$$

where $D_r(x) = \{z \in \mathbb{C}, |z - x| < r\}$ and $r_i = \sum_{j \neq i} |a_{ij}|$.

Proof: Let $Ax = \lambda x$, $|x|_\infty = |x_k| = 1$. If $\lambda \in \bigcup_i D_{r_i}(a_{ii})$, then there is nothing to prove. Suppose not. Now

$$\sum_j a_{kj} x_j = \lambda x_k$$

implies that $(\lambda - a_{kk})x_k = \sum_{j \neq k} a_{kj} x_j$ or

$$|\lambda - a_{kk}| \leq \left| \sum_{j \neq k} a_{kj} x_j \right| \leq \sum_{j \neq k} |a_{kj}| = r_k.$$

Since $\lambda \notin D_{r_k}(a_{kk})$, $|\lambda - a_{kk}| = r_k$ and hence $\lambda \in \partial D_{r_k}(a_{kk})$. We now show that $\lambda \in \partial D_{r_i}(a_{ii})$ for every i.

Since A is irreducible, there is a connection $k = i_0, i_1, \ldots, i_p = i$ such that $a_{i_{j-1}, i_j} \neq 0$. Now

$$r_k = |\lambda - a_{kk}| \leq \sum_{j \neq k} |a_{kj}||x_j| \leq r_k$$

implies that $|x_j| = 1$ for every $j \neq k$ with $a_{kj} \neq 0$. In particular, $a_{k,i_1} \neq 0$, which means that $|x_{i_1}| = 1$. Hence

$$|\lambda - a_{i_1,i_1}| \leq \sum_{j \neq i_1} |a_{i_1,j}||x_j| \leq r_{i_1} \quad \Rightarrow \quad |\lambda - a_{i_1,i_1}| = r_{i_1}.$$

So $\lambda \in \partial D_{r_{i_1}}(a_{i_1,i_1})$. Now repeat the argument $p - 1$ more times to obtain $\lambda \in \partial D_{r_i}(a_{ii})$. This establishes Gershgorin's Theorem. □

Theorem 1.16 Diagonally dominant or irreducibly diagonally dominant matrices are invertible.

Proof: Let A be diagonally dominant, $Ax = \lambda x$ with $|x|_\infty = |x_k| = 1$. Similar to the above proof, we obtain

$$|\lambda - a_{kk}| \leq \sum_{j \neq k} |a_{kj}| < |a_{kk}|$$

since A is diagonally dominant. Thus A can have no zero eigenvalue.

Now suppose A is irreducibly diagonally dominant. If 0 is an eigenvalue, then according to Gershgorin's Theorem, there are two possibilities. First, $0 \in D_{r_i}(a_{ii})$ for some i. Then $|0 - a_{ii}| < r_i \leq |a_{ii}|$ which is a contradiction. The second possibility is that $0 \in \partial D_{r_j}(a_{jj})$ for every j. In particular, take $j = q$, where $r_q < |a_{qq}|$. Now $|0 - a_{qq}| = r_q < |a_{qq}|$ which is, again, absurd. □

1.9 SUMMARY AND EXERCISES

The finite difference scheme replaces linear differential operators by finite differences, with the PDE becoming a linear system of equations. The finite difference matrix is **sparse**, meaning that it has $O(h^{-1})$ non-zero entries, and has a condition number which behaves like $O(h^{-2})$ for second-order PDEs. The scheme is simple in concept and implementation for regular geometry. However, its error analysis is not sharp in the sense that it requires the solution to be smoother than necessary (C^4 for a second-order scheme for the Laplacian and C^6 for a fourth-order scheme of the Laplacian or a second-order scheme for the biharmonic operator). Also, it becomes clumsy when the geometry is not regular, especially for Neumann problems. Discrete versions of the maximum principle, integration by parts, Green's function and energy method are some of the techniques which are inherited from the continuous problem and are useful for the stability analysis of the discrete problem. For a PDE more difficult than the Poisson equation, perhaps only one of the above tools is applicable and so all these techniques are indispensable in a numerical analyst's arsenal.

Exercise:

E1 Consider the scheme $A_h u_h = f_h := R_h f$ to solve $-\Delta u = f$ on $\Omega = (0,1)^2$ with $u = 0$ on $\partial\Omega$. Assume that A_h is non-singular and that the scheme is second-order convergent: $|R_h u - u_h|_\infty \leq ch^2 \|u\|^*_{C^4(\overline{\Omega})}$. Show that the scheme is stable: $|A_h^{-1}|_\infty \leq c$.

E2 Show (1.39). Use this result to show (1.10). Also show the discrete energy error estimate $|e_h|_{1,h} \leq ch^2 \|u\|^*_{C^4(\overline{\Omega})}$, where

$$|v_h|^2_{1,h} = h^2 \left(|v_h|_2^2 + |\delta_h^x v_h|_2^2 + |\delta_h^y v_h|_2^2 \right)$$

for any vector v_h which vanishes on $\partial\Omega_h$.

E3 Show (1.20) and (1.21).

E4 Show (1.22).

E5 For $\tilde{\Delta}_h$, the fourth-order difference approximation of the Laplacian, estimate $|\tilde{\Delta}_h^{-1}|_2$ using a discrete Fourier analysis and $|\tilde{\Delta}_h^{-1}|_\infty$ using a discrete maximum principle. Show that the scheme $-\tilde{\Delta}_h u_h = \tilde{f}_h$ is convergent of order 4 in $|\cdot|_\infty$.

E6 Use the discrete energy method to estimate $|\tilde{\Delta}_h^{-1}|_2$. Show that the scheme $-\tilde{\Delta}_h u_h = \tilde{f}_h$ is convergent of order 4 in the $|\cdot|_h$ norm.

E7 Consider the Poisson equation with a homogeneous boundary condition and the usual finite difference scheme $-\Delta_h u_h = R_h f$. Show that

$$\max_{\bar{\Omega}_h} \left| \frac{4u_{h/2}}{3} - \frac{u_h}{3} - u \right| \leq ch^4$$

for some constant c independent of h. Thus, a fourth-order solution can be obtained by taking a linear combination of two solutions at two different values of h. This is known as **Richardson extrapolation**. Hint: Define $-\Delta v = F := (u_{xxxx} + u_{yyyy})/12$ with $v = 0$ along the boundary and $e_h = h^{-4}(u_h - R_h u - h^2 R_h v)$. Assume that $u \in C^6(\overline{\Omega})$ and $v \in C^4(\overline{\Omega})$. Show $|e_h|_\infty \leq c(\|u\|^*_{C^6(\overline{\Omega})} + \|v\|^*_{C^4(\overline{\Omega})})$.

E8 Discretize the operator $Lu = (pu')'$ for $u \in C^4([0,1])$ vanishing at the boundary using the usual second-order finite differences. (The difference equation at node x_i should only involve the values of u_{i-1}, u_i, u_{i+1}.) Find an expression for the consistency of the scheme. Repeat for $Lu = p'u' + pu''$. Which scheme is better?

E9 Suppose $u = u(x,y)$ is a smooth function. Find a second-order difference approximation for $\partial^2_{xy} u$. Assume an equally spaced grid of size h in both directions. Hint: It is sufficient to take the four points $u_{i\pm1,j+1}$ and $u_{i\pm1,j-1}$.

E10 Suppose $u = u(x,y)$ is a smooth function. Find a second-order difference approximation for u_y at the point $((i+0.5)h, jh)$ in terms of u along the grid points. Here, h is the grid size.

E11 Suppose $u = u(x,y)$ is a smooth function and we take a finite difference approximation $\sum_{-1\leq i,j\leq 1} c_{ij} u_{ij}$ of $\Delta u(0,0)$. Find conditions on the coefficients c_{ij} so that the scheme is second-order consistent. Show that the scheme cannot be third-order consistent.

E12 Let Ω be a bounded domain in $\mathbb{R}^2$ with a smooth boundary. Suppose $\Delta u = 0$ on Ω. Show that $|\nabla u|_2^2$ achieves its maximum on $\partial\Omega$. Prove a discrete version of this fact.

E13 Formulate and prove a discrete maximum principle for the differential operator $\Delta + \partial_x$ discretized by second-order differences. A maximum principle for $\Delta + c$, where c is a negative constant, is the following. If $\Delta v + cv \geq 0$ on Ω, then $\max_{\overline{\Omega}} v \leq \max_{\partial\Omega} v^+$, where $v^+(x) = \max(0, v(x))$ for $v \in C^2(\Omega) \cap C(\overline{\Omega})$. Give an example in one dimension

showing that the conclusion cannot be changed to $\max_{\overline{\Omega}} v = \max_{\partial\Omega} v^+$ or to $\max_{\overline{\Omega}} v \le \max_{\partial\Omega} v$. Formulate and prove a corresponding discrete maximum principle.

E14 The **comparison principle** states that if $u, v \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfy $-\Delta u \le -\Delta v$ on Ω and $u \le v$ on $\partial\Omega$, then $u \le v$ on Ω. Prove this. State and prove a discrete comparison principle.

E15 Let Ω be the unit square and a be a positive constant. Show that the second-order finite difference scheme with a uniform grid for the operator $-au_{xx} - u_{yy}$ satisfies the discrete maximum principle. Assume that u vanishes on the boundary. Show that the scheme is symmetric positive definite.

E16 Prove the stability and first-order convergence of the Neumann scheme $-\Delta^{(1)} u_h = f_h$ for u_h with a zero average.

E17 Prove the stability and convergence of the scheme (1.25) using the discrete energy method. Use a generalized (one-dimensional) version of the discrete Poincaré–Friedrichs inequality $|v_h|_2 \le |\delta_h^x v_h|_2$ which holds for any v_h which vanishes at one end point (but not necessarily at the other end point).

E18 Prove Theorem 1.10.

E19 Repeat Exercise E17 for the one-dimensional version of the scheme (1.26). Set $u_0 = 0$.

E20 Find a second-order convergent scheme for the biharmonic operator with the boundary conditions $u = \Delta u = 0$ on $\partial\Omega$. Prove that it is second-order convergent.

E21 Show the error estimate $|e_h|_\infty \le ch \|u\|^*_{C^6(\overline{\Omega})}$ for the scheme $L_h u_h = f_h$, where L_h is the discrete approximation of the biharmonic operator. Hint: First show that any $N \times N$ matrix A satisfies the inequality $|A|_\infty \le \sqrt{N}\,|A|_2$.

E22 Study the stability (boundedness of the discrete solution independent of h) of the usual second-order finite difference scheme for the PDE $-\Delta u = \cos u$ on $\Omega = (0,1)^2$ with $u = 0$ on $\partial\Omega$. Suppose we solve the discrete nonlinear equation by the simple iteration $-\Delta_h u_h^{(k+1)} = \cos u_h^{(k)}$. Show that this iteration converges to some u_h for all initial iterate $u_h^{(0)}$. Study the convergence of u_h to $R_h u$ as $h \to 0$.

E23 Repeat the above problem for the PDE $-\Delta u = e^{-u}$.

E24 For an arbitrary domain bounded by some disk of diameter d, show that the scheme (1.32) is first-order consistent. (Take an arbitrary smooth diffusion coefficient a.)

E25 Show (1.33) using the discrete maximum principle. Hint: Define the discrete function $w_{ij} = ((ih - p)^2 + (jh - q)^2)/4$ for some $p, q \in \mathbb{R}$ and show that $\overline{\Delta}_h w_h = \mathbf{1}$. Here $\overline{\Delta}_h$, in the same spirit as (1.6), operates on

vectors defined on $\overline{\Omega}_h$ which now includes points on the curved boundary such as points B and C in Figure 1.3.

E26 Consider the discrete Poisson equation $-\Delta_h u_h = f_h$ on $\mathbb{Z}^2$ $(h = 1)$. With $u_{ij} = u(i, j)$, show that

$$u_{pq} = \sum_{i,j\in\mathbb{Z}} G_{p-i,q-j} f_{ij}, \; G_{ij} = \frac{1}{16\pi^2} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{\cos(ix + jy) - 1}{\sin^2 \frac{x}{2} + \sin^2 \frac{y}{2}}\, dx\, dy.$$

E27 Show (1.34) and (1.35) using the discrete Green's function.

E28 For a non-rectangular Ω, estimate the smallest and largest eigenvalues of Δ_h. Hint: To estimate the largest eigenvalue, consider $\tilde{\Omega}$, a square which contains Ω. Recall that $\lambda_{\max}(\tilde{\Omega}) \leq \lambda_{\max}(\Omega)$.

E29 Discuss the consistency, stability and convergence of the ghost fluid method for the 1D boundary value problem in the text.

E30 Recall the discrete Green's function defined in (1.13). Show that $G_h(P, Q) = G_h(Q, P)$, where $P, Q \in \overline{\Omega}_h$. Show (1.16).

E31 Show that the discrete Green's function satisfies $0 \leq G_h(P, Q) \leq -C \ln h$. Hint: First show that

$$G_h(P, Q) = \sum_{i,j=1}^{n-1} \frac{q^{(ij)}(P) q^{(ij)}(Q)}{\lambda_{ij} h^2} \leq c \sum_{i,j=1}^{n-1} \frac{1}{i^2 + j^2};$$

see (1.4). Estimate the right-hand side by an integral over a quarter disk. Use the above result to show that for any vector v_h vanishing on $\partial\Omega_h$,

$$|v_h|_\infty \leq c\,(\ln h^{-1})^{1/2} \left(|\delta_h^x v_h|_h^2 + |\delta_h^y v_h|_h^2\right)^{1/2}.$$

E32 Let ϵ be a positive constant. Apply the usual second-order central difference scheme to $-\epsilon u'' + u' = 0$ with boundary conditions $u(0) = 0$, $u(1) = 1$. Solve the resultant difference equation. Note the occurrence of oscillations if $\mathrm{Pe} > 1$. Now change the boundary condition at the right to $u(1) = 0$. Show that if $\mathrm{Pe} < 1$, then a discrete maximum principle holds and show stability.

E33 Prove a discrete maximum principle for (1.40) with ϵ positive. Prove stability of the upwind scheme for the first derivative discussed in the text using the discrete maximum principle.

E34 Consider the following scheme for the equation $-\epsilon u'' + u' = 0$ with boundary conditions $u(0) = 0$, $u(1) = 1$ and $\epsilon > 0$:

$$\epsilon \frac{2u_j - u_{j-1} - u_{j+1}}{h^2} + \frac{(1-\theta)(u_{j+1} - u_j) + \theta(u_j - u_{j-1})}{h} = 0.$$

Find the value of the parameter θ, where $0 \leq \theta \leq 1$, so that the scheme is exact at the nodes (i.e., $u_j = u(jh)$, $j = 0, \ldots, n$).

E35 Consider the following scheme for $Lu = f$, where L is defined in (1.40) and $f \in C^2([0,1])$:

$$\frac{2+\mathrm{Pe}}{\mathrm{Pe}\,(1+\mathrm{Pe})} f_j + \frac{\mathrm{Pe}}{2(1+\mathrm{Pe})} f_{j-1}$$
$$= \epsilon \frac{2u_j - u_{j-1} - u_{j+1}}{h^2} + \frac{\mathrm{Pe}}{1+\mathrm{Pe}} \frac{u_j - u_{j-1}}{h} + \frac{1}{1+\mathrm{Pe}} \frac{u_{j+1} - u_{j-1}}{2h}.$$

Show that the consistency error is $O(h^2)$ for all $\epsilon > 0$. Investigate the stability and convergence of this scheme. It should be emphasized that all constants in the analysis can be made independent of both h and ϵ.

Note that when $\mathrm{Pe} \gg 1$, the discretization of the first derivative is essentially an upwind scheme. An increase in the order of the consistency error comes from a linear combination of f_j and f_{j-1}; the same idea was used in the section on a fourth-order approximation of Δ.

E36 Consider the PDE

$$-\epsilon \Delta u + b \cdot \nabla u = f$$

on the unit square with a homogeneous Dirichlet boundary condition. Assume b is a constant vector such that $b_1 < 0$, $b_2 > 0$, $|b|_\infty = 1$ and $P_j := |b_j|\, h/(2\epsilon) \geq 1$ for $j = 1, 2$. Consider the discretization

$$-\epsilon(\Delta_h u)_{ij} + (b \cdot \nabla u)_{ij} = \frac{P_1}{2(1+P_1)} f_{i+1,j} + \frac{P_2}{2(1+P_2)} f_{i,j-1},$$

where

$$(b \cdot \nabla u)_{ij} = \frac{b_1}{1+P_1} \Big[P_1 \Big(\omega_1\, \delta_h^x v_{ij} + (1-\omega_1)\, \delta_h^x u_{i,j-1} \Big) + \frac{u_{i+1,j} - u_{i-1,j}}{2h} \Big]$$
$$+ \frac{b_2}{1+P_2} \Big[P_2 \Big(\omega_2\, \delta_h^y u_{i,j-1} + (1-\omega_2)\, \delta_h^y u_{i+1,j-1} \Big) + \frac{u_{i,j+1} - u_{i,j-1}}{2h} \Big]$$

and

$$\omega_1 = 1 - \frac{P_2(1+P_1)}{2P_1(1+P_2)}, \qquad \omega_2 = 1 - \frac{P_1(1+P_2)}{2P_2(1+P_1)}.$$

Show that the scheme is consistent of order two for all $\epsilon > 0$. Show stability and second-order convergence of this method.

Convection-dominated PDEs in higher dimensions are more complicated to solve than their one-dimensional counterpart due to geometric reasons.

For instance, it is desirable to increase the diffusion in the direction of b but not in the direction perpendicular to b.

E37 Consider the PDE $\Delta\psi = 0$ in the domain which is the region in between the circles $|z| = 1$ and $|z - 1| = 5/2$ in the complex plane. The boundary conditions are $\psi = \psi_1$ along the inner circle and $\psi = \psi_2$ along the outer circle. Show that the conformal map $w = f(z) = (z + 1/4)/(z + 4)$ maps the above two circles into two concentric circles in the w-plane. Let $\psi(z) = \phi(f(z))$. Show that $0 = \Delta\psi = |f'(z)|^2(\phi_{\xi\xi} + \phi_{\eta\eta})$, where $w = \xi + i\eta$. Summarize your solution procedure for the original problem.

E38 Consider the domain Ω which is nearly a disk with boundary given by $r = 1 + \epsilon\cos 5\theta$ in polar coordinates with ϵ a positive number. Consider the transformation $(r, \theta) \to (R, \theta)$ with $R = r\,(1 + \epsilon\cos 5\theta)^{-1}$ so that in the new coordinates, the domain is the unit disk. Work out the Poisson equation in the new coordinates.

P1 Solve the Poisson equation on the unit square with a homogeneous Dirichlet boundary condition using the three fourth-order finite difference schemes described in the text. Take f so that the exact solution is $u(x, y) = x^\alpha(1 - x)\sin \pi y$. Monitor the decrease of the error as a function of $n = 1/h$ for $\alpha = 1/2$, $3/2$ and 2. Explain the differences.

P2 Solve the Poisson equation on the unit disk with a homogeneous Dirichlet boundary condition using a symmetric second-order finite difference scheme in polar coordinates. Take f so that the exact solution is $u(r, \theta) = \cos(\pi r/2)\sin\theta$.

P3 Repeat the above exercise using the ghost fluid method.

P4 Solve the Poisson equation inside the ellipse $x^2 + 4y^2 = 4$ with a homogeneous boundary condition. Take as the exact solution $u(x, y) = x(x^2 + 4y^2 - 4)^2$.

P5 Solve the Poisson equation on Ω defined in Exercise E38 with a homogeneous Dirichlet boundary condition and in the new coordinates so that the computational domain becomes a disk.

P6 Solve the biharmonic equation (1.37) with homogeneous Dirichlet boundary conditions using a second-order finite difference scheme with exact solution $u(x, y) = \sin^2 \pi x \sin^2 2\pi y$.

P7 Solve the Poisson equation on Ω which is a right-angled triangle with one side along the x axis and another one along the y axis. Both these sides have length one. Impose the homogeneous Dirichlet boundary condition along these two sides and the homogeneous Neumann boundary condition along the third side. Derive a second-order numerical boundary condition on the third side using fictitious points and make sure that the final scheme is symmetric.

P8 In polar coordinates, define the domain Ω as all (r, θ) so that $0 \le r < 2 + \sin\theta$, $0 \le \theta < 2\pi$. Plot Ω. Solve the Poisson equation on Ω with a homogeneous Dirichlet boundary condition. Take the exact solution $u(r, \theta) = r^4(2 + \sin\theta)^{-4} - 1$. Use the scheme (1.30) in conjunction with the ghost fluid method.

P9 Consider the Poisson equation $-\Delta u = f$ on Ω which is the sphere in $\mathbb{R}^3$ of radius one. Note that there is a solution iff the compatibility condition $\int_\Omega f = 0$ holds. The solution (if it exists) is unique up to an additive constant. In spherical coordinates

$$x = \sin\phi\,\cos\theta, \qquad y = \sin\phi\,\sin\theta, \qquad z = \cos\phi$$

$(0 \le \theta < 2\pi,\ 0 \le \phi \le \pi)$, the PDE reads

$$-\frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial u}{\partial\phi}\right) - \frac{1}{\sin^2\phi}\frac{\partial^2 u}{\partial\theta^2} = f.$$

Observe that there are coordinate singularities at the north and south poles. Since the inner product for this problem has the form

$$\langle f, g\rangle = \int_0^{2\pi}\int_0^{\pi} f(\phi,\theta)\,g(\phi,\theta)\,\sin\phi\,d\phi\,d\theta,$$

the discrete scheme will be symmetric if we discretize the PDE in the form

$$-\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial u}{\partial\phi}\right) - \frac{1}{\sin\phi}\frac{\partial^2 u}{\partial\theta^2} = f\,\sin\phi.$$

Let $u_{jk} = u(\phi_j, \theta_k)$, where $\theta_k = k\tau$, $\tau = 2\pi/m$ for some positive m with $1 \le k \le m$ and, because of the coordinate singularities at $\phi = 0, \pi$, we offset the grid for ϕ by $h/2$, where $h = \pi/n$ is the ϕ grid size for some positive n:

$$\phi_j = \left(j - \frac{1}{2}\right)h, \qquad 1 \le j \le n.$$

Of course, apply the periodic conditions $u_{j0} = u_{jm}$ and $u_{j,m+1} = u_{j1}$. Derive a symmetric finite difference scheme. How would you handle the fact that the matrix is singular? Verify numerically that the error behaves like $O(h^2)$ choosing $m = 2n$.

P10 Solve the PDE

$$-\epsilon\Delta u + b\cdot\nabla u = f$$

on the unit square with a homogeneous Dirichlet boundary condition. Assume b is a constant vector such that $b_1 < 0$, $b_2 > 0$, $|b|_\infty = 1$ and

$P_j := |b_j| h/(2\epsilon) \geq 1$ for $j = 1, 2$. Use (i) central differencing for all terms, (ii) central difference scheme for $-\Delta$ and the upwind scheme

$$(b \cdot \nabla u)_{ij} = b_1 \frac{u_{i+1,j} - u_{ij}}{h} + b_2 \frac{u_{ij} - u_{i,j-1}}{h}$$

and (iii) the scheme defined in Exercise E36. Compare (numerically) the three schemes for $h = 0.1$ and $\epsilon = 0.1,\ 0.01$ for the exact solution $u(x, y) = (1 - x - g(1 - x))(y^2 - g(y))$, where $g(x) = (1 - e^{x/\epsilon})(1 - e^{1/\epsilon})^{-1}$. Explain the difference.

P11 Consider the ODE $-\epsilon u'' + (x - 0.5)u' = x - 0.5$ on $(0, 1)$ with boundary conditions $u(0) = -0.5$, $u(1) = 0.5$. The exact solution is $u(x) = x - 0.5$. This ODE is said to have a turning point because the coefficient of u' vanishes at $x = 0.5$. Solve this ODE using a simple first-order upwind scheme for $\epsilon = 1, 0.1, 0.01, 0.001$ with $n = 101$. What happens? Next try homogeneous boundary conditions. Describe the numerical solution. Compare it with the solution of the usual central difference scheme with $n = 1001$.

P12 Solve the Poisson equation on $(0, 1)^2$ with a homogeneous Dirichlet boundary condition using cubic splines. We illustrate this method for the 1D case. Consider a uniform grid with nodes $x_j = jh$, $0 \leq j \leq n$, where $h = 1/n$. On $[x_j, x_{j+1}]$, for $0 \leq j \leq n - 1$, define the numerical solution as the cubic polynomial $u_j(x) = a_j(x - x_j)^3 + b_j(x - x_j)^2 + c_j(x - x_j) + d_j$. The coefficients are determined by the boundary conditions $u_0(0) = 0$, $u_{n-1}(1) = 0$; continuity of the solution and its derivative: $u_j(x_{j+1}) = u_{j+1}(x_{j+1})$ and $u_j'(x_{j+1}) = u_{j+1}'(x_{j+1})$ for $0 \leq j \leq n - 2$; and satisfaction of the differential equation at the nodes: $-u_j''(x_j) = f(x_j)$, $-u_j''(x_{j+1}) = f(x_{j+1})$ for $0 \leq j \leq n - 1$. The coefficients a_j and b_j can be solved easily. After solving the remaining system for c, write the final system $Ad = F$. How does this differ from the usual finite difference equations? Demonstrate the second-order convergence of the method numerically.

Notes: There are many good books on finite difference methods. We mention just a few: [3], [54], [82], [99]. The books [81], [87] and [92] contain much more thorough studies of convection-diffusion equations. Level set methods can be used to generate a grid on an arbitrary domain. See [95] for an excellent introduction. A reference book on grid generation is [104]. [6] and [23] offer nice discussions of the mathematics involved in image processing. The article [73] introduces the elegant scheme for the disk which avoids the origin, although that scheme is not symmetric. See [83] for more on the ghost fluid method.