Chapter 7

Boundary Value Problems (BVPs)

Definition 7.1. A *partial differential equation* (PDE) is an equation involving an unknown function of two or more variables and some of its partial derivatives.

Definition 7.2. *Laplace equation* is a second-order PDE of the form

$$\Delta u(\mathbf{x}) = 0, \tag{7.1}$$

where the unknown is a function $u : \overline{\Omega} \to \mathbb{R}$, $\overline{\Omega}$ is the closure of an open set $\Omega \subset \mathbb{R}^n$, and the *Laplacian operator* $\Delta : \mathcal{C}^2(\Omega) \to \mathcal{C}(\Omega)$ is

$$\Delta := \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}. \tag{7.2}$$

Example 7.3. The fundamental theorem of vector calculus (a.k.a. Helmholtz's theorem) states that a sufficiently continuous vector field $\mathbf{v}^*$ on a bounded domain can be uniquely decomposed into a divergence-free part and a curl-free part:

$$\begin{cases} \mathbf{v}^* &= \mathbf{v} + \nabla\phi, \\ \nabla\cdot\mathbf{v} &= 0, \qquad \nabla\times\nabla\phi = \mathbf{0}. \end{cases} \tag{7.3}$$

Potential flow is a special type of flow where the velocity $\mathbf{u}$ can be expressed as the gradient of a scalar function:

$$\mathbf{u} = \nabla\varphi,$$

where φ is called the *velocity potential*. For incompressible fluids with $\nabla\cdot\mathbf{u} = 0$, the velocity potential satisfies a Laplace equation $\Delta\varphi = 0$.

Definition 7.4. *Poisson's equation* is a second-order PDE of the form

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \tag{7.4}$$

where the unknown is a function $u : \overline{\Omega} \to \mathbb{R}$, $\overline{\Omega}$ is the closure of an open set $\Omega \subset \mathbb{R}^n$, and the RHS function $f : \Omega \to \mathbb{R}$ is given a priori.

Definition 7.5. A *boundary value problem* (BVP) is a differential equation together with a set of additional constraints, called the *boundary conditions*, that hold only on the domain boundary.

Definition 7.6. Common types of boundary conditions for a one-dimensional interval $\Omega = (a, b)$ are

- *Dirichlet conditions*: $u(a) = \alpha$ and $u(b) = \beta$;
- *Mixed conditions*: $u(a) = \alpha$ and $\left.\frac{\partial u}{\partial x}\right|_b = \beta$;
- *Neumann conditions*: $\left.\frac{\partial u}{\partial x}\right|_a = \alpha$ and $\left.\frac{\partial u}{\partial x}\right|_b = \beta$.

Example 7.7. The Helmholtz decomposition in Example 7.3 is realized by solving the Neumann BVP

$$\Delta\phi = \nabla\cdot\mathbf{v}^* \qquad \text{in } \Omega, \tag{7.5a}$$

$$\mathbf{n}\cdot\nabla\phi = \mathbf{n}\cdot(\mathbf{v}^* - \mathbf{v}) \qquad \text{on } \partial\Omega. \tag{7.5b}$$

7.1 Finite difference (FD) methods

Formula 7.8. In solving a linear BVP, the general procedures of an FD method are as follows.

(FD-1) Discretize the problem domain by a grid.

(FD-2) Approximate each spatial derivative in the PDE with some finite difference formula at every grid point to get a system of linear equations $A\mathbf{U} = \mathbf{F}$ where the vector $\mathbf{U}$ approximates the unknown variable on the grid while the vector $\mathbf{F}$ contains given conditions of the BVP such as boundary conditions and derivatives of the unknown function.

(FD-3) Solve the system of algebraic equations.

Example 7.9 (An FD method for Poisson's equation in a unit interval)**.** Consider the one-dimensional BVP

$$-u''(x) = f(x) \text{ in } \Omega := (0, 1) \tag{7.6}$$

with Dirichlet boundary conditions

$$u(0) = \alpha, \quad u(1) = \beta. \tag{7.7}$$

The general procedures of an FD method based on the central difference are as follows.

(a) Discretize Ω by a Cartesian grid with uniform spacing,

$$x_j = jh, \qquad h = \frac{1}{m+1}, \quad j = 0, 1, \ldots, m+1.$$

Set $U_0 = \alpha$, $U_{m+1} = \beta$ and we will compute m values $U_1, \ldots, U_m$ where each U_j approximates $u(x_j)$.

(b) Approximate the second derivative u'' with a centered difference

$$u''(x_j) = \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2} + O(h^2) \quad (7.8)$$

and we get the following system of linear equations:

$$\begin{aligned} -\frac{\alpha - 2U_1 + U_2}{h^2} &= f(x_1), \\ -\frac{U_{j-1} - 2U_j + U_{j+1}}{h^2} &= f(x_j), \quad j = 2, \cdots, m-1, \\ -\frac{U_{m-1} - 2U_m + \beta}{h^2} &= f(x_m). \end{aligned}$$

These equations are written in the form

$$A\mathbf{U} = \mathbf{F}, \quad (7.9)$$

where

$$\mathbf{U} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_{m-1} \\ U_m \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} f(x_1) + \frac{\alpha}{h^2} \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) + \frac{\beta}{h^2} \end{bmatrix}, \quad (7.10)$$

$$A = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}. \quad (7.11)$$

(c) Solve the linear system (7.9).

7.2 Errors and consistency

Definition 7.10. The *global error* or *solution error* of an FD method in Formula 7.8 is

$$\mathbf{E} = \mathbf{U} - \hat{\mathbf{U}}, \quad (7.12)$$

where $\hat{\mathbf{U}} = [u(x_1), u(x_2), \cdots, u(x_m)]^T$ is the vector of true values and $\mathbf{U}$ the computed solution.

Definition 7.11. A *grid function* is a function $\mathbf{g} : \mathbf{X} \to \mathbb{R}$ on a discrete grid $\mathbf{X}$ that contains a finite number of points.

Definition 7.12. The *q-norm of a grid function* $\mathbf{g}$ on a one-dimensional grid $\mathbf{X} := \{x_1, x_2, \ldots, x_N\}$ is

$$\|\mathbf{g}\|_q = \left(h \sum_{i=1}^{N} |g_i|^q \right)^{\frac{1}{q}}, \quad (7.13)$$

where $\mathbf{g} = (g_1, g_2, \cdots, g_N)$. In particular, the *1-norm* is

$$\|\mathbf{g}\|_1 = h \sum_{i=1}^{N} |g_i| \quad (7.14)$$

and the *max-norm* is

$$\|\mathbf{g}\|_\infty = \max_{1 \le i \le N} |g_i|. \quad (7.15)$$

Exercise 7.13. Suppose a grid function $\mathbf{g} : \mathbf{X} \to \mathbb{R}$ has $\mathbf{X} := \{x_1, x_2, \ldots, x_N\}$, $g_1 = O(h)$, $g_N = O(h)$, and $g_j = O(h^2)$ for all $j = 2, \ldots, N-1$. Show that

$$\|\mathbf{g}\|_\infty = O(h), \ \|\mathbf{g}\|_1 = O(h^2), \ \|\mathbf{g}\|_2 = O(h^{\frac{3}{2}}). \quad (7.16)$$

As the main point of this exercise, the differences in the max-norm, 1-norm, and 2-norm of a grid function often reveal the percentage of components with large magnitude.

Definition 7.14. The *local truncation error* (LTE) of an FD method in Formula 7.8 is the error caused by replacing a continuous derivative with an FD formula.

Example 7.15. When we approximate Δu with

$$D^2 u(x_j) := \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2}, \quad (7.17)$$

the LTE of the FD method in Example 7.9 is

$$\tau_j = -D^2 u(x_j) - \big(-u''(x_j)\big) = -\frac{h^2}{12} u''''(x_j) + O(h^4).$$

Lemma 7.16. The LTE of an FD method in Formula 7.8 is the error of calculating the righthand side function $\mathbf{F}$ in (7.9) by replacing U_j with the exact solution $u(x_j)$,

$$\boldsymbol{\tau} = A\hat{\mathbf{U}} - \mathbf{F}, \quad (7.18)$$

where $\hat{\mathbf{U}}$ is the vector of true solution values.

Proof. By (7.6), any component F_j of $\mathbf{F}$ is

$$F_j = f(x_j) = -\Delta u(x_j).$$

The rest of the proof follows from Definition 7.14. □

Lemma 7.17. The LTE and the global error are related as

$$A\mathbf{E} = -\boldsymbol{\tau}. \quad (7.19)$$

Proof. $A\mathbf{E} = A(\mathbf{U} - \hat{\mathbf{U}}) = \mathbf{F} - (\mathbf{F} + \boldsymbol{\tau}) = -\boldsymbol{\tau}$. □

Definition 7.18. An FD method in Formula 7.8 is said to be *consistent* with the BVP if

$$\lim_{h \to 0} \|\boldsymbol{\tau}^h\| = 0, \quad (7.20)$$

where $\boldsymbol{\tau}^h$ is the LTE.

7.3 Stability and convergence

Definition 7.19. An FD method is *convergent* if

$$\lim_{h \to 0} \|\mathbf{E}^h\| = 0, \quad (7.21)$$

where $\mathbf{E}^h$ is the solution error in Definition 7.10.

Definition 7.20. An FD method in Formula 7.8 is *stable* if

(a) $\lim_{h \to 0} A$ is invertible,

(b) $\lim_{h \to 0} \|A^{-1}\| = O(1)$.

Theorem 7.21. A consistent and stable FD method is convergent.

Proof. Lemma 7.17 and Definitions 7.18 and 7.20 yield

$$\lim \|\mathbf{E}^h\| \le \lim \|(A^h)^{-1}\| \lim \|\boldsymbol{\tau}^h\| \le C \lim \|\boldsymbol{\tau}^h\| = 0. \quad □$$

7.3.1 Convergence in the 2-norm

Definition 7.22 (Matrix norms induced by vector norms)**.** The *norm* of a matrix $A \in \mathbb{R}^{n\times n}$ is defined by

$$\begin{aligned}\|A\| &= \sup\left\{\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}\right\}\\ &= \sup\left\{\|A\mathbf{x}\| : \|\mathbf{x}\| = 1\right\}.\end{aligned}$$

Example 7.23. Commonly used matrix norms include

$$\begin{cases}\|A\|_\infty = \max_{1\le i\le n}\sum_{j=1}^n |a_{ij}|,\\ \|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^n |a_{ij}|,\\ \|A\|_2 = \sqrt{\rho(A^TA)},\end{cases} \tag{7.22}$$

where $\rho(B) := \max_i |\lambda_i(B)|$ is the spectral radius of the matrix B, i.e. the maximum modulus of eigenvalues of B.

Lemma 7.24. The eigenvalues λ_k and eigenvectors $\mathbf{w}_k$ of the matrix A in (7.11) are

$$\lambda_k(A) = \frac{4}{h^2}\sin^2\frac{k\pi}{2(m+1)}, \tag{7.23}$$

$$w_{k,j} = \sin\frac{jk\pi}{m+1}, \tag{7.24}$$

where $j, k = 1, 2, \ldots, m$.

Proof. It is straightforward to verify the conclusions using the trigonometric identity

$$\sin\alpha + \sin\beta = 2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}.$$ □

Theorem 7.25. The FD method in Example 7.9 is second-order convergent in the 2-norm.

Proof. We have $\|A\|_2 = \rho(A)$ since the matrix A is symmetric. Then Lemma 7.24 yields

$$\lim_{h\to0}\|A^{-1}\|_2 = \lim_{h\to0}\frac{1}{\min|\lambda_p|} = \lim_{h\to0}\frac{h^2}{4\sin^2\frac{\pi h}{2}} = \frac{1}{\pi^2} = O(1).$$

By Definition 7.20, the method is stable. The rest of the proof follows from Example 7.15, Definition 7.18, and Theorem 7.21. □

7.3.2 Green's function

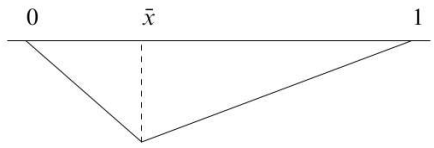
Definition 7.26. For any fixed $\bar{x}\in[0,1]$, the *Green's function* $G(x;\bar{x})$ is the function of x that solves the BVP

$$\begin{cases}u''(x) = \delta(x-\bar{x});\\ u(0) = u(1) = 0,\end{cases} \tag{7.25}$$

where $\delta(x-\bar{x})$ is the Dirac delta function in Definition 5.40.

Lemma 7.27. The Green's function $G(x;\bar{x})$ that solves (7.25) is

$$G(x;\bar{x}) = \begin{cases}(\bar{x}-1)x, & x\in[0,\bar{x}],\\ \bar{x}(x-1), & x\in[\bar{x},1].\end{cases} \tag{7.26}$$



Proof. For any fixed ϵ, we have from (7.25) and (5.39b) that

$$\begin{aligned}\int_{x_0-\epsilon}^{x_0+\epsilon} G''(x)\mathrm{d}x &= \int_{x_0-\epsilon}^{x_0+\epsilon}\delta(x-\bar{x})\mathrm{d}x\\ &= \begin{cases}0, & \bar{x}\notin(x_0-\epsilon, x_0+\epsilon),\\ 1, & \bar{x}\in(x_0-\epsilon, x_0+\epsilon).\end{cases}\end{aligned}$$

Take limit $\epsilon\to0$ of the above equation and we deduce from the second fundamental theorem of calculus (Theorem C.73) that

$$\begin{aligned}&\lim_{\epsilon\to0}G'(x_0+\epsilon) - \lim_{\epsilon\to0}G'(x_0-\epsilon)\\ &= \begin{cases}0 & \text{if } x_0\in(0,\bar{x})\cup(\bar{x},1),\\ 1 & \text{if } x_0 = \bar{x}.\end{cases}\end{aligned} \tag{7.27}$$

Substitute

$$G(x;\bar{x}) = \begin{cases}ax+b, & x\in[0,\bar{x}],\\ cx+d, & x\in[\bar{x},1]\end{cases}$$

into (7.27) and (7.25) and the continuity of $G(x;\bar{x})$ yields

$$\begin{cases}c = a+1\\ b = 0\\ c+d = 0\\ a\bar{x}+b = c\bar{x}+d\end{cases} \Rightarrow \begin{cases}a = \bar{x}-1\\ b = 0\\ c = \bar{x}\\ d = -\bar{x}\end{cases},$$

which completes the proof. □

Corollary 7.28. The solution to the linear BVP

$$\begin{cases}u''(x) = c\delta(x-\bar{x}),\\ u(0) = u(1) = 0.\end{cases}$$

is

$$u(x) = cG(x;\bar{x}).$$

Proof. This follows directly from Lemma 7.27. □

7.3.3 Convergence in the max-norm

Lemma 7.29. For the matrix A in (7.11), any element of its inverse $B = A^{-1}$ is

$$b_{ij} = -hG(x_i;x_j) = \begin{cases}-h(x_j-1)x_i, & i\le j,\\ -hx_j(x_i-1), & i\ge j.\end{cases} \tag{7.28}$$

More explicitly, the matrix B is

$$B = -h\begin{bmatrix}x_1(x_1-1) & x_1(x_2-1) & \cdots & x_1(x_m-1)\\ x_1(x_2-1) & x_2(x_2-1) & \cdots & x_2(x_m-1)\\ \vdots & \vdots & \ddots & \vdots\\ x_1(x_m-1) & x_2(x_m-1) & \cdots & x_m(x_m-1)\end{bmatrix}.$$

Proof. To verify that B is indeed the inverse of A, it suffices to multiply the ith row of h^2A and the jth column of $-\frac{1}{h}B$,

$$\begin{bmatrix}0,\cdots,0,-1,2,-1,0,\cdots,0\end{bmatrix}\begin{bmatrix}x_1(x_j-1)\\ \vdots\\ x_{j-1}(x_j-1)\\ x_j(x_j-1)\\ x_j(x_{j+1}-1)\\ \vdots\\ x_j(x_m-1)\end{bmatrix},$$

the only nonzero case is when $i=j$:

$$2x_j(x_j-1)-x_{j-1}(x_j-1)-x_j(x_{j+1}-1)=-h. \qquad \square$$

Theorem 7.30. The max-norm of $B=A^{-1}$ satisfies

$$\|B\|_\infty=\max_{1\le i\le m}\sum_{j=1}^m|b_{ij}|\le 1. \tag{7.29}$$

Proof. Lemma 7.29 yields

$$\begin{aligned}\sum_{j=1}^m|b_{ij}|&=\sum_{j=1}^i hx_j|x_i-1|+\sum_{j=i+1}^m hx_i|x_j-1|\\ &\le\sum_{j=1}^i h\left(\frac{m}{m+1}\right)^2+\sum_{j=i+1}^m h\left(\frac{m}{m+1}\right)^2\\ &=mh\left(\frac{m}{m+1}\right)^2=\left(\frac{m}{m+1}\right)^3\le 1. \qquad \square\end{aligned}$$

7.4 A solution via Green's function

Lemma 7.31. Suppose $\mathcal{L}$ is an invertible linear differential operator that satisfies

$$\mathcal{L}u(x)=f(x). \tag{7.30}$$

Then we have

$$u(x)=\int G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x}, \tag{7.31}$$

where G is the Green's function satisfying

$$\mathcal{L}G(x;\bar{x})=\delta(x-\bar{x}). \tag{7.32}$$

Proof. Multiply (7.32) by $f(\bar{x})$, integrate w.r.t. $\bar{x}$, and we have

$$\int\mathcal{L}G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x}=\int\delta(x-\bar{x})f(\bar{x})\mathrm{d}\bar{x}=f(x),$$

where the second equality follows from the sifting property of the Dirac delta function (Lemma 5.42). Therefore

$$\mathcal{L}\int G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x}=f(x),$$

which further implies (7.31) since $\mathcal{L}$ is invertible. $\square$

Theorem 7.32. The Dirichlet BVP

$$\begin{cases}u''(x)=f(x),\\ u(0)=\alpha,u(1)=\beta\end{cases} \tag{7.33}$$

is solved by

$$u(x)=\alpha G_0(x)+\beta G_1(x)+\hat{U}(x), \tag{7.34}$$

where $G_0(x),G_1(x)$ and $\hat{U}(x)$ are defined by BVPs as follows

$$\begin{cases}G_0''(x)=0,\\ G_0(0)=1,G_0(1)=0\end{cases}\Rightarrow\quad G_0(x)=1-x, \tag{7.35a}$$

$$\begin{cases}G_1''(x)=0,\\ G_1(0)=0,G_1(1)=1\end{cases}\Rightarrow\quad G_1(x)=x, \tag{7.35b}$$

$$\begin{cases}\hat{U}''(x)=f(x),\\ \hat{U}(0)=0,\hat{U}(1)=0\end{cases}\Rightarrow\quad \hat{U}(x)=\int_0^1 f(\bar{x})G(x;\bar{x})\mathrm{d}\bar{x}. \tag{7.35c}$$

Proof. This follows from the linearity of the BVP (7.33). $\square$

7.5 Other boundary conditions

Example 7.33. Consider the second-order BVP

$$u''(x)=f(x)\quad\text{in }(0,1) \tag{7.36}$$

with mixed boundary conditions

$$u'(0)=\sigma,\quad u(1)=\beta. \tag{7.37}$$

As the crucial difference between this BVP and the BVP with pure Dirichlet conditions in Example 7.9, the value of $u(x)$ at $x=0$ becomes an unknown to be solved for.

The first approach is to use a one-sided expression

$$\frac{U_1-U_0}{h}=\sigma \tag{7.38}$$

to arrive at

$$A_E\mathbf{U}_E=\mathbf{F}_E \tag{7.39}$$

where

$$A_E=\frac{1}{h^2}\begin{bmatrix}-h & h & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1\\ & & & & & 0 & h^2\end{bmatrix},$$

$$\mathbf{U}_E=\begin{bmatrix}U_0\\ U_1\\ U_2\\ \vdots\\ U_{m-1}\\ U_m\\ U_{m+1}\end{bmatrix},\qquad \mathbf{F}_E=\begin{bmatrix}\sigma\\ f(x_1)\\ f(x_2)\\ \vdots\\ f(x_{m-1})\\ f(x_m)\\ \beta\end{bmatrix}.$$

The LTE at $x_0 = 0$ is

$$\begin{aligned}\tau_0 &= \frac{1}{h^2}(hu(x_1) - hu(x_0)) - \sigma \\ &= u'(x_0) + \frac{1}{2}hu''(x_0) + O(h^2) - \sigma \\ &= \frac{1}{2}hu''(x_0) + O(h^2),\end{aligned}$$

which is only first order accurate.

The second approach is to extend the domain with a *ghost cell* $x_{-1} = -h$ and use a central difference to obtain

$$\frac{U_1 - U_{-1}}{2h} = \sigma \tag{7.40}$$

that is second-order accurate for the LTE. We do not have any information for U_{-1}, so we want to eliminate it by

$$\frac{1}{h^2}(U_{-1} - 2U_0 + U_1) = f(x_0). \tag{7.41}$$

(7.40) and (7.41) yield

$$\frac{1}{h}(-U_0 + U_1) = \sigma + \frac{h}{2}f(x_0). \tag{7.42}$$

The resulting matrix is the same as that of the first approach in (7.39) except that the first component of $\mathbf{F}_E$ has an additional term $\frac{h}{2}f(x_0)$.

The third approach is to use U_0, U_1, and U_2 to approximate $u'(0)$ and we can get a second-order FD formula, c.f. Example 6.37,

$$-\frac{1}{h}\left(\frac{3}{2}U_0 - 2U_1 + \frac{1}{2}U_2\right) = \sigma + O(h^2). \tag{7.43}$$

This results in the linear system

$$A_F\mathbf{U}_E = \mathbf{F}_E \tag{7.44}$$

where

$$A_F = \frac{1}{h^2}\begin{bmatrix} -\frac{3}{2}h & 2h & -\frac{1}{2}h & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \\ & & & & & 0 & h^2 \end{bmatrix}.$$

Exercise 7.34. Show that the first column of $B_E = A_E^{-1}$ contains elements that are $O(1)$.

Example 7.35. Consider the second-order BVP

$$u''(x) = f(x) \text{ in } (0,1), \tag{7.45}$$

with pure Neumann conditions

$$u'(0) = \sigma_0, \quad u'(1) = \sigma_1. \tag{7.46}$$

To ensure the existence of a solution, the following compatibility condition on $f(x)$, σ_0, and σ_1 must be satisfied:

$$\int_0^1 f(x)\mathrm{d}x = \int_0^1 u''(x)\mathrm{d}x = u'(1) - u'(0) = \sigma_1 - \sigma_0. \tag{7.47}$$

In fact, if (7.47) holds, there are an infinite number of solutions: if v is a solution of (7.45), $v + \mathbb{R}$ are also solutions.

Using procedures similar to those in Example 7.33, we can discretize (7.45) and (7.46) as

$$A_F\mathbf{U}_E = \mathbf{F}_F, \tag{7.48}$$

where

$$A_F = \frac{1}{h^2}\begin{bmatrix} -h & h & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \\ & & & & & h & -h \end{bmatrix}, \tag{7.49}$$

$$\mathbf{F}_F = \begin{bmatrix} \sigma_0 + \frac{h}{2}f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) \\ -\sigma_1 + \frac{h}{2}f(x_{m+1}) \end{bmatrix}.$$

Lemma 7.36. The matrix A_F in (7.49) satisfies

$$\dim \mathcal{N}(A_F) = 1. \tag{7.50}$$

Proof. Clearly, $\mathbf{e} = (1, 1, \cdots, 1)^T$ is in the null space of A_F. The rest follows from the well-posedness of the BVP with mixed conditions. □

Theorem 7.37 (Solvability condition)**.** The linear system (7.48) has a solution if and only if

$$\frac{h}{2}f(x_0) + h\sum_{i=1}^{m} f(x_i) + \frac{h}{2}f(x_{m+1}) = \sigma_1 - \sigma_0. \tag{7.51}$$

Proof. The fundamental theorem of linear algebra (Theorem B.89) implies

$$\mathbb{R}^{m+2} = \mathcal{R}(A_F) \oplus \mathcal{N}(A_F^T) \tag{7.52}$$

and $\dim \mathcal{N}(A_F^T) = \dim \mathcal{N}(A_F)$. Lemma 7.36 further yields $\dim \mathcal{N}(A_F^T) = 1$. Then it is readily verified that

$$\mathcal{N}(A_F^T) = \operatorname{span}\left\{(1, h, h, \ldots, h, 1)^T\right\}.$$

For sufficiency, the above equation and (7.51) imply that $\mathbf{F}_F$ is orthogonal to $\mathcal{N}(A_F^T)$ and thus (7.52) yields $\mathbf{F}_F \in \mathcal{R}(A_F)$. Hence (7.48) must have a solution. As for necessity, the existence of a solution of (7.48) implies $\mathbf{F}_F \in \mathcal{R}(A_F)$, which, together with the above two equations, implies (7.51). □

7.6 BVPs in two dimensions

Example 7.38 (An FD method for Poisson's equation in a unit square)**.** Consider the two-dimensional BVP

$$-\frac{\partial^2}{\partial x^2}u(x,y) - \frac{\partial^2}{\partial y^2}u(x,y) = f(x,y) \tag{7.53}$$

in $\Omega := (0,1)\times(0,1)$ with homogeneous Dirichlet conditions

$$u(x,y)|_{\partial\Omega} = 0. \tag{7.54}$$

A uniform Cartesian grid can be generated with

$$x_i = ih, \quad y_j = jh, \quad i,j = 1,2,\cdots,m, \tag{7.55}$$

where $h = \Delta x = \Delta y = \frac{1}{m+1}$ is the uniform grid size.

Approximate $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ separately and we have, $\forall i,j = 1,2,\ldots,m$,

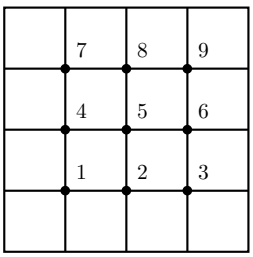
$$-\frac{U_{i-1,j} - 2U_{ij} + U_{i+1,j}}{h^2} - \frac{U_{i,j-1} - 2U_{ij} + U_{i,j+1}}{h^2} = f_{ij}. \tag{7.56}$$

These $m\times m$ equations organize into a single system

$$A_{2D}\mathbf{U} = \mathbf{F}. \tag{7.57}$$

Exercise 7.39. Show that the LTE $\boldsymbol{\tau}$ of the FD method in Example 7.38 is

$$\tau_{i,j} = -\frac{1}{12}h^2\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right)\Big|_{(x_i,y_j)} + O(h^4). \tag{7.58}$$



Example 7.40. For $m = 3$ with ordering as shown above, we have

$$A_{2D} = \frac{1}{h^2}\begin{bmatrix} +4 & -1 & & -1 & & & & & \\ -1 & +4 & -1 & & -1 & & & & \\ & -1 & +4 & & & -1 & & & \\ -1 & & & +4 & -1 & & -1 & & \\ & -1 & & -1 & +4 & -1 & & -1 & \\ & & -1 & & -1 & +4 & & & -1 \\ & & & -1 & & & +4 & -1 & \\ & & & & -1 & & -1 & +4 & -1 \\ & & & & & -1 & & -1 & +4 \end{bmatrix}$$

$$= \frac{1}{h^2}\begin{bmatrix} T & -I & \\ -I & T & -I \\ & -I & T \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} U_{11} \\ U_{21} \\ U_{31} \\ U_{12} \\ U_{22} \\ U_{32} \\ U_{13} \\ U_{23} \\ U_{33} \end{bmatrix},$$

where

$$T = \begin{bmatrix} +4 & -1 & 0 \\ -1 & +4 & -1 \\ 0 & -1 & +4 \end{bmatrix}. \tag{7.59}$$

where $\mathbf{U}$ is obtained by stacking the columns on top of each other.

Lemma 7.41. Let $\mathbf{E} = \mathbf{U} - \hat{\mathbf{U}}$ denote the global error of the linear system (7.57). Then the LTE (7.58) satisfies

$$A_{2D}\mathbf{E} = -\boldsymbol{\tau}. \tag{7.60}$$

Proof. The proof is the same as that of Lemma 7.17. □

7.6.1 Kronecker product

Definition 7.42. The *Kronecker product* of two matrices $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$ is another matrix $A\otimes B \in \mathbb{C}^{mp\times nq}$ given by

$$A\otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}, \tag{7.61}$$

where a_{ij} is the (i,j)th element of A.

Example 7.43.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \otimes \begin{bmatrix} 0 & 5 \\ 6 & 7 \end{bmatrix} = \begin{bmatrix} 0 & 5 & 0 & 10 \\ 6 & 7 & 12 & 14 \\ 0 & 15 & 0 & 20 \\ 18 & 21 & 24 & 28 \end{bmatrix}.$$

Definition 7.44. For $X \in \mathbb{C}^{m\times n}$, $\mathrm{vec}(X)$ is defined to be a column vector of size mn made of the columns of X stacked on top of one another from left to right.

Lemma 7.45. Any $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$, and $X \in \mathbb{R}^{m\times n}$ satisfy

$$\mathrm{vec}(AX) = (I_n \otimes A)\mathrm{vec}(X), \tag{7.62}$$

$$\mathrm{vec}(XB) = (B^T \otimes I_m)\mathrm{vec}(X). \tag{7.63}$$

Proof. We have

$$\begin{aligned} \mathrm{vec}(AX) &= \mathrm{vec}([A\mathbf{X}_1, A\mathbf{X}_2, \cdots, A\mathbf{X}_n]) \\ &= \begin{bmatrix} A\mathbf{X}_1 \\ A\mathbf{X}_2 \\ \vdots \\ A\mathbf{X}_n \end{bmatrix} = \begin{bmatrix} A & & & \\ & A & & \\ & & \ddots & \\ & & & A \end{bmatrix}\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \\ &= (I_n \otimes A)\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}. \end{aligned}$$

Let $Y = XB$, then

$$\mathbf{Y}_j = X\mathbf{b}_j \Rightarrow y_{kj} = \sum_{i=1}^{n} x_{ki}b_{ij}. \tag{7.64}$$

Let $C = B^T \otimes I_m$, then the (i,j)-th sub-block of C is

$$C_{ij} = b_{ji} I_m. \tag{7.65}$$

Let $D = C\mathrm{vec}(X)$, then the j-th block of D is

$$\mathbf{D}_j = \sum_{i=1}^{n} C_{ji}\mathbf{X}_i = \sum_{i=1}^{n} b_{ij} I_m \mathbf{X}_i = \sum_{i=1}^{n} b_{ij}\mathbf{X}_i, \tag{7.66}$$

and the (k,j)-th entry of D is (Here we also use (k,j) to denote the scalar index corresponding to the multi-index (k,j).)

$$d_{(k,j)} = \sum_{i=1}^{n} b_{ij} x_{ki} = \sum_{i=1}^{n} x_{ki} b_{ij}, \tag{7.67}$$

Combining (7.64) and (7.67) yields (7.63). □

7.6.2 Convergence in the 2-norm

Lemma 7.46. The linear system (7.57) is equivalent to

$$AU_{m\times m} + U_{m\times m}A = F_{m\times m}, \tag{7.68}$$

where the (i,j)th element of $U_{m\times m}$ is the computed solution at the (i,j)th grid point, the (i,j)th element of $F_{m\times m}$ is

$$(F_{m\times m})_{ij} = f(ih, jh),$$

and A is the 1D discrete Laplacian in (7.11).

Proof. A direct computation gives

$$\begin{cases} (AU_{m\times m})_{ij} = \dfrac{1}{h^2}(-U_{i-1,j} + 2U_{ij} - U_{i+1,j}), \\ (U_{m\times m}A)_{ij} = \dfrac{1}{h^2}(-U_{i,j-1} + 2U_{ij} - U_{i,j+1}), \end{cases} \tag{7.69}$$

and the *homogeneous* Dirichlet condition yields

$$AU_{m\times m} + U_{m\times m}A = F_{m\times m}.$$ □

Lemma 7.47. The 1D discrete Laplacian A in (7.11) satisfies

$$\mathrm{vec}(AU_{m\times m} + U_{m\times m}A) = (I_m \otimes A + A \otimes I_m)\mathrm{vec}(U_{m\times m}).$$

Proof. By Lemma 7.45, we have

$$\mathrm{vec}(AU_{m\times m}) = (I_m \otimes A)\mathrm{vec}(U_{m\times m}),$$

and

$$\mathrm{vec}(U_{m\times m}A) = (A^T \otimes I_m)\mathrm{vec}(U_{m\times m}) = (A \otimes I_m)\mathrm{vec}(U_{m\times m}),$$

where the second equality follows from the symmetry of A. Adding these two equations gives the desired result. □

Theorem 7.48. With matrix ordering, the linear system (7.57) can be written as

$$A_{2D} = I_m \otimes A + A \otimes I_m,\ \mathbf{U} = \mathrm{vec}(U_{m\times m}),\ \mathbf{F} = \mathrm{vec}(F_{m\times m}).$$

Proof. This follows from Lemma 7.46 and Lemma 7.47. □

Definition 7.49. The *discrete Laplacian* in n-dimensional space analogous to the 1D discrete Laplacian (7.11) is

$$A_{nD} = \sum_{j=0}^{n-1} \underbrace{I_m \otimes \cdots \otimes I_m}_{\#I_m = j} \otimes A \otimes \underbrace{I_m \otimes \cdots \otimes I_m}_{\#I_m = n-j-1}. \tag{7.70}$$

Example 7.50. For $n = 3$, we have

$$A_{3D} = A \otimes I_m \otimes I_m + I_m \otimes A \otimes I_m + I_m \otimes I_m \otimes A.$$

Theorem 7.51. The eigen-pairs of A_{2D} are

$$\lambda_{ij} = \lambda_i + \lambda_j, \quad \mathbf{W}_{ij} = \mathrm{vec}\left(\mathbf{w}_i \mathbf{w}_j^T\right), \tag{7.71}$$

where $i, j = 1, 2, \ldots, m$ and $(\lambda_i, \mathbf{w}_i)$ is an eigen-pair of A in Lemma 7.24.

Proof. By Lemma 7.24, we have

$$\begin{aligned} A\mathbf{w}_i\mathbf{w}_j^T + \mathbf{w}_i\mathbf{w}_j^T A &= \lambda_i \mathbf{w}_i \mathbf{w}_j^T + \mathbf{w}_i \mathbf{w}_j^T A^T \\ &= \lambda_i \mathbf{w}_i \mathbf{w}_j^T + \mathbf{w}_i (A\mathbf{w}_j)^T \\ &= \lambda_i \mathbf{w}_i \mathbf{w}_j^T + \lambda_j \mathbf{w}_i \mathbf{w}_j^T \\ &= (\lambda_i + \lambda_j)\mathbf{w}_i \mathbf{w}_j^T. \end{aligned}$$

Then Theorem 7.48 and Lemma 7.47 yield

$$\begin{aligned} A_{2D}\mathrm{vec}\left(\mathbf{w}_i\mathbf{w}_j^T\right) &= (I_m \otimes A + A \otimes I_m)\mathrm{vec}\left(\mathbf{w}_i\mathbf{w}_j^T\right) \\ &= \mathrm{vec}\left(A(\mathbf{w}_i\mathbf{w}_j^T) + (\mathbf{w}_i\mathbf{w}_j^T)A\right) \\ &= (\lambda_i + \lambda_j)\mathrm{vec}\left(\mathbf{w}_i\mathbf{w}_j^T\right), \end{aligned}$$

and hence $\lambda_i + \lambda_j$ is an eigenvalue of A_{2D} with corresponding eigenvector $\mathbf{W}_{ij}$. □

Theorem 7.52. The FD method in Example 7.38 is second-order convergent in the 2-norm.

Proof. We have $\|A_{2D}\|_2 = \rho(A_{2D})$ since A_{2D} is symmetric. Then Theorem 7.51 yields

$$\lim_{h\to 0} \|A_{2D}^{-1}\|_2 = \lim_{h\to 0} \frac{1}{\min|\lambda_{ij}|} = \lim_{h\to 0} \frac{h^2}{8\sin^2\frac{\pi h}{2}} = \frac{1}{2\pi^2} = O(1).$$

By Definition 7.20, the method is stable. The proof is completed by (7.58), Definition 7.18, Theorem 7.21, and Lemma 7.41. □

7.6.3 Convergence in the max-norm via a discrete maximum principle

Theorem 7.53. The FD method in Example 7.38 is second-order convergent in the max-norm.

Proof. Define a *comparison function* ϕ as

$$\phi(x,y) := \left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2 \tag{7.72}$$

and write $\phi_{i,j} = \phi(ih, jh)$. By Example 7.38, A_{2D} can be expressed componentwise as

$$\begin{aligned} A_{2D}U_{i,j} &:= (A_{2D}\mathbf{U})_{i,j} \\ &= \frac{1}{h^2}(4U_{i,j} - U_{i+1,j} - U_{i-1,j} - U_{i,j+1} - U_{i,j-1}). \end{aligned} \tag{7.73}$$

Then (7.58), (7.72), and (7.73) yield

$$\forall i,j = 1,2,\ldots,m, \quad A_{2D}\phi_{i,j} = -4. \tag{7.74}$$

Construct a grid function

$$\psi_{i,j} := E_{i,j} + \frac{1}{4}\tau_m \phi_{i,j}, \tag{7.75}$$

where $\tau_m := \|\boldsymbol{\tau}\|_\infty$ and $E_{i,j}$ is the solution error at the grid points (ih, jh) with $i,j = 0,1,\cdots,m+1$. At each interior grid point (x_i, y_j) in (7.55), we have

$$A_{2D}\psi_{i,j} = A_{2D}E_{i,j} + \frac{1}{4}\tau_m A_{2D}\phi_{i,j} = -\tau_{i,j} - \tau_m \le 0.$$

By (7.73), $A_{2D}\psi_{i,j} \le 0$ dictates that $\psi_{i,j}$ be no greater than at least one of its neighbors $\psi_{i+1,j}$, $\psi_{i-1,j}$, $\psi_{i,j+1}$, and $\psi_{i,j-1}$. Therefore, *the maximum value of ψ must occur at a boundary point*, i.e., a point with $i = 0, m+1$ or $j = 0, m+1$. Consequently, there exists some constant $C > 0$ such that

$$E_{i,j} \le \psi_{i,j} \le \frac{1}{8}\tau_m < Ch^2,$$

where the first step follows from (7.75) and $\tau_m\phi_{i,j} \ge 0$, the second step from $E_{i,j}$ being zero at all boundary points and the fact that the maximum of ϕ is $\frac{1}{2}$ at the domain corners, and the last step from (7.58).

By similar arguments, the grid function

$$\chi_{i,j} := -E_{i,j} + \frac{1}{4}\tau_m \phi_{i,j} \tag{7.76}$$

satisfies $A_{2D}\chi_{i,j} \le 0$ for all grid points and thus

$$-E_{i,j} \le \chi_{i,j} \le \frac{1}{8}\tau_m < Ch^2.$$

To sum up, we have $|E_{i,j}| = O(h^2)$ for all grid points. □

Notation 4. Consider discretizing a BVP on domain Ω. Denote by $\mathbf{X}_\Omega$ the set of *equation-discretization points* where the BVP is discretized and where values of the unknown function u are sought. Let $\mathbf{X}_{\partial\Omega} \subset \partial\Omega$ be the set of *boundary points* so that each point $Q \in \mathbf{X}_{\partial\Omega}$ satisfies

- the BVP is not discretized at Q,
- $u(Q)$ is prescribed by a Dirichlet condition,
- this Dirichlet condition at Q is involved in discretizing the BVP at some $P \in X_\Omega$.

Example 7.54. A boundary point in Notation 4 must be on the domain boundary $\partial\Omega$, but an equation-discretization point P might also belong to $\partial\Omega$: if we only have a Neumann condition for P, then the *value* of the unknown at P might still be needed. In addition, a grid $\mathbf{X}$ might contain some point that belongs to neither $\mathbf{X}_\Omega$ nor $\mathbf{X}_{\partial\Omega}$.

Lemma 7.55 (Discrete maximum principle)**.** Suppose that an FD discretization of a BVP yields

$$\forall P \in \mathbf{X}_\Omega, \quad L_h U_P - f_P + g_P = 0, \tag{7.77}$$

where f_P corresponds to the RHS of (7.4), g_P corresponds to all boundary data other than Dirichlet conditions, and L_h and $\mathbf{X}_\Omega$ satisfy

(DMP-1) for each equation-discretization point $P \in \mathbf{X}_\Omega$, L_h is of the form

$$L_h U_P = c_P U_P - \sum_{Q \in Q_P} c_Q U_Q, \tag{7.78}$$

where $Q_P \subset \mathbf{X}_\Omega \cup \mathbf{X}_{\partial\Omega}$, $c_P > 0$ and each $c_Q > 0$. The set $\{P\} \cup Q_P$ in (7.78) is called the *P-stencil* or the *stencil of L_h at P*;

(DMP-2) $\forall P \in \mathbf{X}_\Omega$, $c_P \ge \sum_{Q \in Q_P} c_Q$;

(DMP-3) $\mathbf{X}_\Omega$ is *connected*, i.e.,

$$\begin{aligned}&\forall P_0, P_m \in \mathbf{X}_\Omega, \exists P_1,\ P_2,\ \ldots,\ P_{m-1} \text{ s.t.}\\ &\forall r = 1,2,\ldots,m,\ P_r \text{ is in the } P_{r-1}\text{-stencil;}\end{aligned} \tag{7.79}$$

(DMP-4) at least one equation (7.78) involves a boundary value U_Q given by a Dirichlet condition.

Then for any grid function $\psi : \mathbf{X} \to \mathbb{R}$ satisfying

$$\begin{cases} \max_{P \in \mathbf{X}} \psi_P \ge 0, \\ \forall P \in \mathbf{X}_\Omega, \quad L_h \psi_P \le 0, \end{cases} \tag{7.80}$$

we have

$$\max_{P \in \mathbf{X}_\Omega} \psi_P \le \max_{Q \in \mathbf{X}_{\partial\Omega}} \psi_Q. \tag{7.81}$$

Proof. Suppose

$$M_\Omega := \max_{Q \in \mathbf{X}_\Omega} \psi_Q > M_{\partial\Omega} := \max_{Q \in \mathbf{X}_{\partial\Omega}} \psi_Q$$

and let P be the point where ψ attains M_Ω. Then

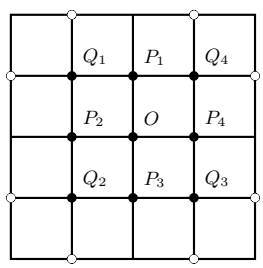
$$(*): \quad M_\Omega = \psi_P \le \frac{1}{c_P}\sum_Q c_Q \psi_Q \le \frac{1}{c_P}\sum_Q c_Q M_\Omega \le M_\Omega,$$

where the first inequality follows from (7.80) and (7.78), the second from the definition of M_Ω, and the third from (DMP-2) and $M_\Omega \ge 0$. For the second "$\le$" in $(*)$ to be "$=$," we must have $\psi_Q = \psi_P$ for each Q. By (DMP-3), U takes the same value M_Ω on all equation-discretization points. Then (DMP-4) implies $M_\Omega = M_{\partial\Omega}$, which contradicts the starting point $M_\Omega > M_{\partial\Omega}$. □

Example 7.56. Suppose the concept of connectedness in (DMP-3) were defined as

$$\begin{aligned}&\forall P_0, P_m \in \mathbf{X}_\Omega, \exists P_1,\ P_2,\ \ldots,\ P_{m-1} \text{ s.t. } \forall r = 1,2,\ldots,m,\\ &\text{both } U_{P_r} \text{ and } U_{P_{r-1}} \text{ appear in some equation (7.77).}\end{aligned} \tag{7.82}$$

Then the conclusion (7.81) of Lemma 7.55 would not hold. The following is a counter-example.



Let L_h be

$$\begin{aligned} L_h\psi_{P_1} &= 3\psi_{P_1} - \psi_{P_2} - \psi_{P_3} - \psi_{P_4}, \\ L_h\psi_{P_2} &= 3\psi_{P_2} - \psi_{P_1} - \psi_{P_3} - \psi_{P_4}, \\ L_h\psi_{P_3} &= 3\psi_{P_3} - \psi_{P_1} - \psi_{P_2} - \psi_{P_4}, \\ L_h\psi_{P_4} &= 3\psi_{P_4} - \psi_{P_1} - \psi_{P_2} - \psi_{P_3}, \\ L_h\psi_O &= 4\psi_O - \psi_{P_1} - \psi_{P_2} - \psi_{P_3} - \psi_{P_4}, \end{aligned}$$

and the expressions of L_h at Q_i's be similar with that of L_h at O such that (DMP-1,2,4) hold. (7.82) also holds. The following distribution of the grid function ψ,

$$\psi_{\mathbf{x}} = \begin{cases} 10 & \text{if } \mathbf{x} = P_1, P_2, P_3, P_4, \\ 1 & \text{otherwise.} \end{cases}$$

satisfies (7.80), but the conclusion (7.81) does not hold.

The key point of this example is that, in order for two points P_0 and P_m to be called connected in the context of solving elliptic equations, there must simultaneously exist a path from P_0 to P_m and a path from P_m to P_0. The definition of connectedness in (7.82) is not satisfactory because it fails to capture the direction of the path. In contrast, (7.79) captures the dependence of path connectedness on the direction of the path.

Theorem 7.57. Suppose an FD discretization of a BVP satisfies the conditions (DMP-1,2,3,4) in Lemma 7.55. Then the solution error $E_P := U_P - u(P)$ of the FD method (7.77) is bounded by

$$\forall P \in \mathbf{X}, \quad |E_P| \le T_{\max}\left(\max_{Q\in\mathbf{X}_{\partial\Omega}} \phi(Q)\right), \tag{7.83}$$

where $T_{\max} = \max_{P\in\mathbf{X}_\Omega} |T_P|$, T_P is the LTE at P satisfying $L_hE_P = -T_P$, and $\phi : \mathbf{X} \to \mathbb{R}$ is a nonnegative grid function satisfying

$$\forall P \in \mathbf{X}_\Omega, \quad L_h\phi_P \le -1. \tag{7.84}$$

Proof. Lemma 7.17 implies $L_hE_P = -T_P$. Define

$$\psi_P := E_P + T_{\max}\phi_P$$

and we know from (7.84) that

$$L_h\psi_P \le -T_P - T_{\max} \le 0.$$

Furthermore, we have $\max_{P\in\mathbf{X}} \psi_P \ge 0$ because $\phi_P \ge 0$ and

$$(*): \quad \forall Q \in \mathbf{X}_{\partial\Omega},\ E_Q = 0,$$

c.f. Notation 4. Then we have

$$\begin{aligned} E_P &\le \max_{P\in\mathbf{X}_\Omega} (E_P + T_{\max}\phi_P) \\ &\le \max_{Q\in\mathbf{X}_{\partial\Omega}} (E_Q + T_{\max}\phi_Q) = T_{\max} \max_{Q\in\mathbf{X}_{\partial\Omega}} (\phi_Q), \end{aligned}$$

where the first step follows from $T_{\max}\phi_P \ge 0$, the second from Lemma 7.55, and the third from $(*)$.

Repeat the above arguments with $\psi_P = -E_P + T_{\max}\phi_P$ and we have

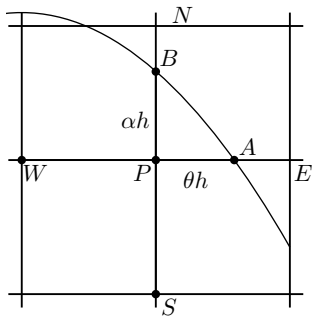
$$-E_P \le T_{\max} \max_{Q\in\mathbf{X}_{\partial\Omega}} (\phi_Q). \qquad \square$$

7.6.4 Convergence on irregular domains

Example 7.58 (An FD method for Poisson's equation in 2D irregular domains)**.** Consider the BVP

$$-\frac{\partial^2}{\partial x^2}u(x,y) - \frac{\partial^2}{\partial y^2}u(x,y) = f(x,y) \tag{7.85}$$

in a 2D irregular domain Ω with Dirichlet conditions.



An equation-discretization point is said to be *regular* if the standard 5-point stencil is applicable; otherwise it is *irregular*. For an irregular point, we modify the FD discretization in Example 7.38 to incorporate the info of local geometry and Dirichlet conditions. For example, in the above plot, the discrete operator becomes

$$L_hU_P := \frac{(1+\theta)U_P - U_A - \theta U_W}{\frac{1}{2}\theta(1+\theta)h^2} + \frac{(1+\alpha)U_P - U_B - \alpha U_S}{\frac{1}{2}\alpha(1+\alpha)h^2}. \tag{7.86}$$

In the resulting linear system $L_hU_P - f_P = 0$, the form of L_hU_P as in (7.86) is different from that in (7.56) at a regular equation-discretization point. Consequently, a global analysis of this linear system is difficult.

Exercise 7.59. Show that, in Example 7.58, the LTE at an irregular equation-discretization point is $O(h)$ while the LTE at a regular equation-discretization point is $O(h^2)$.

Theorem 7.60. Suppose that, in the notation of Theorem 7.57, the set $\mathbf{X}_\Omega$ of equation-discretization points can be partitioned as

$$\mathbf{X}_\Omega = \mathbf{X}_1 \cup \mathbf{X}_2, \quad \mathbf{X}_1 \cap \mathbf{X}_2 = \emptyset,$$

the nonnegative function $\phi : \mathbf{X} \to \mathbb{R}$ satisfies

$$\forall P \in \mathbf{X}_1, \quad L_h\phi_P \le -C_1 < 0; \tag{7.87a}$$
$$\forall P \in \mathbf{X}_2, \quad L_h\phi_P \le -C_2 < 0, \tag{7.87b}$$

and the LTE of (7.77) satisfy

$$\forall P \in \mathbf{X}_1, \quad |T_P| < T_1; \tag{7.88a}$$
$$\forall P \in \mathbf{X}_2, \quad |T_P| < T_2. \tag{7.88b}$$

Then the solution error $E_P := U_P - u(P)$ of the FD method (7.77) is bounded by

$$\forall P \in \mathbf{X}, \quad |E_P| \le \left(\max_{Q\in\mathbf{X}_{\partial\Omega}} \phi(Q)\right) \max_{P\in\mathbf{X}_\Omega} \left\{\frac{T_1}{C_1}, \frac{T_2}{C_2}\right\}.$$

Exercise 7.61. Prove Theorem 7.60 by choosing a function ψ to which Lemma 7.55 applies.

Theorem 7.62. The FD method in Example 7.58 is second-order convergent in the max-norm.

Proof. Define a comparison function ϕ as

$$\phi(x,y) := \begin{cases} F_1\left[(x-p)^2+(y-q)^2\right] & \text{if } (x,y)\in \mathbf{X}_\Omega; \\ F_1\left[(x-p)^2+(y-q)^2\right]+F_2 & \text{if } (x,y)\in \mathbf{X}_{\partial\Omega}, \end{cases}$$

where (p,q) is the geometric center of Ω and $F_1, F_2 > 0$ are constants to be chosen later. Both regular points and irregular points belong to $\mathbf{X}_\Omega$. Their difference is that, for a regular point Q, we have

$$L_h\phi_Q = -4F_1$$

while for an irregular point P shown in Example 7.58 the coefficient of U_A is

$$-\frac{2}{\theta(1+\theta)h^2} < -\frac{1}{h^2}$$

because $\theta \in (0,1)$. Similarly, the coefficient of U_B satisfies $-\frac{2}{\alpha(1+\alpha)h^2} < -\frac{1}{h^2}$. Therefore,

$$L_h\phi_P < -4F_1 - \frac{2}{h^2}F_2 < -\frac{2}{h^2}F_2. \tag{7.89}$$

By Exercise 7.59, we write the maximum LTEs on regular and irregular points as $T_1 = K_1h^2$ and $T_2 = K_2h$, respectively. Then Theorem 7.60 implies

$$|E_P| \le (F_1R^2+F_2)\max\left\{\frac{K_1h^2}{4F_1}, \frac{K_2h^3}{2F_2}\right\}, \tag{7.90}$$

where R is the maximum distance of a point in Ω to the geometric center of Ω. The RHS of the above equation is minimized when we choose $\frac{F_1}{F_2} = \frac{K_1}{2K_2h}$ so that the two terms in max{} equal. It follows that

$$\forall P\in\mathbf{X},\quad |E_P| \le \frac{1}{4}K_1R^2h^2 + \frac{1}{2}K_2h^3. \qquad \square$$

7.7 Problems

7.7.1 Programming Assignments

Write a C++ package to solve the two-dimensional Poisson equation in Definition 7.4 by the FD methods in Examples 7.38 and 7.58.

I. Your package must give the user the following options:

   (a) the problem domain: either $\Omega = [0,1]^2$ or $\Omega\setminus\mathbf{D}$ where $\mathbf{D}$ is a circular disk, of which the radius and the center, specified by the user, must keep $\Omega\setminus\mathbf{D}$ connected and must make $\Omega\setminus\mathbf{D}$ cover at least four grid points (you must check the validity of the input parameters);

   (b) boundary conditions: Dirichlet, Neumann, or mixed (partly Dirichlet and partly Neumann).

   Use a direct method such as the LU factorization to solve the linear system, with your favorite implementation of a BLAS or LAPACK.

II. For the function

$$u(x,y) = \exp(y+\sin(x)), \tag{7.91}$$

   derive the corresponding $f(x,y)$ and the boundary conditions. Test your solver for all combinations of (a,b) in I on grids with $n = 8, 16, 32, 64$ along each dimension, report the 1-, 2-, and ∞-norms of the errors and the corresponding convergence rates on the four grids. You should also design at least two of your own test functions and carry out the same process.

III. The user-specified parameters must be clearly listed in an input file with a key-value syntax; I recommend `json` (`https://www.json.org/json-en.html`), but you may use or write your own parser. The main program is supposed to read the input file, create objects from the input, call your BVP solver, generate test results, and report errors and convergence rates.

IV. The generation of pictures on solution and errors can be performed outside your program using another tool of your choice.

V. Write a GNU `makefile` under your root directory so that the command "`make run`" would trigger the compilation of your source code, the production of the executable, and the running of your tests.

VI. Write a report to summarize the main points of your numerical experiments, which should be designed to verify the analytic results in the notes. To some extent, your grade depends on the number of key conclusions in the notes confirmed by your numerical experiments.