Chapter 2
Finite Difference Methods for Elliptic Equations

Remark 2.1. Model problem. The model problem in this chapter is the Poisson equation with Dirichlet boundary conditions

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $\Omega \subset \mathbb{R}^2$. This chapter follows in wide parts Samarskij (1984). □

2.1 Basics on Finite Differences

Remark 2.2. Grid. This section considers the one-dimensional situation. Consider the interval $[0, 1]$ that is decomposed by an equidistant grid

$$\begin{aligned} x_i &= ih, \quad i = 0, \dots, n, \quad h = 1/n, \text{ – nodes,} \\ \omega_h &= \{x_i \ : \ i = 0, \dots, n\} \text{ – grid.} \end{aligned}$$

□

Definition 2.3. Grid function. A vector $\underline{u}_h = (u_0, \dots, u_n)^T \in \mathbb{R}^{n+1}$ that assigns every grid point a function value is called grid function. □

Definition 2.4. Finite differences. Let $v(x)$ be a sufficiently smooth function and denote by $v_i = v(x_i)$, where $\{x_i\}$ are the nodes of the grid. The following quotients are called

$$\begin{aligned} v_{x,i} &= \frac{v_{i+1} - v_i}{h} \text{ – forward difference,} \\ v_{\overline{x},i} &= \frac{v_i - v_{i-1}}{h} \text{ – backward difference,} \\ v_{\mathring{x},i} &= \frac{v_{i+1} - v_{i-1}}{2h} \text{ – central difference,} \\ v_{\overline{x}x,i} &= \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} \text{ – second order difference,} \end{aligned}$$

see Figure 2.1. □

Remark 2.5. Some properties of the finite differences. It is (*exercise*)

$$v_{\mathring{x},i} = \frac{1}{2}(v_{x,i} + v_{\overline{x},i}), \quad v_{\overline{x}x,i} = (v_{\overline{x},i})_{x,i}.$$

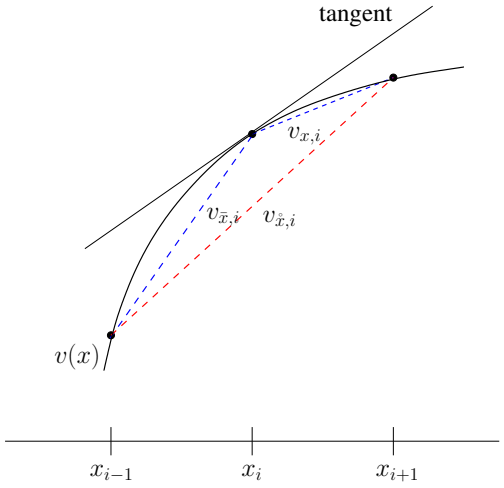


Fig. 2.1 Illustration of the finite differences.

Using the Taylor series expansion for $v(x)$ at the node x_i, one gets (*exercise*)

$$\begin{aligned} v_{x,i} &= v'(x_i) + \frac{1}{2}hv''(x_i) + \mathcal{O}\left(h^2\right), \\ v_{\bar{x},i} &= v'(x_i) - \frac{1}{2}hv''(x_i) + \mathcal{O}\left(h^2\right), \\ v_{\hat{x},i} &= v'(x_i) + \mathcal{O}\left(h^2\right), \\ v_{\bar{x}x,i} &= v''(x_i) + \mathcal{O}\left(h^2\right). \end{aligned}$$

□

Definition 2.6. Consistent difference operator. Let L be a differential operator. The difference operator $L_h : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is called consistent with L of order k if

$$\max_{0 \le i \le n} |(Lu)(x_i) - (L_h u_h)_i| = \|Lu - L_h u_h\|_{\infty,\omega_h} = \mathcal{O}\left(h^k\right)$$

for all sufficiently smooth functions $u(x)$. □

Example 2.7. Consistency orders. The order of consistency measures the quality of approximation of L by L_h.

The difference operators $v_{x,i}, v_{\bar{x},i}, v_{\hat{x},i}$ are consistent to $L = \frac{\mathrm{d}}{\mathrm{d}x}$ with order $1, 1$, and 2, respectively. The operator $v_{\bar{x}x,i}$ is consistent of second order to $L = \frac{\mathrm{d}^2}{\mathrm{d}x^2}$, see Remark 2.5. □

Example 2.8. Approximation of a more complicated differential operator by difference operators. Consider the differential operator

$$Lu = \frac{\mathrm{d}}{\mathrm{d}x}\left(k(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right),$$

where $k(x)$ is assumed to be continuously differentiable. Define the difference operator L_h as follows

$$(L_h u_h)_i = (a u_{\bar{x},i})_{x,i} = \frac{1}{h}\Big(a(x_{i+1})u_{\bar{x},i}(x_{i+1}) - a(x_i)u_{\bar{x},i}(x_i)\Big) = \frac{1}{h}\left(a_{i+1}\frac{u_{i+1}-u_i}{h} - a_i\frac{u_i - u_{i-1}}{h}\right), \quad (2.2)$$

where a is a grid function that has to be determined appropriately. One gets with the product rule

$$(Lu)_i = k'(x_i)(u')_i + k(x_i)(u'')_i$$

and with a Taylor series expansion for u_{i-1}, u_{i+1}, which is inserted in (2.2),

$$(L_h u_h)_i = \frac{a_{i+1} - a_i}{h}(u')_i + \frac{a_{i+1} + a_i}{2}(u'')_i + \frac{h(a_{i+1} - a_i)}{6}(u''')_i + \mathcal{O}\left(h^2\right).$$

Thus, the difference of the differential operator and the difference operator is

$$(Lu)_i - (L_h u_h)_i = \left(k'(x_i) - \frac{a_{i+1} - a_i}{h}\right)(u')_i + \left(k(x_i) - \frac{a_{i+1} + a_i}{2}\right)(u'')_i - \frac{h(a_{i+1} - a_i)}{6}(u''')_i + \mathcal{O}\left(h^2\right). \tag{2.3}$$

In order to define L_h so that it is consistent of second order to L, one has to satisfy the following two conditions

$$\frac{a_{i+1} - a_i}{h} = k'(x_i) + \mathcal{O}\left(h^2\right), \quad \frac{a_{i+1} + a_i}{2} = k(x_i) + \mathcal{O}\left(h^2\right).$$

From the first requirement, it follows that $a_{i+1} - a_i = \mathcal{O}(h)$. Hence, the third term in the consistency error equation (2.3) is already of order $\mathcal{O}\left(h^2\right)$. Possible choices for the grid function are (*exercise*)

$$a_i = \frac{k_i + k_{i-1}}{2}, \quad a_i = k\left(x_i - \frac{h}{2}\right), \quad a_i = (k_i k_{i-1})^{1/2}.$$

Notice that the 'natural' choice, $a_i = k_i$, leads only to first order consistency. (*exercise*) □

2.2 Finite Difference Approximation of the Laplacian in Two Dimensions

Remark 2.9. The five point stencil. The Laplacian in two dimensions is defined by

$$\Delta u(\boldsymbol{x}) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \partial_{xx} u + \partial_{yy} u = u_{xx} + u_{yy}, \quad \boldsymbol{x} = (x, y).$$

The simplest approximation uses for both second order derivatives the second order differences. One obtains the so-called five point stencil and the approximation

$$(\Delta u)_{ij} \approx (\Lambda u)_{ij} = u_{\overline{x}x,i} + u_{\overline{y}y,j} = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h_x^2} + \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h_y^2}, \tag{2.4}$$

see Figure 2.2. From the consistency order of the second order difference, it follows immediately that Λu approximates the Laplacian of order $\mathcal{O}(h_x^2 + h_y^2)$. □

Remark 2.10. The five point stencil on curvilinear boundaries. There is a difficulty if the five point stencil is used in domains with curvilinear boundaries. The approximation of the second derivative requires three function values in each coordinate direction

$$(x - h_x^-, y), (x, y), (x + h_x^+, y), \quad (x, y - h_y^-), (x, y), (x, y + h_y^+),$$

see Figure 2.3. A guideline of defining the approximation is that the five point stencil is recovered in the case $h_x^- = h_x^+$ and $h_y^- = h_y^+$. Consider just the x-direction. A possible approximation is

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{1}{\overline{h}_x}\left(\frac{u(x + h_x^+, y) - u(x, y)}{h_x^+} - \frac{u(x, y) - u(x - h_x^-, y)}{h_x^-}\right) \tag{2.5}$$

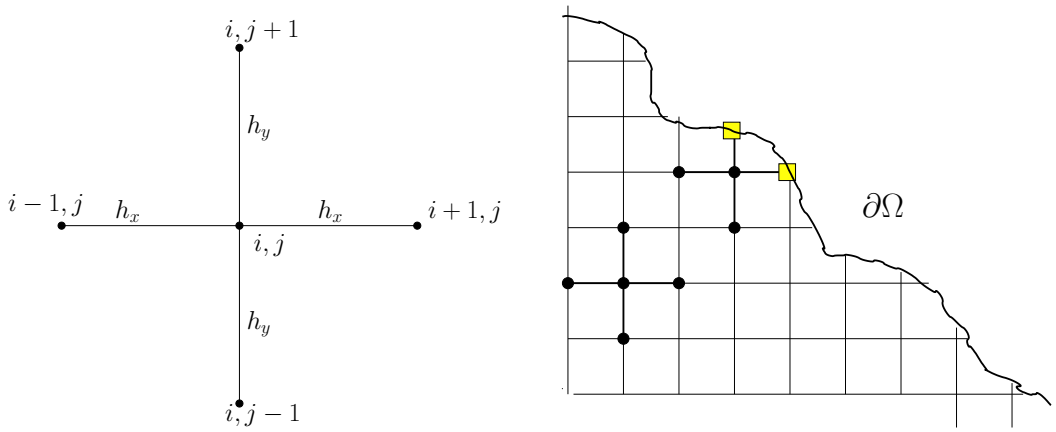


Fig. 2.2 Five point stencils.

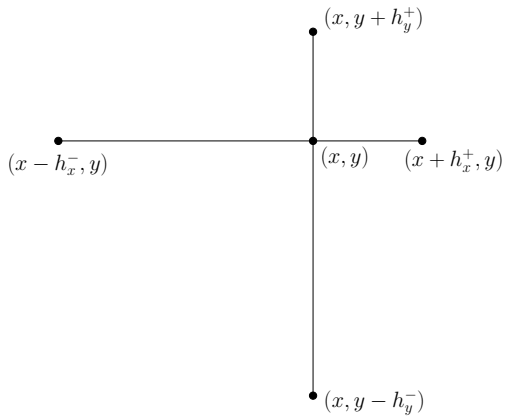


Fig. 2.3 Sketch to Remark 2.10.

with $\overline{h}_x = (h_x^+ + h_x^-)/2$. Using a Taylor series expansion, one finds that the error of this approximation is

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{\overline{h}_x}\left(\frac{u(x+h_x^+, y) - u(x, y)}{h_x^+} - \frac{u(x, y) - u(x-h_x^-, y)}{h_x^-}\right) = -\frac{1}{3}(h_x^+ - h_x^-)\frac{\partial^3 u}{\partial x^3} + \mathcal{O}\left(\overline{h}_x^2\right).$$

For $h_x^+ \neq h_x^-$, this approximation is of first order.

A different way consists to use

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{1}{\tilde{h}_x}\left(\frac{u(x+h_x^+, y) - u(x, y)}{h_x^+} - \frac{u(x, y) - u(x-h_x^-, y)}{h_x^-}\right)$$

with $\tilde{h}_x = \max\{h_x^+, h_x^-\}$. However, this approximation possesses only the order zero, i.e., there is actually no approximation.

Altogether, there is a loss of order of consistency at curvilinear boundaries. □

Example 2.11. The Dirichlet problem. Consider the Poisson equation that is equipped with Dirichlet boundary conditions (2.1). First, $\mathbb{R}^2$ is decomposed by a grid with rectangular mesh cells $x_i = ih_x, y_j = jh_y$, $h_x, h_y > 0$, $i, j \in \mathbb{Z}$. Denote by

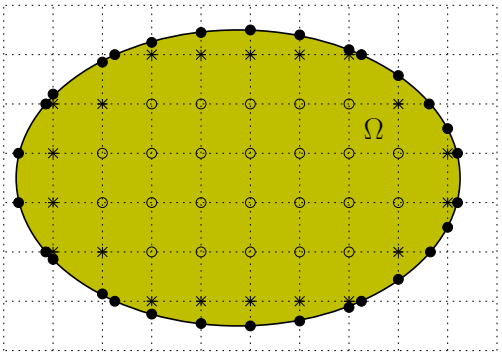


Fig. 2.4 Different types of nodes in the grid.

$w_h^\circ = \{\circ\}$	inner nodes, five point stencil does not contain any boundary node,
$w_h^* = \{*\}$	inner nodes that are close to the boundary, five point stencil contains boundary nodes,
$\gamma_h = \{\bullet\}$	boundary nodes,
$\omega_h = w_h^\circ \cup w_h^*$	inner nodes,
$\omega_h \cup \gamma_h$	grid,

see Figure 2.4.

The finite difference approximation of problem (2.1) that will be studied in the following consists to find a mesh function $u(\boldsymbol{x})$ such that

$$\begin{aligned} -\Lambda u(\boldsymbol{x}) &= \phi(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in w_h^\circ, \\ -\Lambda^* u(\boldsymbol{x}) &= \phi(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in w_h^*, \\ u(\boldsymbol{x}) &= g(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \gamma_h, \end{aligned} \tag{2.6}$$

where $\phi(\boldsymbol{x})$ is a grid function that approximates $f(\boldsymbol{x})$, and Λ^* is an approximation of the Laplacian for nodes that are close to the boundary, e.g., defined by (2.5). For the five point stencil,

$$\phi(\boldsymbol{x}) = f(\boldsymbol{x}) \tag{2.7}$$

is chosen. The discrete problem is a large sparse linear system of equations. The most important questions are:

- Which properties possesses the solution of (2.6)?
- Converges the solution of (2.6) to the solution of the Poisson problem and if yes, with which order in the norm $\|\cdot\|_{\infty,\omega_h}$?

□

2.3 The Discrete Maximum Principle for a Finite Difference Approximation

Remark 2.12. Contents of this section. Solutions of the Laplace problem, i.e., of (2.1) with $f(\boldsymbol{x}) = 0$, fulfill so-called maximum principles. This section shows that certain finite difference approximation of this operator, where the five point stencil of the Laplacian is a special case, satisfies a discrete analog of one of the maximum principles, under an assumption on the grid. The analysis proceeds along the classical lines, see Samarskij (1984) or (Samarskii, 2001, Chapter 4) □

Theorem 2.13. Maximum principles for harmonic functions. *Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$, be a bounded domain and $u \in C^2(\Omega) \cap C(\overline{\Omega})$ be harmonic in Ω, i.e., $u(\boldsymbol{x})$ solves the Laplace equation $-\Delta u = 0$ in Ω.*

- *Weak maximum principle. It holds*

$$\max_{\boldsymbol{x} \in \overline{\Omega}} u(\boldsymbol{x}) = \max_{\boldsymbol{x} \in \partial\Omega} u(\boldsymbol{x}).$$

  That means, $u(\boldsymbol{x})$ takes its maximal value at the boundary.
- *Strong maximum principle. If Ω is connected and if the maximum is taken in Ω (notice that Ω is open), i.e., $u(\boldsymbol{x}_0) = \max_{\boldsymbol{x} \in \overline{\Omega}} u(\boldsymbol{x})$ for a point $\boldsymbol{x}_0 \in \Omega$, then $u(\boldsymbol{x})$ is constant*

$$u(\boldsymbol{x}) = \max_{\boldsymbol{x} \in \overline{\Omega}} u(\boldsymbol{x}) = u(\boldsymbol{x}_0) \quad \forall\ \boldsymbol{x} \in \overline{\Omega}.$$

Proof. See the literature, e.g., (Evans, 2010, p. 27, Theorem 4) or the course on the theory of partial differential equations. ■

Remark 2.14. Interpretation of the maximum principle.

- The Laplace problem models the temperature distribution of a heated body without heat sources in Ω. Then, the weak maximum principle just states that the temperature in the interior of the body cannot be higher than the highest temperature at the boundary.
- There are maximum principles also for more complicated operators than the Laplacian, e.g., see Evans (2010).
- Since the solution of boundary value problems with partial differential equations will be only approximated by a discretization like a finite difference method, one has to expect that basic physical properties are satisfied by the numerical solution also only approximately. However, in applications, it is often very important that such properties are satisfied exactly. The development and study of so-called physically-consistent discretizations is an active field of research.

□

Remark 2.15. The difference equation. In this section, a difference equation of the form

$$a(\boldsymbol{x})u(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in S(\boldsymbol{x})} b(\boldsymbol{x}, \boldsymbol{y})u(\boldsymbol{y}) + F(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h \cup \gamma_h, \tag{2.8}$$

will be considered. In (2.8), for each node $\boldsymbol{x}$, the set $S(\boldsymbol{x})$ is the set of all nodes on which the sum has to be performed, but $\boldsymbol{x} \notin S(\boldsymbol{x})$. That means, $a(\boldsymbol{x})$ describes the contribution of the finite difference scheme of a node $\boldsymbol{x}$ to itself and $b(\boldsymbol{x}, \boldsymbol{y})$ describes the contributions from the neighbors. The algebraic formulation of (2.8) is a linear system of equations. Then, the diagonal entries are determined by $a(\boldsymbol{x})$ and the off-diagonal entries by $-b(\boldsymbol{x}, \boldsymbol{y})$, where the minus sign occurs because the term with $b(\boldsymbol{x}, \boldsymbol{y})$ is on the right-hand side of (2.8).

It will be assumed that the grid ω_h of inner nodes is connected, i.e., for all $\boldsymbol{x}_a, \boldsymbol{x}_e \in \omega_h$ exist $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \omega_h$ with $\boldsymbol{x}_1 \in S(\boldsymbol{x}_a), \boldsymbol{x}_2 \in S(\boldsymbol{x}_1), \ldots, \boldsymbol{x}_e \in S(\boldsymbol{x}_m)$. For instance, the situation depicted in Figure 2.5 is not allowed. The algebraic interpretation of this assumption, together with (2.9) below, is that the restriction of the system matrix to the inner nodes is an irreducible matrix.

It will be assumed that the coefficients $a(\boldsymbol{x})$ and $b(\boldsymbol{x}, \boldsymbol{y})$ satisfy the following conditions:

$$\begin{aligned} &a(\boldsymbol{x}) > 0,\ b(\boldsymbol{x}, \boldsymbol{y}) > 0,\ \forall\ \boldsymbol{x} \in \omega_h, \forall\ \boldsymbol{y} \in S(\boldsymbol{x}), \\ &a(\boldsymbol{x}) = 1,\ b(\boldsymbol{x}, \boldsymbol{y}) = 0\ \forall\ \boldsymbol{x} \in \gamma_h \text{ (Dirichlet boundary condition).} \end{aligned} \tag{2.9}$$

The values of the Dirichlet boundary condition are incorporated in (2.8) in the function $F(\boldsymbol{x})$. Thus, the linear system of equations will have the form

$$\begin{pmatrix} A^{\mathrm{i}} & A^{\mathrm{ib}} \\ 0 & I \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{u_g} \end{pmatrix} = \begin{pmatrix} \underline{\phi} \\ \underline{g} \end{pmatrix}, \tag{2.10}$$

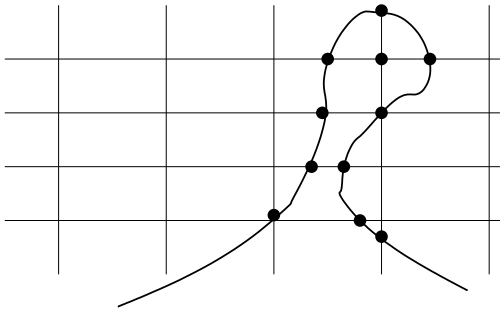


Fig. 2.5 Grid that is not allowed in Section 2.3.

where I is the identity matrix, $\underline{u}$ is the vector that corresponds to the inner nodes, $\underline{u_g}$ the vector for the boundary nodes, $\underline{\phi}$ the vector for the right-hand side in the inner nodes, and $\underline{g}$ the vector from the given boundary conditions. The matrix block A^{i} contains the connections among the inner nodes and the block A^{ib} the connections of the inner nodes close to the boundary to the boundary nodes. □

Example 2.16. Five point stencil for approximating the Laplacian. Inserting the approximation of the Laplacian with the five point stencil (2.4) for $\boldsymbol{x} = (x, y) \in \omega_h^\circ$ in scheme (2.8) gives

$$\frac{2(h_x^2 + h_y^2)}{h_x^2 h_y^2} u(x,y) = \left[\frac{1}{h_x^2} u(x + h_x, y) + \frac{1}{h_x^2} u(x - h_x, y) + \frac{1}{h_y^2} u(x, y + h_y) + \frac{1}{h_y^2} u(x, y - h_y) \right] + \phi(x,y).$$

It follows that

$$\begin{aligned} a(\boldsymbol{x}) &= \frac{2(h_x^2 + h_y^2)}{h_x^2 h_y^2} > 0, \\ b(\boldsymbol{x}, \boldsymbol{y}) &\in \{h_x^{-2}, h_y^{-2}\} > 0, \\ S(\boldsymbol{x}) &= \{(x - h_x, y), (x + h_x, y), (x, y - h_y), (x, y + h_y)\}. \end{aligned}$$

For inner nodes that are close to the boundary, only the one-dimensional case (2.5) will be considered for simplicity. Let $x + h_x^+ \in \gamma_h$, then it follows by inserting (2.5) in (2.8)

$$\frac{1}{\overline{h}_x} \left(\frac{1}{h_x^+} + \frac{1}{h_x^-} \right) u(x,y) = \frac{u(x - h_x^-, y)}{\overline{h}_x h_x^-} + \underbrace{\frac{u(x + h_x^+, y)}{\overline{h}_x h_x^+}}_{\text{on } \gamma_h \to A^{\mathrm{ib}}} + \phi(x), \tag{2.11}$$

such that

$$\begin{aligned} a(x) &= \frac{1}{\overline{h}_x} \left(\frac{1}{h_x^+} + \frac{1}{h_x^-} \right) > 0, \\ b(x) &\in \left\{ \frac{1}{\overline{h}_x h_x^-}, \frac{1}{\overline{h}_x h_x^+} \right\} > 0, \\ S(x) &= \{(x - h_x^-, y), (x + h_x^+, y)\}. \end{aligned}$$

Hence, the assumptions (2.9) on the coefficients are satisfied. □

Remark 2.17. Reformulation of the difference scheme. Scheme (2.8) can be reformulated in the form

$$d(\boldsymbol{x})u(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y})\big(u(\boldsymbol{y}) - u(\boldsymbol{x})\big) + F(\boldsymbol{x}) \tag{2.12}$$

with $d(\boldsymbol{x}) = a(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y})$. Algebraically, $d(\boldsymbol{x})$ is the sum of the matrix entries of the row that corresponds to the node $\boldsymbol{x}$. □

Example 2.18. Five point stencil for approximating the Laplacian. Using the five point stencil for approximating the Laplacian, the form (2.12) of the scheme is obtained with

$$d(\boldsymbol{x}) = \frac{2\left(h_x^2 + h_y^2\right)}{h_x^2 h_y^2} - \frac{2}{h_x^2} - \frac{2}{h_y^2} = 0 \tag{2.13}$$

for $\boldsymbol{x} \in \omega_h^\circ$. Thus, the corresponding row sums of the matrix are zero.

For nodes close to the boundary $\boldsymbol{x} \in \omega_h^*$, again only the one-dimensional situation as in Example 2.16 is considered. One obtains

$$d(x) = \frac{1}{\overline{h}_x}\left(\frac{1}{h_x^+} + \frac{1}{h_x^-}\right) - \frac{1}{\overline{h}_x h_x^-} - \frac{1}{\overline{h}_x h_x^+} = 0,$$

i.e., also for such nodes, the corresponding row sum vanishes.

The coefficients $a(\boldsymbol{x})$ and $b(\boldsymbol{x},\boldsymbol{y})$ are the weights of the finite difference stencil for approximating the Laplacian. A minimal condition for consistency is that this approximation vanishes for constant functions since the derivatives of constant functions vanish. The algebraic formulation of this consistency condition is just that all row sums of the rectangular matrix $(A^{\mathrm{i}}\ A^{\mathrm{ib}})$ vanish, since a constant function is represented by a constant vector. If the row sums vanish, then the multiplication of the matrix with a constant vector gives the zero vector. □

Lemma 2.19. Discrete maximum principle (DMP) for inner nodes. *Let $u(\boldsymbol{x}) \neq const$ on ω_h and $d(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \omega_h$. Then, it follows from*

$$L_h u(\boldsymbol{x}) := d(\boldsymbol{x})u(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y})\big(u(\boldsymbol{y}) - u(\boldsymbol{x})\big) \leq 0 \tag{2.14}$$

(or $L_h u(\boldsymbol{x}) \geq 0$, respectively) on ω_h that $u(\boldsymbol{x})$ does not possess a positive maximum (or negative minimum, respectively) on ω_h.

Proof. The proof is performed by contradiction. Let $L_h u(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \omega_h$ and assume that $u(\boldsymbol{x})$ has a positive maximum on ω_h at $\overline{\boldsymbol{x}}$, i.e., $u(\overline{\boldsymbol{x}}) = \max_{\boldsymbol{x}\in\omega_h} u(\boldsymbol{x}) > 0$.

For the node $\overline{\boldsymbol{x}}$, using (2.9), it holds that

$$L_h u(\overline{\boldsymbol{x}}) = \underbrace{d(\overline{\boldsymbol{x}})}_{\geq 0}\underbrace{u(\overline{\boldsymbol{x}})}_{>0} - \sum_{\boldsymbol{y}\in S(\overline{\boldsymbol{x}})} \underbrace{b(\overline{\boldsymbol{x}},\boldsymbol{y})}_{>0}\ \underbrace{\big(u(\boldsymbol{y}) - u(\overline{\boldsymbol{x}})\big)}_{\leq 0 \text{ by definition of } \overline{\boldsymbol{x}}} \geq d(\overline{\boldsymbol{x}})u(\overline{\boldsymbol{x}}) \geq 0. \tag{2.15}$$

Taking the assumptions of the lemma into account, it follows that $L_h u(\overline{\boldsymbol{x}}) = 0$ and, in particular, that all terms of $L_h u(\overline{\boldsymbol{x}})$ have to vanish. For the first term, it follows that $d(\overline{\boldsymbol{x}}) = 0$. For the terms in the sum to vanish, it must hold

$$u(\boldsymbol{y}) = u(\overline{\boldsymbol{x}}) \quad \forall\ \boldsymbol{y} \in S(\overline{\boldsymbol{x}}). \tag{2.16}$$

From the assumption $u(\boldsymbol{x}) \neq const$, it follows that there exists a node $\hat{\boldsymbol{x}} \in \omega_h$ with $u(\overline{\boldsymbol{x}}) > u(\hat{\boldsymbol{x}})$. Because the grid is connected, there is a path $\overline{\boldsymbol{x}}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, \hat{\boldsymbol{x}}$ in ω_h such that, using (2.16) for all nodes of this path,

$$\begin{aligned}
&\boldsymbol{x}_1 \in S(\overline{\boldsymbol{x}}),\ \ u(\boldsymbol{x}_1) = u(\overline{\boldsymbol{x}}),\\
&\boldsymbol{x}_2 \in S(\boldsymbol{x}_1),\ u(\boldsymbol{x}_2) = u(\boldsymbol{x}_1) = u(\overline{\boldsymbol{x}}),\\
&\cdots\\
&\hat{\boldsymbol{x}} \in S(\boldsymbol{x}_m),\ u(\boldsymbol{x}_m) = u(\boldsymbol{x}_{m-1}) = \ldots = u(\overline{\boldsymbol{x}}) > u(\hat{\boldsymbol{x}}).
\end{aligned}$$

The last inequality is a contradiction to (2.16) for $\boldsymbol{x}_m$. ■

Remark 2.20. On L_h. Notice that L_h is defined for the inner nodes, i.e., this operator corresponds to the rectangular matrix $(A^{\mathrm{i}}\ A^{\mathrm{ib}})$ from (2.10). □

Corollary 2.21. DMP for the finite difference boundary value problem. *Let $u(\boldsymbol{x}) \leq 0$ for $\boldsymbol{x} \in \gamma_h$ and $L_h u(\boldsymbol{x}) \leq 0$ (or $u(\boldsymbol{x}) \geq 0$ for $\boldsymbol{x} \in \gamma_h$ and $L_h u(\boldsymbol{x}) \geq 0$, respectively) on ω_h. Assume that there is at least one inner node close to the boundary $\boldsymbol{x}^*$ and one node $\boldsymbol{x}_\gamma$ on the boundary with $b(\boldsymbol{x}^*, \boldsymbol{x}_\gamma) > 0$, i.e., the matrix block A_2 in (2.10) is not the zero matrix. Then, the grid function $u(\boldsymbol{x})$ is non-positive (or non-negative, respectively) for all $\boldsymbol{x} \in \omega_h \cup \gamma_h$.*

Proof. Let $L_h u(\boldsymbol{x}) \leq 0$ on ω_h. Assume that there is a node $\overline{\boldsymbol{x}} \in \omega_h$ with $u(\overline{\boldsymbol{x}}) > 0$. Then, the grid function has either a positive maximum on ω_h and it is not constant, which is a contradiction to the DMP for the inner nodes, Lemma 2.19, or $u(\boldsymbol{x})$ has to be constant, i.e., $u(\boldsymbol{x}) = u(\overline{\boldsymbol{x}}) > 0$ for all $\boldsymbol{x} \in \omega_h$. For the second case, consider the boundary-connected inner node $\boldsymbol{x}^* \in \omega_h^*$. Using the same calculations as in (2.15) and taking into account that the values of u at the boundary are non-positive, one obtains

$$L_h u(\boldsymbol{x}^*) = \underbrace{d(\boldsymbol{x}^*)}_{\geq 0}\underbrace{u(\boldsymbol{x}^*)}_{>0} - \sum_{\boldsymbol{y}\in S(\boldsymbol{x}^*),\boldsymbol{y}\notin\gamma_h} \underbrace{b(\boldsymbol{x}^*,\boldsymbol{y})}_{>0}\underbrace{\big(u(\boldsymbol{y}) - u(\boldsymbol{x}^*)\big)}_{=0} - \sum_{\boldsymbol{y}\in S(\boldsymbol{x}^*),\boldsymbol{y}\in\gamma_h} \underbrace{b(\boldsymbol{x}^*,\boldsymbol{y})}_{>0}\underbrace{\big(u(\boldsymbol{y}) - u(\boldsymbol{x}^*)\big)}_{<0} > 0. \tag{2.17}$$

In the last sum, there is at least one term since $\boldsymbol{x}_\gamma \in S(\boldsymbol{x}^*)$. Altogether, (2.17) is a contradiction to the assumption on L_h. ■

Corollary 2.22. Unique solution of the discrete Laplace problem with homogeneous right-hand side and homogeneous Dirichlet boundary conditions. *Under the assumptions of Corollary 2.21, the discrete Laplace problem $L_h u(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \omega_h$ and $u(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \gamma_h$ possesses only the trivial solution $u(\boldsymbol{x}) = 0$.*

Proof. The statement of the corollary follows by applying Corollary 2.21 both for $L_h u(\boldsymbol{x}) \leq 0$ and $L_h u(\boldsymbol{x}) \geq 0$. ■

Theorem 2.23. Existence and uniqueness of a solution of the finite difference problem (2.6)**.** *Under the assumptions of Corollary 2.22, the finite difference problem* (2.6) *possesses a unique solution.*

Proof. Corollary 2.22 shows that the homogeneous linear system of equations (2.10) has a unique solution. Hence, the system matrix is invertible and it follows that (2.10) is uniquely solvable for all right-hand sides, where (2.10) is just the matrix-vector representation of (2.6). ■

Corollary 2.24. Comparison lemma. *Let the assumptions of Corollary 2.21 be satisfied and let*

$$\begin{aligned} L_h u(\boldsymbol{x}) &= f(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \omega_h; \quad u(\boldsymbol{x}) = g(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \gamma_h, \\ L_h \overline{u}(\boldsymbol{x}) &= \overline{f}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \omega_h; \quad \overline{u}(\boldsymbol{x}) = \overline{g}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \gamma_h, \end{aligned}$$

with $|f(\boldsymbol{x})| \leq \overline{f}(\boldsymbol{x})$, $\boldsymbol{x} \in \omega_h$, and $|g(\boldsymbol{x})| \leq \overline{g}(\boldsymbol{x})$, $\boldsymbol{x} \in \gamma_h$. Then, it is $|u(\boldsymbol{x})| \leq \overline{u}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \omega_h \cup \gamma_h$. The function $\overline{u}(\boldsymbol{x})$ is called majorizing function.

Proof. Exercise. ■

Remark 2.25. Remainder of this section. The remaining corollaries presented in this section will be applied in the stability proof in Section 2.4. In this proof, the homogeneous problem (right-hand side vanishes) and the problem with homogeneous Dirichlet boundary conditions will be analyzed separately. □

Corollary 2.26. Homogeneous problem. *For the solution of the problem*

$$\begin{aligned} L_h u(\boldsymbol{x}) &= 0, \quad \boldsymbol{x} \in \omega_h, \\ u(\boldsymbol{x}) &= g(\boldsymbol{x}), \ \boldsymbol{x} \in \gamma_h, \end{aligned}$$

with $d(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \omega_h$, it holds that

$$\|u\|_{l^\infty(\omega_h \cup \gamma_h)} \leq \|g\|_{l^\infty(\gamma_h)}.$$

Proof. Consider the problem

$$\begin{aligned} L_h\overline{u}(\boldsymbol{x}) &= 0, && \boldsymbol{x}\in\omega_h,\\ \overline{u}(\boldsymbol{x}) &= \overline{g}(\boldsymbol{x}) = const = \|g\|_{l^\infty(\gamma_h)}, && \boldsymbol{x}\in\gamma_h. \end{aligned}$$

Since the row sums for $\boldsymbol{x}\in\omega_h$ vanish, $\overline{u}(\boldsymbol{x}) = \|g\|_{l^\infty(\gamma_h)} = const$ is a solution of this problem.[1] By Corollary 2.22, this solution is unique.

Now, the application of Corollary 2.24 gives $\overline{u}(\boldsymbol{x}) \geq |u(\boldsymbol{x})|$ for all $\boldsymbol{x}\in\omega_h\cup\gamma_h$, so that

$$\|u\|_{l^\infty(\omega_h\cup\gamma_h)} \leq \overline{u}(\boldsymbol{x}) = \|g\|_{l^\infty(\gamma_h)},$$

which is the statement of the corollary. ∎

Corollary 2.27. Problem with homogeneous boundary condition and inhomogeneous right-hand side close to the boundary. *Consider*

$$\begin{aligned} L_h u(\boldsymbol{x}) &= f(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h,\\ u(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x}\in\gamma_h, \end{aligned}$$

with $f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}\in\omega_h^\circ$. Define[2]

$$\tilde{d}(\boldsymbol{x}) = a(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\notin\gamma_h} b(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{x}) + \sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\in\gamma_h} b(\boldsymbol{x},\boldsymbol{y}) \quad \boldsymbol{x}\in\omega_h.$$

With respect to the finite difference scheme, it will be assumed that $\tilde{d}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}\in\omega_h^\circ$, and $\tilde{d}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}\in\omega_h^*$. Then, the following estimate is valid

$$\|u\|_{l^\infty(\omega_h\cup\gamma_h)} \leq \left\|D^+ f\right\|_{l^\infty(\omega_h)}$$

with $D^+ = diag(0,\tilde{d}(\boldsymbol{x})^{-1})$. The zero entries appear for $\boldsymbol{x}\in\omega_h^\circ$ and the entries $\tilde{d}(\boldsymbol{x})^{-1}$ for $\boldsymbol{x}\in\omega_h^*$.

Proof. Let $\overline{f}(\boldsymbol{x}) = |f(\boldsymbol{x})|$, $\boldsymbol{x}\in\omega_h$, and $\overline{g}(\boldsymbol{x}) = 0, \boldsymbol{x}\in\gamma_h$. The solution of the corresponding finite difference problem $\overline{u}(\boldsymbol{x})$ is non-negative, $\overline{u}(\boldsymbol{x})\geq 0$ for all $\boldsymbol{x}\in\omega_h\cup\gamma_h$, see the DMP for the boundary value problem, Corollary 2.21. Define $\overline{\boldsymbol{x}}$ by

$$\overline{u}(\overline{\boldsymbol{x}}) = \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)}.$$

One can choose $\overline{\boldsymbol{x}}\in\omega_h^*$, because if $\overline{\boldsymbol{x}}\in\omega_h^\circ$, then it holds that

$$\underbrace{d(\overline{\boldsymbol{x}})}_{=0}\overline{u}(\overline{\boldsymbol{x}}) - \sum_{\boldsymbol{y}\in S(\overline{\boldsymbol{x}})} \underbrace{b(\overline{\boldsymbol{x}},\boldsymbol{y})}_{>0}\underbrace{\big(\overline{u}(\boldsymbol{y}) - \overline{u}(\overline{\boldsymbol{x}})\big)}_{\leq 0} = \overline{f}(\overline{\boldsymbol{x}}) = 0,$$

i.e., $\overline{u}(\overline{\boldsymbol{x}}) = \overline{u}(\boldsymbol{y})$ for all $\boldsymbol{y}\in S(\overline{\boldsymbol{x}})$. Let $\hat{\boldsymbol{x}}\in\omega_h^*$ and $\overline{\boldsymbol{x}},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\hat{\boldsymbol{x}}$ be a connection with $\boldsymbol{x}_i\notin\omega_h^*$, $i=1,\ldots,m$. For $\boldsymbol{x}_m$, it holds analogously that

$$\overline{u}(\boldsymbol{x}_m) = \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)} = \overline{u}(\boldsymbol{y})\ \forall\ \boldsymbol{y}\in S(\boldsymbol{x}_m).$$

Hence, it follows in particular that $\overline{u}(\hat{\boldsymbol{x}}) = \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)}$ so that one can choose $\overline{\boldsymbol{x}} = \hat{\boldsymbol{x}}$.

Using the definition of $\tilde{d}(\hat{\boldsymbol{x}})$ and the homogeneous values at the boundary yields

$$\begin{aligned} d(\hat{\boldsymbol{x}})\overline{u}(\hat{\boldsymbol{x}}) - \sum_{\boldsymbol{y}\in S(\hat{\boldsymbol{x}})} b(\hat{\boldsymbol{x}},\boldsymbol{y})\big(\overline{u}(\boldsymbol{y}) - \overline{u}(\hat{\boldsymbol{x}})\big) &= \overline{f}(\hat{\boldsymbol{x}}) \quad \Longleftrightarrow\\ d(\hat{\boldsymbol{x}})\overline{u}(\hat{\boldsymbol{x}}) + \sum_{\boldsymbol{y}\in S(\hat{\boldsymbol{x}}),\boldsymbol{y}\in\gamma_h} b(\hat{\boldsymbol{x}},\boldsymbol{y})\overline{u}(\hat{\boldsymbol{x}}) - \sum_{\boldsymbol{y}\in S(\hat{\boldsymbol{x}}),\boldsymbol{y}\notin\gamma_h} b(\hat{\boldsymbol{x}},\boldsymbol{y})\big(\overline{u}(\boldsymbol{y}) - \overline{u}(\hat{\boldsymbol{x}})\big) - \sum_{\boldsymbol{y}\in S(\hat{\boldsymbol{x}}),\boldsymbol{y}\in\gamma_h} b(\hat{\boldsymbol{x}},\boldsymbol{y})\overline{u}(\boldsymbol{y}) &= \overline{f}(\hat{\boldsymbol{x}}) \quad \Longleftrightarrow\\ \underbrace{\tilde{d}(\hat{\boldsymbol{x}})}_{>0}\underbrace{\overline{u}(\hat{\boldsymbol{x}})}_{=\|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)}} - \sum_{\boldsymbol{y}\in S(\hat{\boldsymbol{x}}),\boldsymbol{y}\notin\gamma_h} \underbrace{b(\hat{\boldsymbol{x}},\boldsymbol{y})}_{>0}\underbrace{\big(\overline{u}(\boldsymbol{y}) - \overline{u}(\hat{\boldsymbol{x}})\big)}_{\leq 0} &= \overline{f}(\hat{\boldsymbol{x}}). \end{aligned}$$

[1] For the Poisson problem, the corresponding continuous problem is $-\Delta u = 0$ in Ω, $u = const = \|g\|_{l^\infty(\gamma_h)}$ on $\partial\Omega$. It is clear that $u = \|g\|_{l^\infty(\gamma_h)}$ is the solution of this problem. It is shown that the discrete analog holds, too.

[2] The value of $\tilde{d}(\boldsymbol{x})$ is just the row sum of the matrix block A^{i} from (2.10).

It follows, using also Corollary 2.24, that

$$\|u\|_{l^\infty(\omega_h\cup\gamma_h)} \le \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)} \le \frac{\overline{f}(\hat{\boldsymbol{x}})}{\tilde{d}(\hat{\boldsymbol{x}})} \le \max_{\boldsymbol{x}\in\omega_h^*} \frac{\overline{f}(\boldsymbol{x})}{\tilde{d}(\boldsymbol{x})} \le \left\|D^+ f\right\|_{l^\infty(\omega_h)}.$$

■

2.4 Stability and Convergence of the Finite Difference Approximation of the Poisson Problem with Dirichlet Boundary Conditions

Remark 2.28. Decomposition of the solution. A short form to write (2.6) with $\phi(\boldsymbol{x}) = f(\boldsymbol{x})$ is

$$L_h u(\boldsymbol{x}) = f(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h, \quad u(\boldsymbol{x}) = g(\boldsymbol{x}),\ \boldsymbol{x}\in\gamma_h.$$

The solution of (2.6) can be decomposed into

$$u(\boldsymbol{x}) = u_1(\boldsymbol{x}) + u_2(\boldsymbol{x}),$$

with

$$\begin{aligned} &L_h u_1(\boldsymbol{x}) = f(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h, \quad u_1(\boldsymbol{x}) = 0,\ \boldsymbol{x}\in\gamma_h \text{ (homogeneous bdry. cond.)},\\ &L_h u_2(\boldsymbol{x}) = 0,\ \boldsymbol{x}\in\omega_h, \quad u_2(\boldsymbol{x}) = g(\boldsymbol{x}),\ \boldsymbol{x}\in\gamma_h \text{ (homogeneous rhs)}. \end{aligned}$$

□

Stability with Respect to the Boundary Condition

Remark 2.29. Stability with respect to the boundary condition. From Corollary 2.26, it follows that

$$\|u_2\|_{l^\infty(\omega_h)} \le \|g\|_{l^\infty(\gamma_h)}. \tag{2.18}$$

□

Stability with Respect to the Right-Hand Side

Remark 2.30. Decomposition of the right-hand side. The right-hand side will be decomposed into

$$f(\boldsymbol{x}) = f^\circ(\boldsymbol{x}) + f^*(\boldsymbol{x})$$

with

$$f^\circ(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}), \boldsymbol{x}\in\omega_h^\circ, \\ 0, \quad \boldsymbol{x}\in\omega_h^*, \end{cases} \qquad f^*(\boldsymbol{x}) = f(\boldsymbol{x}) - f^\circ(\boldsymbol{x}).$$

Since the considered finite difference scheme is linear, also the function $u_1(\boldsymbol{x})$ can be decomposed into

$$u_1(\boldsymbol{x}) = u_1^\circ(\boldsymbol{x}) + u_1^*(\boldsymbol{x})$$

with

$$L_h u_1^\circ(\boldsymbol{x}) = f^\circ(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h, \quad u_1^\circ(\boldsymbol{x}) = 0,\ \boldsymbol{x}\in\gamma_h,$$

$$L_h u_1^*(\boldsymbol{x}) = f^*(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \quad u_1^*(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \gamma_h.$$

From here, only the finite difference discretization of the Poisson problem with the five point stencil is considered. □

Remark 2.31. Estimate for the inner nodes. Let $B((0,0), R)$ be a circle with center $(0,0)$ and radius R, which is chosen so that $R \geq \|\boldsymbol{x}\|_2$ for all $\boldsymbol{x} \in \overline{\Omega}$. Consider the function

$$\overline{u}(\boldsymbol{x}) = \alpha\left(R^2 - x^2 - y^2\right) \quad \text{with } \alpha > 0,$$

that takes for $(x, y) \in \Omega$ only positive values. Applying the definition of the five point stencil, it follows that

$$\begin{aligned}
\Lambda\overline{u}(\boldsymbol{x}) &= -\alpha\Lambda(x^2 + y^2 - R^2) \\
&= -\alpha\left(\frac{(x+h_x)^2 - 2x^2 + (x-h_x)^2}{h_x^2} + \frac{(y+h_y)^2 - 2y^2 + (y-h_y)^2}{h_y^2}\right) \\
&= -4\alpha =: -\overline{f}(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h^\circ,
\end{aligned}$$

and

$$\begin{aligned}
\Lambda^*\overline{u}(\boldsymbol{x}) &= -\alpha\left[\frac{1}{\overline{h}_x}\left(\frac{(x+h_x^+)^2 - x^2}{h_x^+} - \frac{x^2 - (x-h_x^-)^2}{h_x^-}\right) + \frac{1}{\overline{h}_y}\left(\frac{(y+h_y^+)^2 - y^2}{h_y^+} - \frac{y^2 - (y-h_y^-)^2}{h_y^-}\right)\right] \\
&= -\alpha\left(\frac{h_x^+ + h_x^-}{\overline{h}_x} + \frac{h_y^+ + h_y^-}{\overline{h}_y}\right) =: -\overline{f}(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h^*.
\end{aligned}$$

Hence, $\overline{u}(\boldsymbol{x})$ is the solution of the finite difference problem

$$\begin{aligned}
L_h\overline{u}(\boldsymbol{x}) &= \overline{f}(\boldsymbol{x}), && \boldsymbol{x} \in \omega_h, \\
\overline{u}(\boldsymbol{x}) &= \alpha\left(R^2 - x^2 - y^2\right) \geq 0, && \boldsymbol{x} \in \gamma_h.
\end{aligned}$$

By construction, it is $\overline{u}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \gamma_h$. Choosing $\alpha = \frac{1}{4}\left\|f^\circ\right\|_{l^\infty(\omega_h)}$, one obtains

$$\begin{aligned}
&\overline{f}(\boldsymbol{x}) = 4\alpha = \left\|f^\circ\right\|_{l^\infty(\omega_h)} \geq \left|f^\circ(\boldsymbol{x})\right|,\ \boldsymbol{x} \in \omega_h^\circ, \\
&\overline{f}(\boldsymbol{x}) \geq 0 = \left|f^\circ(\boldsymbol{x})\right|,\ \boldsymbol{x} \in \omega_h^*.
\end{aligned}$$

Now, Corollary 2.24 (Comparison Lemma) can be applied, which leads to

$$\left\|u_1^\circ\right\|_{l^\infty(\omega_h)} \leq \|\overline{u}\|_{l^\infty(\omega_h)} \leq \alpha R^2 = \frac{R^2}{4}\left\|f^\circ\right\|_{l^\infty(\omega_h)}. \tag{2.19}$$

One gets the last 'lower or equal' estimate because $(0,0)$ does not need to belong to Ω or ω_h. □

Remark 2.32. Estimate for the nodes that are close to the boundary. Corollary 2.27 can be applied to estimate $u_1^*(\boldsymbol{x})$. For $\boldsymbol{x} \in \omega_h^*$, one has

$$\tilde{d}(\boldsymbol{x}) = a(\boldsymbol{x}) - \sum_{\boldsymbol{y} \in S(\boldsymbol{x}), \boldsymbol{y} \notin \gamma_h} b(\boldsymbol{x}, \boldsymbol{y}).$$

Consider again for simplicity the one-dimensional case. With the approach from Example 2.18, one finds, using the definition of $\overline{h}_x$ and $h_x^- = h_x \geq h_x^+$ that

$$\tilde{d}(x) = \frac{1}{\overline{h}_x}\left(\frac{1}{h_x^+} + \frac{1}{h_x^-}\right) - \frac{1}{\overline{h}_x h_x^-} = \frac{1}{\overline{h}_x h_x^+} = \frac{2}{h_x h_x^+ + h_x^+ h_x^+} \geq \frac{2}{h_x h_x + h_x h_x} = \frac{1}{h_x h_x} > 0.$$

Hence, it is

$$\tilde{d}(\boldsymbol{x}) \geq \frac{1}{h^2}$$

with $h = \max\{h_x, h_y\}$. One obtains with Corollary 2.27 that

$$\left\|u_1^*\right\|_{l^\infty(\omega_h)} \leq \left\|D^+ f^*\right\|_{l^\infty(\omega_h)} \leq h^2 \left\|f^*\right\|_{l^\infty(\omega_h)}. \tag{2.20}$$

□

Lemma 2.33. Stability estimate. *The solution of the discrete Dirichlet problem* (2.6) *with* $\phi(\boldsymbol{x}) = f(\boldsymbol{x})$ *satisfies*

$$\|u\|_{l^\infty(\omega_h \cup \gamma_h)} \leq \|g\|_{l^\infty(\gamma_h)} + \frac{R^2}{4} \|f\|_{l^\infty(\omega_h^\circ)} + h^2 \|f\|_{l^\infty(\omega_h^*)} \tag{2.21}$$

with $R \geq \|\boldsymbol{x}\|_2$ *for all* $\boldsymbol{x} \in \Omega$ *and* $h = \max\{h_x, h_y\}$*, i.e., the solution* $u(\boldsymbol{x})$ *can be bounded in the norm* $\|\cdot\|_{l^\infty(\omega_h \cup \gamma_h)}$ *by the data of the problem.*

Proof. The statement of the lemma is obtained by combining the estimates (2.18), (2.19), and (2.20). ■

Convergence

Theorem 2.34. Convergence. *Let* $u(\boldsymbol{x})$ *be the solution of the Poisson problem* (2.1) *and* $u_h(\boldsymbol{x})$ *be the finite difference approximation given by the solution of* (2.6) *with* $\phi(\boldsymbol{x}) = f(\boldsymbol{x})$*. Then, it is*

$$\|u - u_h\|_{l^\infty(\omega_h \cup \gamma_h)} \leq Ch^2$$

with $h = \max\{h_x, h_y\}$.

Proof. The error in the node (x_i, y_j) is defined by $e_{ij} = u(x_i, y_j) - u_h(x_i, y_j)$. With the consistency relation $-\Lambda u(x_i, y_j) = -\Delta u(x_i, y_j) + \mathcal{O}\left(h^2\right)$, the Poisson equation (2.1) and the finite difference problem (2.6), one obtains for interior nodes

$$\begin{aligned} -\Lambda e(x_i, y_j) &= -\Lambda u(x_i, y_j) + \Lambda u_h(x_i, y_j) = -\Delta u(x_i, y_j) + \mathcal{O}\left(h^2\right) - f(x_i, y_i) \\ &= f(x_i, y_i) + \mathcal{O}\left(h^2\right) - f(x_i, y_i) = \mathcal{O}\left(h^2\right). \end{aligned}$$

Performing a similar calculation for the nodes close to the boundary leads to the following problem for the error

$$\begin{aligned} -\Lambda e(\boldsymbol{x}) &= \psi(\boldsymbol{x}),\ \boldsymbol{x} \in w_h^\circ,\ \psi(\boldsymbol{x}) = \mathcal{O}\left(h^2\right), \\ -\Lambda^* e(\boldsymbol{x}) &= \psi(\boldsymbol{x}),\ \boldsymbol{x} \in w_h^*,\ \psi(\boldsymbol{x}) = \mathcal{O}\left(h\right), \\ e(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \in \gamma_h, \end{aligned}$$

where $\psi(\boldsymbol{x})$ is the consistency error, see Section 2.2. Applying the stability estimate (2.21) to this problem, one obtains immediately

$$\|e\|_{l^\infty(\omega_h \cup \gamma_h)} \leq \frac{R^2}{4} \|\psi\|_{l^\infty(\omega_h^\circ)} + h^2 \|\psi\|_{l^\infty(\omega_h^*)} = \mathcal{O}\left(h^2\right).$$

■

2.5 An Efficient Solver for the Dirichlet Problem in the Rectangle

Remark 2.35. Contents of this section. This section considers the Poisson problem (2.1) in the special case $\Omega = (0, l_x) \times (0, l_y)$. In this case, a modification of the difference stencil in a neighborhood of the boundary

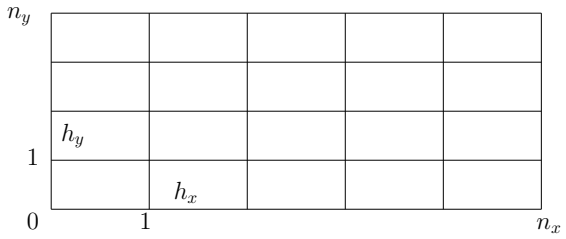


Fig. 2.6 Grid for the Dirichlet problem in the rectangular domain.

of the domain is not needed. The convergence of the finite difference approximation was already established in Theorem 2.34. Applying this approximation results in a large linear system of equations $A\boldsymbol{u} = \underline{f}$ which has to be solved. This section discusses some properties of the matrix A and it presents an approach for solving this system in the case of a rectangular domain in an almost optimal way.

A number of result obtained here will be needed also in Section 2.6. □

Remark 2.36. The considered problem and its approximation. The considered continuous problem consists in solving

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega = (0, l_x) \times (0, l_y), \\ u &= g \text{ on } \partial\Omega, \end{aligned}$$

and the corresponding discrete problem in solving

$$\begin{aligned} -\Lambda u(\boldsymbol{x}) &= f(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \\ u(\boldsymbol{x}) &= g(\boldsymbol{x}),\ \boldsymbol{x} \in \gamma_h, \end{aligned}$$

where the discrete Laplacian is of the form (for simplicity of notation, the subscript h is omitted)

$$\Lambda u = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h_x^2} + \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h_y^2} =: \Lambda_x u + \Lambda_y u, \tag{2.22}$$

with $h_x = l_x/n_x, h_y = l_y/n_y$, $i = 0, \ldots, n_x, j = 0, \ldots, n_y$, see Figure 2.6. □

Remark 2.37. The linear system of equations. The difference scheme (2.22) is equivalent to a linear system of equations $A\underline{u} = \underline{f}$.

For assembling the matrix and the right-hand side of the system, often a lexicographical enumeration of the nodes of the grid is used. The nodes are called enumerated lexicographically if the node (i_1, j_1) has a smaller number than the node (i_2, j_2), if for the corresponding coordinates, it is

$$y_1 < y_2 \ \text{ or } (y_1 = y_2) \wedge (x_1 < x_2).$$

Using this lexicographical enumeration of the nodes, one obtains for the inner nodes a system of the form

$$\begin{aligned} A &= \text{BlockTriDiag}(C, B, C) \in \mathbb{R}^{(n_x-1)(n_y-1)\times(n_x-1)(n_y-1)}, \\ B &= \text{TriDiag}\left(-\frac{1}{h_x^2}, \frac{2}{h_x^2} + \frac{2}{h_y^2}, -\frac{1}{h_x^2}\right) \in \mathbb{R}^{(n_x-1)\times(n_x-1)}, \\ C &= \text{Diag}\left(-\frac{1}{h_y^2}\right) \in \mathbb{R}^{(n_x-1)\times(n_x-1)}, \end{aligned}$$

$$\underline{f} = \begin{cases} f(\boldsymbol{x}), & \boldsymbol{x} \in \omega_h^\circ, \\ f(\boldsymbol{x}) + \dfrac{g(x \pm h_x, y)}{h_x^2}, & \boldsymbol{x} \in \omega_h^*, \text{ close to right or left boundary,} \\ f(\boldsymbol{x}) + \dfrac{g(x, y \pm h_y)}{h_y^2}, & \boldsymbol{x} \in \omega_h^*, \text{ close to upper or lower boundary,} \\ f(\boldsymbol{x}) + \dfrac{g(x \pm h_x, y)}{h_x^2} + \dfrac{g(x, y \pm h_y)}{h_y^2}, & \boldsymbol{x} \in \omega_h^*, \text{ corner of inner nodes.} \end{cases} \tag{2.23}$$

In this approach, the known Dirichlet boundary values are already substituted into the system and they appear in the right-hand side vector. The matrices B and C possess some modifications for nodes that have a neighbor on the boundary.

The linear system of equations has the following properties:

- high dimension: $N = (n_x - 1)(n_y - 1) \sim 10^3 \cdots 10^7$,
- sparse: per row and column of the matrix there are only 3, 4, or 5 non-zero entries,
- symmetric: hence, all eigenvalues are real,
- positive definite: all eigenvalues are positive. It holds that

$$\begin{aligned} \lambda_{\min} = \lambda_{(1,1)} &\sim \pi^2 \left(\frac{1}{l_x^2} + \frac{1}{l_y^2} \right) = \mathcal{O}(1), \\ \lambda_{\max} = \lambda_{(n_x-1, n_y-1)} &\sim \pi^2 \left(\frac{1}{h_x^2} + \frac{1}{h_y^2} \right) = \mathcal{O}\left(h^{-2}\right), \end{aligned} \tag{2.24}$$

  with $h = \max\{h_x, h_y\}$, see Remark 2.38 below.
- high condition number: For the spectral condition number of a symmetric and positive definite matrix, it is

$$\kappa_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}} = \mathcal{O}\left(h^{-2}\right).$$

Since the dimension of the matrix is large and the matrix is sparse, iterative solvers are an appropriate approach for solving the linear system of equations. The main costs for iterative solvers are the matrix-vector multiplications (often one per iteration). The cost of one matrix-vector multiplication is for sparse matrices proportional to the number of unknowns. Hence, an optimal solver with respect to the number of floating point operations is given if the number of operations for solving the linear system of equations is proportional to the number of unknowns. It is known that the number of iterations of many iterative solvers depends on the condition number of the matrix:

- *(damped) Jacobi method, SOR, SSOR.* The number of iteration is proportional to $\kappa_2(A)$. That means, if the grid is refined once, $h \to h/2$, then the number of unknowns is increased by around the factor 4 in two dimensions and also the number of iterations increases by a factor of around 4. Altogether, for one refinement step, the total costs increase by a factor of around 16.
- *(preconditioned) conjugate gradient (PCG) method.* The number of iterations is bounded in the worst case proportional to $\sqrt{\kappa_2(A)}$, see the corresponding theorem from the class Numerical Mathematics II. Then, the total costs increase by a factor of around 8 if the grid is refined once.
- *multigrid methods.* For multigrid methods, the number of iterations on each grid is bounded by a constant that is independent of the grid. Hence, the total costs are proportional to the number of unknowns and these methods are optimal. However, the implementation of multigrid methods is involved.

□

Remark 2.38. An eigenvalue problem. The derivation of an alternative direct solver is based on the eigenvalues and eigenvectors of the discrete Laplacian. It is possible to computed these quantities only in special

situations, e.g., if the Poisson problem with Dirichlet boundary conditions is considered, the domain is rectangular, and the Laplacian is approximated with the five point stencil.

Consider the following eigenvalue problem

$$\begin{aligned} -\Lambda v(\boldsymbol{x}) &= \lambda v(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \\ v(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \in \gamma_h. \end{aligned}$$

Denote the node $\boldsymbol{x} = (x_i, y_j)$ by $\boldsymbol{x}_{ij}$ and grid functions in a similar way. The solution of this problem is sought in (tensor-)product form (separation of variables)

$$v_{ij}^{(\boldsymbol{k})} = v_i^{(k_x),x} v_j^{(k_y),y}, \quad \boldsymbol{k} = (k_x, k_y)^T.$$

It is

$$\Lambda v_{ij}^{(\boldsymbol{k})} = \left(\Lambda_x v_i^{(k_x),x}\right) v_j^{(k_y),y} + v_i^{(k_x),x} \left(\Lambda_y v_j^{(k_y),y}\right) = -\lambda_{\boldsymbol{k}} v_i^{(k_x),x} v_j^{(k_y),y},$$

where $i = 0, \ldots, n_x$, $j = 0, \ldots, n_y$ refers to the nodes and $k_x = 1, \ldots, n_x - 1$, $k_y = 1, \ldots, n_y - 1$ refers to the eigenvalues. Notice that the number of eigenvalues is equal to the number of inner nodes, i.e., it is $(n_x - 1)(n_y - 1)$. In this ansatz, also a splitting of the eigenvalues in a contribution from the x coordinate and a contribution from the y coordinate is included. From the boundary condition, it follows that

$$v_0^{(k_x),x} = v_{n_x}^{(k_x),x} = v_0^{(k_y),y} = v_{n_y}^{(k_y),y} = 0.$$

Dividing by $v_i^{(k_x),x} v_j^{(k_y),y}$ and rearranging terms, the eigenvalue problem can be split

$$\frac{\Lambda_x v_i^{(k_x),x}}{v_i^{(k_x),x}} + \lambda_{k_x}^{(x)} = -\frac{\Lambda_y v_j^{(k_y),y}}{v_j^{(k_y),y}} - \lambda_{k_y}^{(y)}$$

with $\lambda_{\boldsymbol{k}} = \lambda_{k_x}^{(x)} + \lambda_{k_y}^{(y)}$. Both sides of this equation have to be constant since one of them depends only on i, i.e., on x, and the other one only on j, i.e., on y. The splitting of $\lambda_{\boldsymbol{k}}$ can be chosen so that the constant is zero. Then, one gets

$$\Lambda_x v_i^{(k_x),x} + \lambda_{k_x}^{(x)} v_i^{(k_x),x} = 0, \quad \Lambda_y v_j^{(k_y),y} + \lambda_{k_y}^{(y)} v_j^{(k_y),y} = 0.$$

The solution of these eigenvalue problems is known (exercise)

$$\begin{aligned} v_i^{(k_x),x} &= \sqrt{\frac{2}{l_x}} \sin\left(\frac{k_x \pi i}{n_x}\right), \quad \lambda_{k_x}^{(x)} = \frac{4}{h_x^2} \sin^2\left(\frac{k_x \pi}{2n_x}\right), \\ v_j^{(k_y),y} &= \sqrt{\frac{2}{l_y}} \sin\left(\frac{k_y \pi j}{n_y}\right), \quad \lambda_{k_y}^{(y)} = \frac{4}{h_y^2} \sin^2\left(\frac{k_y \pi}{2n_y}\right). \end{aligned}$$

It follows that the solution of the full eigenvalue problem is

$$v_{ij}^{(\boldsymbol{k})} = \frac{2}{\sqrt{l_x l_y}} \sin\left(\frac{k_x \pi i}{n_x}\right) \sin\left(\frac{k_y \pi j}{n_y}\right), \tag{2.25}$$

$$\lambda_{\boldsymbol{k}} = \frac{4}{h_x^2} \sin^2\left(\frac{k_x \pi}{2n_x}\right) + \frac{4}{h_y^2} \sin^2\left(\frac{k_y \pi}{2n_y}\right), \tag{2.26}$$

with $i = 0, \ldots, n_x, j = 0, \ldots, n_y$ and $k_x = 1, \ldots, n_x - 1, k_y = 1, \ldots, n_y - 1$. For every index $\boldsymbol{k} = (k_x, k_y)$, the eigenvalue is given by (2.26) and the entry of the corresponding eigen-grid-function $v(\boldsymbol{x})$ in the node $\boldsymbol{x}_{ij}$ is given by (2.25).

Using a Taylor series expansion, one obtains now the asymptotic behavior of the eigenvalues as given in (2.24). Note that because of the splitting of the eigenvalues into the directional contributions, the number of individual terms for computing the eigenvalues is only proportional to $(n_x + n_y)$. □

Remark 2.39. On the eigenvectors, weighted Euclidean inner product. Since the matrix corresponding to Λ is symmetric, the eigenvectors are orthogonal with respect to the Euclidean vector product. They become orthonormal with respect to the weighted Euclidean vector product

$$\langle u, v \rangle = h_x h_y \sum_{\boldsymbol{x} \in \omega_h \cup \gamma_h} u(\boldsymbol{x}) v(\boldsymbol{x}) = h_x h_y \sum_{i=0}^{n_x} \sum_{j=0}^{n_y} u_{ij} v_{ij}, \tag{2.27}$$

with

$$h_x = \frac{l_x}{n_x}, h_y = \frac{l_y}{n_y},$$

i.e., then it is

$$\langle v^{(\boldsymbol{k})}, v^{(\boldsymbol{m})} \rangle = \delta_{\boldsymbol{k},\boldsymbol{m}}. \tag{2.28}$$

This property can be checked by using the relation

$$\sum_{i=0}^{n} \sin^2 \left(\frac{i\pi}{n} \right) = \frac{n}{2}, \quad n > 1.$$

The norm induced by the weighted Euclidean vector product is given by

$$\|v\|_h = \langle v, v \rangle^{1/2} = \left(h_x h_y \sum_{i=0}^{n_x} \sum_{j=0}^{n_y} v_{ij}^2 \right)^{1/2}. \tag{2.29}$$

The weights are such that this norm can be bounded for constant grid functions independently of the mesh, i.e.,

$$\|1\|_h = \left(h_x h_y (n_x + 1)(n_y + 1) \right)^{1/2} = \left(l_x l_y \frac{n_x + 1}{n_x} \frac{n_y + 1}{n_y} \right)^{1/2} \leq 2 \left(l_x l_y \right)^{1/2}. \tag{2.30}$$

□

Remark 2.40. Solver based on the eigenvalues and eigenvectors. Let $\phi(\boldsymbol{x})$ be the grid function corresponding to the right-hand side vector $\underline{f}$, see (2.23). Then, one uses the ansatz

$$\phi(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \langle \phi, v^{(\boldsymbol{k})} \rangle v^{(\boldsymbol{k})}(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \phi_{\boldsymbol{k}} v^{(\boldsymbol{k})}(\boldsymbol{x}) \tag{2.31}$$

with the Fourier coefficients

$$\phi_{\boldsymbol{k}} = \left\langle \phi, v^{(\boldsymbol{k})} \right\rangle = \frac{2 h_x h_y}{\sqrt{l_x l_y}} \sum_{i=0}^{n_x} \sum_{j=0}^{n_y} \phi_{ij} \sin \left(\frac{k_x \pi i}{n_x} \right) \sin \left(\frac{k_y \pi j}{n_y} \right), \quad \boldsymbol{k} = (k_x, k_y),$$

with $\phi_{ij} = \phi(\boldsymbol{x}_{ij})$. The solution $u(\boldsymbol{x})$ of (2.22) is sought as a linear combination of the eigenfunctions

$$u(\boldsymbol{x}) = \sum_{\boldsymbol{k}} u_{\boldsymbol{k}} v^{(\boldsymbol{k})}(\boldsymbol{x})$$

with unknown coefficients $u_{\boldsymbol{k}}$. With this ansatz, one obtains for the finite difference operator

$$\Lambda u = \sum_{\boldsymbol{k}} u_{\boldsymbol{k}} \Lambda v^{(\boldsymbol{k})} = \sum_{\boldsymbol{k}} u_{\boldsymbol{k}} \lambda_{\boldsymbol{k}} v^{(\boldsymbol{k})}.$$

Since the eigenfunctions form a basis of the space of the grid functions, a comparison of the coefficients with the right-hand side (2.31) gives

$$-u_{\boldsymbol{k}} \lambda_{\boldsymbol{k}} = \phi_{\boldsymbol{k}} \quad \Longleftrightarrow \quad u_{\boldsymbol{k}} = -\frac{\phi_{\boldsymbol{k}}}{\lambda_{\boldsymbol{k}}}$$

or, for each component, using (2.25),

$$u_{ij} = -\sum_{\boldsymbol{k}} \frac{\phi_{\boldsymbol{k}}}{\lambda_{\boldsymbol{k}}} v_{ij}^{(\boldsymbol{k})} = -\frac{2h_x h_y}{\sqrt{l_x l_y}} \sum_{k_x=1}^{n_x-1} \sum_{k_y=1}^{n_y-1} \frac{\phi_{\boldsymbol{k}}}{\lambda_{\boldsymbol{k}}} \sin\left(\frac{k_x \pi i}{n_x}\right) \sin\left(\frac{k_y \pi j}{n_y}\right),$$

$i = 0, \ldots, n_x$, $j = 0, \ldots, n_y$.

It is possible to implement this approach with the Fast Fourier Transform (FFT) with

$$\mathcal{O}\left(n_x n_y \log_2 n_x + n_x n_y \log_2 n_y\right) = \mathcal{O}\left(N \log_2 N\right), \; N = (n_x - 1)(n_y - 1),$$

operations. Hence, this method is almost, up to a logarithmic factor, optimal. □

2.6 A Higher Order Discretization

Remark 2.41. Contents. The five point stencil is a second order discretization of the Laplacian. In this section, a discretization of higher order will be studied. In these studies, only the case of a rectangular domain $\Omega = (0, l_x) \times (0, l_y)$ and Dirichlet boundary conditions will be considered. □

Remark 2.42. Derivation of a fourth order approximation. Let $u(\boldsymbol{x})$ be the solution of the Poisson problem (2.1) and assume that $u(\boldsymbol{x})$ is sufficiently smooth. It is

$$Lu(\boldsymbol{x}) = \Delta u(\boldsymbol{x}) = L_x u(\boldsymbol{x}) + L_y u(\boldsymbol{x}), \quad L_\alpha u := \frac{\partial^2 u}{\partial x_\alpha^2}.$$

Let the five point stencil be represented by the following operator

$$\Lambda u = \Lambda_x u + \Lambda_y u.$$

Applying a Taylor series expansion and using the notation $L_\alpha^2 u = L_\alpha(L_\alpha u)$, one finds that

$$\Lambda u - \Delta u = \frac{h_x^2}{12} L_x^2 u + \frac{h_y^2}{12} L_y^2 u + \mathcal{O}\left(h^4\right). \tag{2.32}$$

From the equation $-Lu = f$, it follows with differentiation that

$$L_x^2 u = -L_x f - L_x L_y u, \quad L_y^2 u = -L_y f - L_y L_x u.$$

Inserting these expressions in (2.32) gives

$$\Lambda u - \Delta u = -\frac{h_x^2}{12} L_x f - \frac{h_y^2}{12} L_y f - \frac{h_x^2 + h_y^2}{12} L_x L_y u + \mathcal{O}\left(h^4\right). \tag{2.33}$$

The operator $L_x L_y = \frac{\partial^4}{\partial x^2 \partial y^2}$ can be approximated as follows

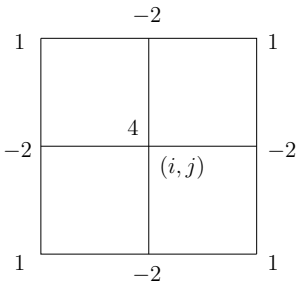


Fig. 2.7 Contribution from $\Lambda_x\Lambda_y$ to the nine point stencil.

$$L_xL_yu \approx \Lambda_x\Lambda_yu = u_{\overline{x}x\overline{y}y}.$$

The difference operator in this approximation requires nine points, see Figure 2.7,

$$\begin{aligned}\Lambda_x\Lambda_yu = \frac{1}{h_x^2h_y^2}\Big(&u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1} - 2u_{i+1,j} + 4u_{ij}\\ &-2u_{i-1,j} + u_{i+1,j-1} - 2u_{i,j-1} + u_{i-1,j-1}\Big).\end{aligned}$$

Therefore it is called a nine point stencil. One checks, as usual by using a Taylor series expansion, that this approximation is of second order

$$L_xL_yu - \Lambda_x\Lambda_yu = \mathcal{O}\left(h^2\right).$$

Inserting this expansion in (2.33) and using the expansion (2.32) for the differential operator shows that the difference equation

$$-\left(\Lambda + \frac{h_x^2+h_y^2}{12}\Lambda_x\Lambda_y\right)u = \left(f + \frac{h_x^2}{12}L_xf + \frac{h_y^2}{12}L_yf\right)$$

is a fourth order approximation of the differential equation (2.1). In addition, one can replace the derivatives of $f(\boldsymbol{x})$ also by finite differences

$$L_xf = \Lambda_xf + \mathcal{O}\left(h_x^2\right), \quad L_yf = \Lambda_yf + \mathcal{O}\left(h_y^2\right).$$

Finally, using (2.1) to replace the Laplacian, one obtains a finite difference equation $-\Lambda'u = \phi$ with

$$\Lambda' = \Lambda_x + \Lambda_y + \frac{h_x^2+h_y^2}{12}\Lambda_x\Lambda_y, \quad \phi = f + \frac{h_x^2}{12}\Lambda_xf + \frac{h_y^2}{12}\Lambda_yf.$$

Deriving the actual form of the nine point stencil is part of a programming exercise problem. Compared to the five point stencil problem (2.6), (2.7) one has to change both sides of the finite difference problem, not only the five point stencil has to be replaced by the nine point stencil but also the right-hand side has to be approximated more accurately than it is necessary for the five point stencil problem. □

Remark 2.43. On the convergence of the fourth order approximation. The finite difference problem with the higher order approximation property can be written with the help of the second order differences. Since the convergence proof is based on the five point stencil, the following lemma considers this stencil. It will be proved that one can estimate the values of the grid function by the second order differences. This result will be used in the convergence proof for the fourth order approximation. □

Lemma 2.44. Stability estimate. *Let*

$$\omega_h = \{(ih_x, jh_y) \ : \ i = 1, \ldots, n_x - 1, j = 1, \ldots, n_y - 1\},$$

and let y be a grid function on $\omega_h \cup \gamma_h$ with $y(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \gamma_h$. Then, the following estimate holds

$$\|y\|_{l^\infty(\omega_h \cup \gamma_h)} \leq M \, \|\Lambda y\|_h \,,$$

with the mesh-independent constant $M = \frac{\max\{l_x^2, l_y^2\}}{2\sqrt{l_x l_y}}$, Λ is also used to symbolize the matrix obtained by using the five point stencil $\Lambda = \Lambda_x + \Lambda_y$ for approximating the second derivatives, and the norm on the right-hand side is defined in (2.29)*.*

Proof. Let $\{v^{\boldsymbol{k}}(\boldsymbol{x})\}$, $\boldsymbol{k} = (k_x, k_y)$, be the orthonormal basis with

$$v^{\boldsymbol{k}}(\boldsymbol{x}_{ij}) = v_{ij}^{\boldsymbol{k}} = \frac{2}{\sqrt{l_x l_y}} \sin\left(\frac{k_x \pi i}{n_x}\right) \sin\left(\frac{k_y \pi j}{n_y}\right),$$

which was derived in Remark 2.38. Then, there is a unique representation of the grid function $y = \sum_{\boldsymbol{k}} y_{\boldsymbol{k}} v^{\boldsymbol{k}}$ and it holds with (2.27) and (2.28)

$$\Lambda y = \sum_{\boldsymbol{k}} y_{\boldsymbol{k}} \lambda_{\boldsymbol{k}} v^{\boldsymbol{k}}, \quad \|\Lambda y\|_h^2 = \sum_{\boldsymbol{k}} y_{\boldsymbol{k}}^2 \lambda_{\boldsymbol{k}}^2. \tag{2.34}$$

It follows for $\boldsymbol{x} \in \omega_h$, because of $|\sin(x)| \leq 1$ for all $x \in \mathbb{R}$, that

$$|y(\boldsymbol{x})| = \left| \sum_{\boldsymbol{k}} y_{\boldsymbol{k}} v^{\boldsymbol{k}}(\boldsymbol{x}) \right| \leq \sum_{\boldsymbol{k}} |y_{\boldsymbol{k}}| \left| v^{\boldsymbol{k}}(\boldsymbol{x}) \right| \leq \frac{2}{\sqrt{l_x l_y}} \sum_{\boldsymbol{k}} |y_{\boldsymbol{k}}| \,.$$

Using this estimate, applying the Cauchy–Schwarz inequality for sums, and utilizing (2.34) gives

$$|y(\boldsymbol{x})|^2 \leq \frac{4}{l_x l_y} \left(\sum_{\boldsymbol{k}} |y_{\boldsymbol{k}}| \right)^2 = \frac{4}{l_x l_y} \left(\sum_{\boldsymbol{k}} |\lambda_{\boldsymbol{k}} y_{\boldsymbol{k}}| \, \frac{1}{\lambda_{\boldsymbol{k}}} \right)^2 \leq \frac{4}{l_x l_y} \sum_{\boldsymbol{k}} \lambda_{\boldsymbol{k}}^2 y_{\boldsymbol{k}}^2 \sum_{\boldsymbol{k}} \frac{1}{\lambda_{\boldsymbol{k}}^2} = \frac{4}{l_x l_y} \|\Lambda y\|_h^2 \sum_{\boldsymbol{k}} \frac{1}{\lambda_{\boldsymbol{k}}^2}. \tag{2.35}$$

Now, one has to estimate the last sum. It is already known that

$$\lambda_{\boldsymbol{k}} = \frac{4}{h_x^2} \sin^2\left(\frac{k_x \pi}{2n_x}\right) + \frac{4}{h_y^2} \sin^2\left(\frac{k_y \pi}{2n_y}\right), \quad k_x = 1, \ldots, n_x - 1, \ k_y = 1, \ldots, n_y - 1.$$

Setting $l = \max\{l_x, l_y\}$ and $h_\alpha = l_\alpha / n_\alpha, \phi_\alpha = \frac{k_\alpha \pi}{2n_\alpha} \in (0, \pi/2)$, $\alpha \in \{x, y\}$, leads to

$$\lambda_{\boldsymbol{k}} = \frac{k_x^2 \pi^2}{l_x^2} \left(\frac{\sin \phi_x}{\phi_x} \right)^2 + \frac{k_y^2 \pi^2}{l_y^2} \left(\frac{\sin \phi_y}{\phi_y} \right)^2 \geq 4 \left(\frac{k_x^2}{l_x^2} + \frac{k_y^2}{l_y^2} \right) \geq \frac{4}{l^2} \left(k_x^2 + k_y^2 \right). \tag{2.36}$$

In performing this estimate, it was used that the function $\sin(\phi)/\phi$ is monotonically decreasing on $(0, \pi/2)$, see Figure 2.8, and that

$$\frac{\sin \phi}{\phi} \geq \frac{\sin(\pi/2)}{\pi/2} = \frac{2}{\pi} \quad \forall \ \phi \in (0, \pi/2).$$

The estimate will be continued by constructing a function that majorizes $(k_x^2 + k_y^2)^{-2}$ and that can be easily integrated. Let $G = \{(x, y) \ : \ x > 0, y > 0, x^2 + y^2 > 1\}$ be the first quadrant of the complex plane without the part that belongs to the unit circle, see Figure 2.9. The function $(x^2 + y^2)^{-2}$ has its smallest value in the square $[k_x - 1, k_x] \times [k_y - 1, k_y]$ in the point (k_x, k_y). Using the lower estimate (2.36) of $\lambda_{\boldsymbol{k}}$, one obtains

$$\begin{aligned} \sum_{\boldsymbol{k}, \boldsymbol{k} \neq (1,1)} \frac{1}{\lambda_{\boldsymbol{k}}^2} \quad &\leq \quad \frac{l^4}{16} \sum_{\boldsymbol{k}, \boldsymbol{k} \neq (1,1)} \left(k_x^2 + k_y^2 \right)^{-2} \\ &= \quad \frac{l^4}{16} \sum_{\boldsymbol{k}, \boldsymbol{k} \neq (1,1)} \underbrace{\left(k_x^2 + k_y^2 \right)^{-2}}_{\text{smallest value in square}} \underbrace{\int_{k_x - 1}^{k_x} \int_{k_y - 1}^{k_y} dy dx}_{=1} \end{aligned}$$

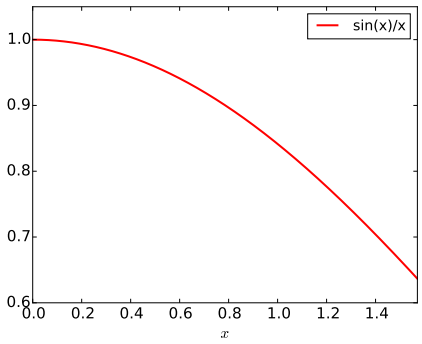


Fig. 2.8 The function $\sin(\phi)/\phi$.

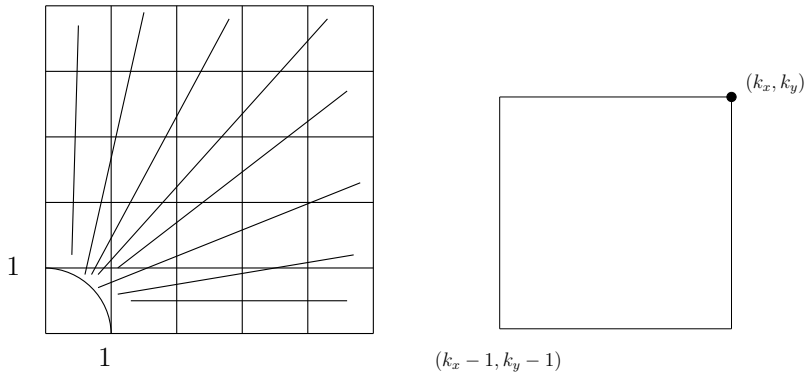


Fig. 2.9 Illustration to the proof of Lemma 2.44.

$$
\begin{aligned}
&= \frac{l^4}{16} \sum_{\boldsymbol{k},\boldsymbol{k}\neq(1,1)} \int_{k_x-1}^{k_x} \int_{k_y-1}^{k_y} \left(k_x^2 + k_y^2\right)^{-2} \, dydx \\
&\leq \frac{l^4}{16} \sum_{\boldsymbol{k},\boldsymbol{k}\neq(1,1)} \int_{k_x-1}^{k_x} \int_{k_y-1}^{k_y} \left(x^2 + y^2\right)^{-2} \, dydx \\
&\leq \frac{l^4}{16} \int_G \left(x^2 + y^2\right)^{-2} \, dxdy \\
&\overset{\text{polar coord.}}{=} \frac{l^4}{16} \int_1^\infty \int_0^{\pi/2} \frac{\rho}{\rho^4} \, d\phi d\rho = \frac{l^4}{16} \frac{\pi}{2} \left(-\frac{\rho^{-2}}{2} \Big|_{\rho=1}^{\rho=\infty} \right) = \frac{\pi l^4}{64}.
\end{aligned}
$$

For performing this computation, one has to exclude $\rho \to 0$.

For $\lambda_{(1,1)}$, it is

$$
\begin{aligned}
\lambda_{(1,1)} &= \frac{4}{h_x^2} \sin^2\left(\frac{\pi}{2n_x}\right) + \frac{4}{h_y^2} \sin^2\left(\frac{\pi}{2n_y}\right) = \frac{4}{h_x^2} \sin^2\left(\frac{h_x\pi}{2l_x}\right) + \frac{4}{h_y^2} \sin^2\left(\frac{h_y\pi}{2l_y}\right) \\
&= \frac{\pi^2}{l_x^2} \left(\frac{2l_x}{h_x\pi}\right)^2 \sin^2\left(\frac{h_x\pi}{2l_x}\right) + \frac{\pi^2}{l_y^2} \left(\frac{2l_y}{h_y\pi}\right)^2 \sin^2\left(\frac{h_y\pi}{2l_y}\right) \\
&\geq \frac{\pi^2}{l_x^2} \frac{8}{\pi^2} + \frac{\pi^2}{l_y^2} \frac{8}{\pi^2} \geq \frac{16}{l^2}.
\end{aligned} \tag{2.37}
$$

For this estimate, the following relations and the monotonicity of $\sin(x)/x$, see Figure 2.8, were used

$$h_\alpha \le \frac{l_\alpha}{2}, \quad \phi_\alpha = \frac{h_\alpha \pi}{2 l_\alpha} \le \frac{\pi}{4}, \quad \left(\frac{\sin \phi_\alpha}{\phi_\alpha}\right)^2 \ge \left(\frac{\sin(\pi/4)}{\pi/4}\right)^2 = \frac{8}{\pi^2}.$$

Collecting all estimates gives

$$\sum_{\boldsymbol{k}} \frac{1}{\lambda_{\boldsymbol{k}}^2} = \lambda_{(1,1)}^{-2} + \sum_{\boldsymbol{k}, \boldsymbol{k} \neq (1,1)} \frac{1}{\lambda_{\boldsymbol{k}}^2} \le \frac{l^4}{256} + \frac{\pi l^4}{64} \le \frac{l^4}{16}.$$

Inserting this estimate in (2.35), the final bound has the form

$$\|y\|_{l^\infty(\omega_h \cup \gamma_h)} \le \frac{2}{\sqrt{l_x l_y}} \|\Lambda y\|_h \frac{l^2}{4} =: M \|\Lambda y\|_h .$$

■

Theorem 2.45. Convergence of the higher order finite difference scheme. *Let $\Omega = (0, l_x) \times (0, l_y)$. The finite difference scheme*

$$\begin{aligned} -\Lambda' u(\boldsymbol{x}) &= \phi(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \\ u(\boldsymbol{x}) &= g(\boldsymbol{x}),\ \boldsymbol{x} \in \gamma_h, \end{aligned}$$

with

$$\Lambda' = \Lambda_x + \Lambda_y + \frac{h_x^2 + h_y^2}{12} \Lambda_x \Lambda_y, \quad \phi = f + \frac{h_x^2}{12} \Lambda_x f + \frac{h_y^2}{12} \Lambda_y f,$$

converges of fourth order.

Proof. Analogously as in the proof of Theorem 2.34, one finds that the following equation holds for the error $e = u(x_i, y_j) - u_{ij}$:

$$\begin{aligned} -\Lambda' e(\boldsymbol{x}) &= \psi(\boldsymbol{x}),\ \psi = \mathcal{O}\left(h^4\right), \boldsymbol{x} \in \omega_h, \\ e(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \in \gamma_h. \end{aligned}$$

Let Ω_h be the vector space of grid functions, which are non-zero only in the interior, i.e., at the nodes from ω_h, and which vanish on γ_h. Let $A_\alpha y = -\Lambda_\alpha y$, $y \in \Omega_h$, $\alpha \in \{x, y\}$. The operators $A_\alpha : \Omega_h \to \Omega_h$ are linear and they have the following properties:

- They are symmetric and positive definite, i.e., $A_\alpha = A_\alpha^* > 0$, where A_α^* is the adjoint (transposed) of A_α, and $(A_\alpha u, v) = (u, A_\alpha v), \forall\, u, v \in \Omega_h$. The square root can be defined in the same way as it is known for symmetric positive definite matrices.
- They are elliptic, i.e., $(A_\alpha u, u) \ge \lambda_1^{(\alpha)} (u, u)$, $\forall u \in \Omega_h$, with

$$\lambda_1^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2\left(\frac{\pi h_\alpha}{2 l_\alpha}\right) \ge \frac{8}{l_\alpha^2},$$

  see (2.37).
- They are bounded, $\|A_\alpha\|_2 \le 4/h_\alpha^2$, since the spectral norm of a symmetric positive definite matrix equals the largest eigenvalue. In addition, using the Rayleigh quotient, it holds $(A_\alpha u, u)/(u, u) \le \lambda_{n_\alpha - 1}^{(\alpha)}$ with

$$\lambda_{n_\alpha - 1}^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2\left(\frac{k_\alpha \pi}{2 n_\alpha}\right) \le \frac{4}{h_\alpha^2} \quad \Longrightarrow \quad (A_\alpha u, u) \le \frac{4}{h_\alpha^2} (u, u). \tag{2.38}$$

- They commute, i.e., it is $A_x A_y = A_y A_x$.
- It holds $A_x A_y = \left(A_x A_y\right)^*$.

The error equation on ω_h is given by

$$A_x e + A_y e - (\kappa_x + \kappa_y) A_x A_y e = A' e = \psi \quad \text{with} \quad \kappa_\alpha = \frac{h_\alpha^2}{12}. \tag{2.39}$$

Using the commutativity of the operators, one finds with (2.38) for all $v \in \Omega_h$ that

$$\begin{aligned} \left(\kappa_x A_x A_y v + \kappa_y A_x A_y v, v\right) &= \left(\left(\kappa_x A_y\right) A_x v, v\right) + \left(\left(\kappa_y A_x\right) A_y v, v\right) \\ &= \kappa_x \left(A_x A_y^{1/2} v, A_y^{1/2} v\right) + \kappa_y \left(A_y A_x^{1/2} v, A_x^{1/2} v\right) \\ &\le \frac{h_x^2}{12} \frac{4}{h_x^2} \left(A_y v, v\right) + \frac{h_y^2}{12} \frac{4}{h_y^2} \left(A_x v, v\right) \end{aligned}$$

$$= \frac{1}{3}\left(\left(A_x + A_y\right)v, v\right).$$

Now, it follows for all $v \in \Omega_h$ that

$$(A'v, v) = \left(\left(A_x + A_y\right)v, v\right) - \left(\kappa_x A_x A_y v + \kappa_y A_x A_y v, v\right) \geq \frac{2}{3}\left(\left(A_x + A_y\right)v, v\right) \geq 0.$$

The matrices on both sides of this inequality are symmetric and because the matrix on the lower estimate is positive definite, also the matrix at the upper estimate is positive definite. The matrices commute since the order of applying the finite differences in x and y direction does not matter. Using these properties, one gets (*exercise*)

$$\left\| \frac{2}{3}\left(A_x + A_y\right)e \right\|_h \leq \left\|A'e\right\|_h = \|\psi\|_h, \tag{2.40}$$

where the last equality follows from (2.39). The application of Lemma 2.44 to the error, (2.40), and (2.30) yields

$$\begin{aligned}
\|e\|_{l^\infty(\omega_h \cup \gamma_h)} &\leq \frac{l^2}{2\sqrt{l_x l_y}} \left\|\left(\Lambda_x + \Lambda_y\right)e\right\|_h \leq \frac{3l^2}{4\sqrt{l_x l_y}} \left\|A'e\right\|_h = \frac{3l^2}{4\sqrt{l_x l_y}} \|\psi\|_h \\
&\leq \frac{3l^2}{4\sqrt{l_x l_y}} \left(h_x h_y (n_x+1)(n_y+1)\right)^{1/2} \|\psi\|_{l^\infty(\omega_h \cup \gamma_h)} \\
&= \frac{3l^2}{4} \left(\frac{n_x+1}{n_x}\frac{n_y+1}{n_y}\right)^{1/2} \|\psi\|_{l^\infty(\omega_h \cup \gamma_h)} = \mathcal{O}\left(h^4\right).
\end{aligned}$$

■

Remark 2.46. On the discrete maximum principle. Reformulation of the finite difference scheme $-\Lambda' u = \phi$ in the form studied for the discrete maximum principle gives for the node (i, j)

$$\begin{aligned}
a(\boldsymbol{x})u(\boldsymbol{x}) &= \sum_{\boldsymbol{y} \in S(\boldsymbol{x})} b(\boldsymbol{x}, \boldsymbol{y})u(\boldsymbol{y}) + \phi(\boldsymbol{x}), \\
a(\boldsymbol{x}) &= \frac{2}{h_x^2} + \frac{2}{h_y^2} - \frac{1}{12}\left(h_x^2 + h_y^2\right)\frac{4}{h_x^2 h_y^2} = \frac{5}{3}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) > 0, \\
b(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{h_x^2} - \frac{1}{12}\left(h_x^2 + h_y^2\right)\frac{2}{h_x^2 h_y^2} = \frac{1}{6}\left(\frac{5}{h_x^2} - \frac{1}{h_y^2}\right), \ i \pm 1, j, \\
&\qquad\qquad \text{(left, right node)} \\
b(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{6}\left(-\frac{1}{h_x^2} + \frac{5}{h_y^2}\right), \ i, j \pm 1, \ \text{(bottom, top node)} \\
b(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{12}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right), \ i \pm 1, j \pm 1, \ \text{(other neighbors)}.
\end{aligned}$$

Hence, the assumptions for the discrete maximum principle, see Remark 2.15, are satisfied only if

$$\frac{1}{\sqrt{5}} < \frac{h_x}{h_y} < \sqrt{5}.$$

Consequently, the ratio of the grid widths has to be bounded and it has to be of order one. In this case, one speaks of an isotropic grid. □

2.7 Summary

Remark 2.47. Summary.

- Finite difference methods are the simplest approach for discretizing boundary value problems with partial differential equations. The derivatives are just approximated by difference quotients.
- They are very popular in the engineering community.
- One large drawback is the difficulty to approximate domains that are not of tensor-product type. However, in the engineering communities, a number of strategies have been developed to deal with this issue in practice.
- Another drawback arises from the point of view of numerical analysis. The numerical analysis of finite difference methods is mainly based on Taylor series expansions. For this tool to be applicable, one has to assume a high regularity of the solution. These assumptions are generally not realistic.
- In Numerical Mathematics, one considers often other schemes than finite difference methods. However, there are problems in practice, where finite difference methods can compete with other discretizations.

□