2.7.2. Finite Difference Schemes in 2D. Compared to the 1D case, higher dimensional problems lead to additional complications, because

1) the elliptic operator L is more complicated, in particular contains mixed derivatives,
2) the spatial domain Ω may be very complicated.

As far as these topics are concerned, it is sufficient to consider the 2D case. Also, the main ideas can be best explained for this setup which is the reason why we study 2D problems as representatives of higher dimensional ones.

To avoid too many difficulties entering at the same time, we first consider a simple bounded domain in $\mathbb{R}^2$, namely the unit square $\Omega = (0,1)^2$ and start with the Dirichlet Poisson problem, i.e.

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

We discretize the unit square Ω and the boundary $\Gamma = \partial\Omega$ by an equidistant grid in both directions,

$$x_i = ih, \quad y_j = jh, \quad i,j = 0,\dots,n, \quad h = \frac{1}{n}$$

and call Ω_h the set of interior grid points and Γ_h the set of grid points on the boundary $\partial\Omega$. For a given function u, we abbreviate $u_{i,j} = u(x_i, y_j)$. Then, if u is smooth enough,

$$\begin{aligned} u_{xx}(x_i, y_j) &= \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \mathcal{O}(h^2) \\ u_{yy}(x_i, y_j) &= \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2} + \mathcal{O}(h^2) \end{aligned}$$

such that

$$-\Delta u(x_i, y_j) = \frac{-u_{i-1,j} - u_{i,j-1} + 4u_{i,j} - u_{i,j+1} - u_{i+1,j}}{h^2} + \mathcal{O}(h^2).$$

(Here, u_{xx} and u_{yy} denote second derivatives with respect to x and y.) Beside the difference quotient, one can also use a more intuitive notation to represent the discretization. The weight of the value $u_{i+k,j+l}$ is taken as entry in a matrix at position (k,l). The center of the matrix is the point $(0,0)$. For the Laplacian, the stencil has the form

$$\frac{1}{h^2}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

(usually, zeroes in the stencil are suppressed). Using this, so called five–point discretization, we arrive at the discrete problem to find values $u_{i,j}$ for $i,j = 0,\dots,n$ such that

$$\begin{aligned} \frac{-u_{i-1,j} - u_{i,j-1} + 4u_{i,j} - u_{i,j+1} - u_{i+1,j}}{h^2} &= f_{i,j} && (x_i, y_j) \in \Omega_h, \\ u_{i,j} &= g_{i,j} && (x_i, y_j) \in \Gamma_h \end{aligned}$$

where $f_{i,j} = f(x_i, y_j)$ and $g_{i,j} = g(x_i, y_j)$. Again, the discretization is linear and we may write it in matrix vector form. In contrast to the 1D case, however, there is no natural ordering of the indices so that it is not obvious, how to proceed:

a) one may use a row–wise ordering, i.e. ordering first in x–direction, which yields

$$\mathbf{u}_h = \begin{pmatrix} u_{00} \\ \vdots \\ u_{n0} \\ u_{01} \\ \vdots \end{pmatrix}$$

b) the column–wise ordering, i.e. ordering first in y–direction, is given by

$$\mathbf{u}_h = \begin{pmatrix} u_{00} \\ \vdots \\ u_{0n} \\ u_{10} \\ \vdots \end{pmatrix}$$

c) checkerboard ordering and others.

The structure of the matrix strongly depends on the ordering and this may be exploited to speed up numerical solvers for linear systems. In our example we use row–wise ordering and get the block tridiagonal matrix

$$\mathcal{L}_h = \begin{pmatrix} I & & & & \\ B & A & B & & \\ & \ddots & \ddots & \ddots & \\ & & B & A & B \\ & & & & I \end{pmatrix}$$

with $(n+1)\times(n+1)$ block matrices of the form

$$B = \frac{1}{h^2}\begin{pmatrix} 0 & & & & \\ & -1 & & & \\ & & \ddots & & \\ & & & -1 & \\ & & & & 0 \end{pmatrix}, \qquad A = \frac{1}{h^2}\begin{pmatrix} h^2 & & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & & h^2 \end{pmatrix}$$

(I denotes the $(n+1)\times(n+1)$ identity matrix). (Note that column–wise ordering leads to the same matrix for our special geometry.) In compact form, the linear system reads

$$\mathcal{L}_h \mathbf{u}_h = \mathbf{d}_h$$

where the right hand side is given by

$$\mathbf{d}_h = \begin{pmatrix} g_{00} \\ \vdots \\ g_{n0} \\ g_{01} \\ f_{11} \\ \vdots \\ f_{n-1,1} \\ g_{n,n} \\ g_{02} \\ f_{12} \\ \vdots \\ f_{n-1,n-1} \\ g_{n,n-1} \\ g_{0n} \\ \vdots \\ g_{nn} \end{pmatrix}$$

In order to show convergence of the so defined scheme, we cannot proceed exactly like in the 1D case. In fact, the unit square is not a $C^{2,\alpha}$ domain so that the theoretical result does not apply. Consequently, the crucial construction of a vector $y \gg 0$ does not carry over, because we used consistency of the scheme in the solution of the auxiliary problem

$$\begin{aligned} -\Delta v &= 1 \qquad \text{in } \Omega \\ v &= 1 \qquad \text{on } \partial\Omega \end{aligned}$$

which need not be sufficiently smooth on the square. Nevertheless, we can still construct the vector $y \gg 0$ by another argument. We just take the function $v(x,y) = 1 + x(1-x)/2$ which satisfies

$$\sup_{\bar{\Omega}} |v| = 1 + \frac{1}{8}, \qquad v \geq 1 \quad \text{on } \bar{\Omega}, \qquad -\Delta v = 1.$$

Since v is a polynomial of degree two and our five–point approximation is second order accurate it follows immediately that the discretization $\mathbf{v}_h$ satisfies also

$$\frac{-v_{i-1,j} - v_{i,j-1} + 4v_{i,j} - v_{i,j+1} - v_{i+1,j}}{h^2} = 1.$$

Together with the estimate $v \geq 1$ on $\bar{\Omega}$ this implies

$$\mathbf{v}_h \geq e, \qquad \mathcal{L}_h \mathbf{v}_h \geq e, \qquad e = (1, \ldots, 1)^T.$$

Since also $\|\mathbf{v}_h\|_{h,\infty} \leq 1 + 1/8$, our considerations from the last section carry over and show that

Lemma 2.32. *For h sufficiently small, the matrix $\mathcal{L}_h^{-1}$ exists and is positive. Moreover, we have stability with*

$$\|\mathcal{L}_h^{-1}\|_{h,\infty} \leq 1 + \frac{1}{8}.$$

In order to get convergence, we now only need consistency of the scheme in the solution of the Dirichlet Poisson problem (according to Theorem 2.24). Consequently,

Theorem 2.33. *The finite difference approximation of the Dirichlet Poisson problem converges to the exact solution u, if u is sufficiently smooth.*

Finally, we mention that our considerations on the eigenvalues of $\mathcal{L}_h$ can be taken over from the 1D case. Also, the unknowns on the boundary can be removed from the system which leads to the modified vector of unknowns

$$\bar{\mathbf{u}}_h = \begin{pmatrix} u_{1,1} \\ \vdots \\ u_{n-1,1} \\ u_{1,2} \\ \vdots \\ u_{n-1,2} \\ u_{1,n-1} \\ \vdots \\ u_{n-1,n-1} \end{pmatrix},$$

the matrix

$$\bar{\mathcal{L}}_h = \frac{1}{h^2} \underbrace{\begin{pmatrix} D & -I & & & \\ -I & D & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & D & -I \\ & & & -I & D \end{pmatrix}}_{n-1 \text{ blocks}}$$

with $(n-1)\times(n-1)$-matrices D given by

$$D = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}$$

and right hand side $\bar{\mathbf{d}}_h$

$$\bar{\mathbf{d}}_h = \begin{pmatrix} f_{11} + \frac{1}{h^2}(g_{10} + g_{01}) \\ f_{21} + \frac{1}{h^2}g_{20} \\ \vdots \\ f_{n-1,1} + \frac{1}{h^2}(g_{n-1,0} + g_{n,1}) \\ f_{12} + \frac{1}{h^2}g_{02} \\ \vdots \\ f_{n-1,2} + \frac{1}{h^2}g_{n,2} \\ \vdots \\ f_{1,n-1} + \frac{1}{h^2}(g_{0,n-1} + g_{1,n}) \\ f_{2,n-1} + \frac{1}{h^2}g_{2,n} \\ \vdots \\ f_{n-1,n-1} + \frac{1}{h^2}(g_{n,n-1} + g_{n-1,n}) \end{pmatrix}$$

Since $\bar{\mathcal{L}}_h$ is positive definite and symmetric, the modified system $\bar{L}_h\bar{\mathbf{u}}_h = \bar{\mathbf{d}}_h$ is favorable for a numerical treatment.

Before we continue with more general differential operators, let us briefly mention some remarkable similarities between the Laplacian Δ and its discrete counterpart Δ_h defined by

$$(\Delta_h\mathbf{u}_h)_{ij} = \frac{-u_{i-1,j} - u_{i,j-1} + 4u_{i,j} - u_{i,j+1} - u_{i+1,j}}{h^2}.$$

In accordance with the definition of harmonic functions, one can speak of discrete harmonic functions $\mathbf{u}_h$ if $\Delta_h\mathbf{u}_h = 0$ in Ω_h, or more precisely,

$$(\Delta_h\mathbf{u}_h)_{ij} = 0 \qquad \text{for } i,j \text{ such that } (x_i, y_j) \in \Omega_h. \tag{2.18}$$

Rewriting equation (2.18) we find

$$u_{i,j} = \frac{1}{4}(u_{i-1,j} + u_{i,j-1} + u_{i,j+1} + u_{i+1,j}) \tag{2.19}$$

which shows that the value of a discrete harmonic function at some point (x_i, y_j) is equal to the average over four neighboring values. This is a *discrete mean value property* which can be compared with

$$\Delta u = 0 \qquad \Longrightarrow \qquad u(x) = \frac{1}{|\partial B_R(x)|}\int_{\partial B_R(x)} u(y)\,d\sigma(y).$$

Similar to our treatment in Section 1.2.2, we can infer a maximum principle from the mean value property. Indeed, if $\mathbf{u}_h$ is a discrete harmonic function which has a maximum in (x_i, y_j), then (2.19) implies that the maximal value u_{ij} has to be found also in all four neighboring sites since otherwise the average would be smaller. Repeating the argument for the neighbors one reaches next neighbors and eventually all points of Ω_h (if the grid is connected) and most points of Γ_h ((observe that the corners of a discretized square are not reached in that way). In any case, one obtains the estimate

$$\max_{(x_i,y_j)\in\Omega_h} u_{ij} \le \max_{(x_i,y_j)\in\Gamma_h} u_{ij}$$

and repeating the argument also for the minimal values of $\mathbf{u}_h$

$$\min_{(x_i,y_j)\in\Omega_h} u_{ij} \ge \min_{(x_i,y_j)\in\Gamma_h} u_{ij}.$$

An important consequence of the maximum principle in the continuous case is uniqueness of the solution to Dirichlet Laplace problems. The same holds in the discrete case. Assume that $\mathbf{u}_h$ and $\mathbf{v}_h$ are two solutions $\mathcal{L}_h\mathbf{u}_h = \mathbf{d}_h$ and $\mathcal{L}_h\mathbf{v}_h = \mathbf{d}_h$ of the discrete problem. Then, the difference satisfies $\mathcal{L}_h\mathbf{w}_h = 0$ (i.e. $\mathbf{w}_h$ is in the kernel of $\mathcal{L}_h$). Since $\mathcal{L}_h\mathbf{w}_h = 0$ implies $\Delta_h\mathbf{w}_h = 0$ as well as $\mathbf{w}_h = 0$ on Γ_h, we find that

$$0 = \min_{\Gamma_h}\mathbf{w}_h \le \min_{\Omega_h}\mathbf{w}_h \le \max_{\Omega_h}\mathbf{w}_h \le \max_{\Gamma_h}\mathbf{w}_h = 0$$

and thus conclude $\mathbf{w}_h = 0$ which means $\mathbf{v}_h = \mathbf{u}_h$ is unique. For the operator $\mathcal{L}_h$, the argument tells us

$$\ker\mathcal{L}_h = \{0\}.$$

Since $\mathcal{L}_h$ is a square matrix this already implies invertibility (i.e. uniqueness of solutions implies existence). Please note how much effort is necessary to show existence in the continuous case (Sections 2.3 and 2.4) and how easily it follows form the maximum principle in the finite dimensional case. Continuous dependence on the data of the problem can be shown almost exactly as in the continuous case since it is another consequence of the maximum principle (the important point is that the bounds can be shown to be independent of h – of course for fixed h, continuity is simple because in finite dimensional spaces any linear operator is bounded). The approach in Section 2.1.1 to get a maximum principle for general operators also has its parallel in the discrete case. In fact, a necessary condition for $\mathbf{u}_h$ to have a maximum at (x_i, y_j) is that

$$u_{ij} \ge u_{i-1,j}, \qquad u_{ij} \ge u_{i+1,j}, \qquad u_{ij} \ge u_{i,j-1}, \qquad u_{ij} \ge u_{i,j+1}$$

which implies

$$\begin{aligned} -h^2(\Delta_h\mathbf{u}_h)_{ij} = (u_{ij} - u_{i-1,j}) + (u_{ij} - u_{i+1,j}) \\ + (u_{ij} - u_{i,j-1}) + (u_{ij} - u_{i,j+1}) \ge 0. \end{aligned}$$

Consequently, if $\Delta_h\mathbf{u}_h \ll 0$, $\mathbf{u}_h$ cannot have a maximum in Ω_h. For $\Delta_h\mathbf{u}_h \le 0$, perturbation arguments have to be used to arrive at similar conclusions as in Theorem 2.2. The last similarity between discrete and continuous formulation of the Dirichlet Laplace problem which we mention here gives some insight into the structure of $\mathcal{L}_h^{-1}$. Multiplying the matrix with the (i,j)–unit vector

$$e_{kl}^{(ij)} = \delta_{ik}\delta_{jl} = \begin{cases} 1 & i = k,\, j = l \\ 0 & \text{otherwise} \end{cases}$$

yields the corresponding column of $\mathcal{L}_h^{-1}$. Note that $\mathbf{e}_h^{(ij)}$ is the discretization of a function which is zero everywhere up to a small neighborhood of the point (x_i, y_j) where it is one. If we scale $\mathbf{e}_h^{(ij)}$ by $1/h^2$, the discrete integral

$$\sum_{k,l} \left(\frac{1}{h^2} e_{kl}^{(ij)} \right) h^2$$

is equal to one. Hence, the scaled unit vector $\mathbf{e}_h^{(ij)}$ has similar properties as the Dirac Delta distribution. Setting

$$\delta_h^{(ij)} := \frac{1}{h^2} \mathbf{e}_h^{(ij)},$$

the (scaled) (kl)–column of $\mathcal{L}_h^{-1}$ is

$$\mathbf{G}_h^{(kl)} = \mathcal{L}_h^{-1} \delta_h^{(kl)}.$$

Obviously, for $(x_k, y_l) \in \Omega_h$, this problem is related to the continuous one

$$\begin{aligned} -\Delta G &= \delta_{(x_k,y_l)} && \text{in } \Omega \\ G &= 0 && \text{on } \partial\Omega \end{aligned}$$

The (generalized) solution G to this problem is called Green's function. As we shall see in Section 2.8, G is essentially the fundamental solution presented in Section 1.2.2 (up to a harmonic contribution which takes care of the boundary conditions). In a sloppy way, we conclude that the discrete solution $\mathbf{G}_h^{(kl)}$ has a similar behavior as G which is essentially of the form

$$G((x_k, y_l), (x, y)) \sim \frac{1}{2\pi} \ln \sqrt{(x_k - x)^2 + (y_l - y)^2}.$$

In particular, we expect that $G_{ij}^{(kl)}$ is, for most (ij), different from zero which implies that $\mathcal{L}_h^{-1}$ is a full matrix. Based on the discrete Green's function, we can write the solution of $\mathcal{L}_h \mathbf{u}_h = \mathbf{d}_h$ in a characteristic way. Assuming that the boundary values are zero, we notice that the data vector $\mathbf{d}_h$ can be written as a discrete convolution with $\delta_h^{(kl)}$ over Ω_h

$$d_{ij} = \sum_{(x_k,y_l)\in\Omega_h} d_{kl} \delta_{ij}^{(kl)} h^2$$

which corresponds to

$$f(x) = \int_\Omega f(y) \delta(x - y) \, dy$$

(note that $g \equiv 0$ are the boundary values). Then, the solution $\mathbf{u}_h = \mathcal{L}_h^{-1} \mathbf{d}_h$ satisfies

$$u_{ij} = \sum_{(x_k,y_l)\in\Omega_h} d_{kl} \left(\mathcal{L}_h^{-1} \delta_h^{(kl)} \right)_{ij} h^2 = \sum_{(x_k,y_l)\in\Omega_h} d_{kl} G_{ij}^{(kl)} h^2.$$

In Section 2.8 we will see that the counterpart of this relation is the representation formula

$$u(x) = \int_\Omega G(x,y) f(y)\, dy$$

for the solution u of $-\Delta u = f$ in Ω, $u = 0$ on $\partial\Omega$ in terms of the Green's function G.

After this inter-mezzo on the similarities between Δ and Δ_h, we continue with the general 2–D Dirichlet problem on $\Omega = (0,1)^2$, i.e.

$$Lu = -(a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy}) + b_1 u_x + b_2 u_y + cu$$

based on the positive definite, symmetric matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$$

In particular, we have $a_{11}, a_{22} > 0$. In the special case $a_{12} = 0$, the operator L does not contain mixed derivatives and we find the splitting

$$Lu = -a_{11}u_{xx} + b_1 u_x + \frac{c}{2}u - a_{22}u_{yy} + b_2 u_y + \frac{c}{2}u = (L^x + L^y)u$$

into two 1–D operators:

1) for fixed y_j, we can write

$$L^x u(x_i, y_j) \approx l^x_{ij} u_{i-1,j} + d^x_{ij} u_{i,j} + r^x_{ij} u_{i+1,j}$$

2) for fixed x_i,

$$L^y u(x_i, y_j) \approx l^y_{ij} u_{i,j-1} + d^y_{ij} u_{i,j} + r^y_{ij} u_{i,j+1}$$

Hence, from the 1D case, we get a second order consistency, if $u \in C^4(\bar\Omega)$. Altogether, we obtain the numerical approximation in the form

$$(Lu)(x_i, y_j) \approx l^y_{ij} u_{i,j-1} + l^x_{ij} u_{i-1,j} + (d^x_{ij} + d^y_{ij}) u_{ij} + r^x_{ij} u_{i+1,j} + r^y_{ij} u_{i,j+1}$$

with corresponding stencil

$$\begin{bmatrix} & r^y_{ij} & \\ l^x_{ij} & d^x_{ij} + d^y_{ij} & r^x_{ij} \\ & l^y_{ij} & \end{bmatrix} \tag{2.20}$$

and

$$\begin{aligned} d^x_{ij} &= \frac{2a_{11}}{h^2} + c & \quad d^y_{ij} &= \frac{2a_{22}}{h^2} + c \\ l^x_{ij} &= -\frac{a_{11}}{h^2} - \frac{b_1}{2h} & \quad l^y_{ij} &= -\frac{a_{22}}{h^2} - \frac{b_2}{2h} \\ r^x_{ij} &= -\frac{a_{11}}{h^2} + \frac{b_1}{2h} & \quad r^y_{ij} &= -\frac{a_{22}}{h^2} + \frac{b_2}{2h} \end{aligned} \tag{2.21}$$

where all coefficients a_{11}, a_{22}, b_1, b_2 and c depend on the grid point (x_i, y_j), i.e. $a_{11} = a_{11}(x_i, y_j)$ etc.

Assuming enough regularity on the solutions $u = \mathcal{L}^{-1}\binom{f}{g}$ of the Dirichlet problem and of the auxiliary problem $v = \mathcal{L}^{-1}\binom{1}{1}$ we can mimic the 1D approach and obtain

Theorem 2.34. *Assume that $a_{12} = 0$ and that the above scheme is consistent in the solution of the auxiliary problem $v = \mathcal{L}^{-1}\binom{1}{1}$ If the Dirichlet problem has a solution u in $C^4(\bar{\Omega})$ then the above scheme is second order convergent.*

If we extend our considerations to the case $a_{12} \neq 0$, we need an approximation for the mixed derivatives $\partial^2 u/(\partial x \partial y)$. The simplest idea is, of course,

$$\begin{aligned} u_{xy}(x_i, y_j) &\approx \left(\frac{u_{i+1,j+1} - u_{i+1,j-1}}{2h} - \frac{u_{i-1,j+1} - u_{i-1,j-1}}{2h} \right) /2h \\ &= \frac{1}{4h^2}(u_{i+1,j+1} - u_{i+1,j-1} - u_{i-1,j+1} + u_{i-1,j-1}) \end{aligned}$$

which yields a second order approximation if u is smooth enough. After multiplying with a_{12}, the stencil is

$$\frac{a_{12}}{4h^2} \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{bmatrix} \tag{2.22}$$

Hence, the complete approximation of the elliptic operator L yields blocks with the structure

$$\begin{bmatrix} +m_{ij} & r^y_{ij} & -m_{ij} \\ l^x_{ij} & d^x_{ij} + d^y_{ij} & r^x_{ij} \\ -m_{ij} & l^y_{ij} & +m_{ij} \end{bmatrix}$$

where $l^{x,y}_{ij}, r^{x,y}_{ij}, d^{x,y}_{ij}$ are given in (2.21) and

$$m_{ij} = \frac{a_{12}(x_i, y_j)}{2h^2}$$

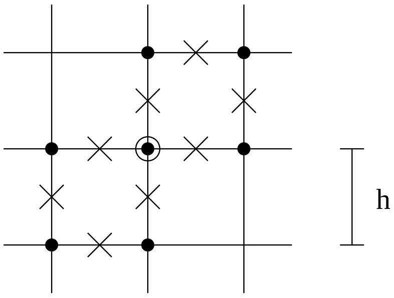
Unfortunately, the corresponding matrix $\mathcal{L}_h$ has lost the important positivity properties because we have alternating signs in the off–diagonals. Negative signs in the off–diagonals, however, are crucial to show stability so that our main objective is now to construct other approximations of $\frac{\partial^2 u}{\partial x \partial y}$ which lead to a good distribution of signs in the stencils. Since (x_i, y_j) is half way inbetween (x_{i+1}, y_{j+1}) and (x_{i-1}, y_{j-1}), the mixed derivative at (x_i, y_j) is (up to second order) just the average of the derivatives at the diagonal neighbors. Writing

(2.22) at the diagonal neighbors as stencil for (x_i, y_j), we get the approximation

$$
(2.23)\quad \frac{1}{8h^2}\begin{bmatrix} & & -1 & 0 & 1 \\ & & 0 & & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & & 0 & & \\ 0 & 0 & 0 & & \end{bmatrix} + \frac{1}{8h^2}\begin{bmatrix} & & 0 & 0 & 0 \\ & & 0 & & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & & 0 & & \\ 1 & 0 & -1 & & \end{bmatrix}
$$

$$
= \frac{1}{8h^2}\begin{bmatrix} & & -1 & 0 & 1 \\ & & 0 & & 0 \\ -1 & 0 & 2 & 0 & -1 \\ 0 & & 0 & & \\ 1 & 0 & -1 & & \end{bmatrix}.
$$

The disadvantage of this stencil is its size. Due to the structure of (2.23), however, it is possible to reduce the size considerably. The trick is to apply this approximation to a grid which is twice as fine as the original one. If we superimpose this fine grid, we introduce additional nodes (some of them are indicated by crosses in the figure below).



The evaluation of (2.23) at a node of the original grid (marked by a circle), does not involve the intermediate nodes due to the zeros in (2.23). Hence, we can transform (2.23) into a stencil on the original grid by setting

$$
\frac{1}{8\left(\frac{h}{2}\right)^2}\begin{bmatrix} 0 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 0 \end{bmatrix} = \frac{1}{2h^2}\begin{bmatrix} 0 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 0 \end{bmatrix}.
$$

Similarly, by taking the other diagonal neighbors, we get the alternative stencil

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

Please observe that for $a_{12} > 0$, the approximation of $-a_{12}\frac{\partial^2 u}{\partial x \partial y}$ given by

$$
\frac{a_{12}}{2h^2}\begin{bmatrix} 0 & 1 & -1 \\ 1 & -2 & 1 \\ -1 & 1 & 0 \end{bmatrix}
$$

is at least negative at those points where the operator without mixed derivatives (2.20) has zero entries. Under certain conditions on the coefficients, the wrong signs around the center and in the center itself can be compensated by the other participating weights. In the case $a_{12} \le 0$, we use the other approximation for the mixed derivative. With the standard abbreviations $a_{12}^- = \min\{a_{12}, 0\}$ and $a_{12}^+ = \max\{a_{12}, 0\}$ we can write the stencil for the general differential operator L in the following form

$$\frac{1}{h^2}\begin{bmatrix} a_{12}^- & |a_{12}| - a_{22} & -a_{12}^+ \\ |a_{12}| - a_{11} & 2(a_{11} + a_{22} - |a_{12}|) & |a_{12}| - a_{11} \\ -a_{12}^+ & |a_{12}| - a_{22} & a_{12}^- \end{bmatrix}$$
$$+ \frac{1}{2h}\begin{bmatrix} & b_2 & \\ -b_1 & & b_1 \\ & -b_2 & \end{bmatrix} + \begin{bmatrix} & 0 & \\ 0 & c & 0 \\ & 0 & \end{bmatrix}.$$

In order to get the correct signs in the corresponding matrix $\mathcal{L}_h$, we have to require that $|a_{12}| < \min\{a_{11}, a_{22}\}$. Then, for h sufficiently small, only the center weight is positive and all others are negative. More precisely, we have to require that

$$\frac{1}{h^2}\left(a_{kk} - |a_{12}|\right) + \frac{1}{2h}|b| \le 0, \qquad k = 1, 2$$

which holds for example if

$$h \le h_0 := \frac{2(\min\{a_{11}, a_{22}\} - |a_{12}|)}{\|b\|_\infty}. \tag{2.24}$$

If $\|b\|_\infty$ is very large compared to the entries of (a_{ij}) (i.e. convection dominates diffusion), h_0 can be very small. In this case, it is possible that the discretization has to be very fine in order to get solvability of $\mathcal{L}_h \mathbf{u}_h = \mathbf{d}_h$. For such problems, it can be advisable to use upwind discretizations for the convective terms. More precisely, if $b_1 > 0$, we choose the backward difference given by the stencil

$$\frac{b_1}{h}\begin{bmatrix} -1 & 1 & 0 \end{bmatrix}$$

and if $b_1 \le 0$ we take

$$\frac{b_1}{h}\begin{bmatrix} 0 & -1 & 1 \end{bmatrix}.$$

The stencils can be combined as

$$\frac{1}{h}\begin{bmatrix} -b_1^+ & |b_1| & b_1^- \end{bmatrix}$$

respectively for both directions

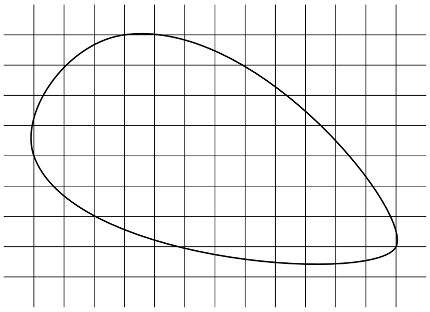
$$\frac{1}{h}\begin{bmatrix} & b_2^- & \\ -b_1^+ & |b_1| + |b_2| & b_1^- \\ & -b_2^+ & \end{bmatrix}.$$

The advantage is that all weights automatically have the right signs to derive invertibility of $\mathcal{L}_h$. However, the order of consistency is reduced by one compared to the discretization with centered differences.

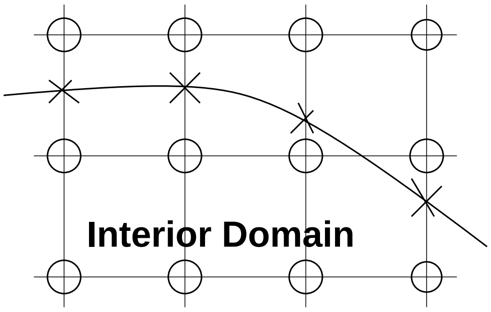
To show that $\mathcal{L}_h^{-1}$ is stable, we need to find again a vector $y \gg 0$ such that $\mathcal{L}_h y \gg 0$ and $\|y\|_{h,\infty} \leq C$, $(\mathcal{L}_h y)_i \geq \mu > 0$ with μ and C independent of h. In Exercise 2.14, such a vector is constructed for the general Dirichlet problem. We collect our results on general operators in

Lemma 2.35. *If the first order terms in the general operator L are approximated by upwind differencing, then the finite difference scheme is stable for sufficiently small h. If central differencing is used, h should satisfy* (2.24) *to obtain a stable scheme.*

Finally, we turn to the question how finite difference schemes for elliptic equations are applied on arbitrary domains $\Omega \subset \mathbb{R}^2$, which do not fit into the rectangular grid structure, like the domain given in the figure below.



If the computational domain Ω is not rectangular, we will directly run into problems, when discretizing the boundary conditions, even if the domain is smooth, because the boundary does not match with our spatial grid: for simplicity, let us consider Laplace's together with Dirichlet boundary conditions. For our standard scheme we need the values at "$\bigcirc$", having only those at "$\times$".



We may overcome this problem, if we use a non–equidistant grid close to the boundary, i.e. instead of using the points $\bigcirc$ outside of Ω, we use difference quotients which are based on the boundary points $\times$. If $\sigma_e, \sigma_w \in (0,1]$ denotes the distance (in h–units) to the east and west neighbors, we modify the standard difference quotient according to

$$\frac{\partial^2 u}{\partial x^2}(x,y)$$
$$\approx \left(\frac{u(x+\sigma_e h, y) - u(x,y)}{\sigma_e h} - \frac{u(x,y) - u(x-\sigma_w h, y)}{\sigma_w h} \right) \Big/ \left(\frac{\sigma_e h + \sigma_w h}{2} \right).$$

Using the same approximation for $\partial^2 u/\partial y^2$, we obtain the so called *Shortley–Weller approximation*

$$
\begin{aligned}
(2.25)\qquad -\Delta u(x,y) \approx \frac{1}{h^2}\Bigg\{ &\left(\frac{2}{\sigma_e\sigma_w}+\frac{2}{\sigma_n\sigma_s}\right)u(x,y)\\
&-\frac{2}{\sigma_e(\sigma_e+\sigma_w)}u(x+\sigma_e h,y)-\frac{2}{\sigma_w(\sigma_e+\sigma_w)}u(x-\sigma_w h,y)\\
&-\frac{2}{\sigma_n(\sigma_n+\sigma_s)}u(x,y+\sigma_n h)-\frac{2}{\sigma_n(\sigma_n+\sigma_s)}u(x,y-\sigma_s h)\Bigg\}
\end{aligned}
$$

We observe that the corresponding matrix $\mathcal{L}_h$ is typically not symmetric but the sign structure is correct (positive diagonal, negative off–diagonal entries). The crucial vector $y \gg 0$ (to get stability) can be constructed by discretizing the function

$$v(x,y)=\frac{1}{2}(x-x_0)(x_0+d-x)+1$$

where we assume that Ω is contained in the slab $\{(x,y)\,:\,x_0\le x\le x_0+d\}$. In points where all factors $\sigma_e,\sigma_w,\sigma_n,\sigma_s$ are equal to one, (2.25) reduces to the standard five point stencil which is second order accurate. (We collect all these points in $\Omega_h^{(2)}$.) The consistency is only first order in points where one of the factors is less than one (giving rise to $\Omega_h^{(1)}=\Omega_h\backslash\Omega_h^{(2)}$). Hence, it seems that the scheme is in general only first order convergent for smooth solutions. If, however, the grid is non–regular only at the boundary, one can show that the order of convergence is still two [**6**]. To explain the basic argument, we recall that the solution error $\mathbf{e}_h$ can be written in terms of the local discretization error $\mathbf{l}_h$ according to

$$\mathbf{e}_h=\mathcal{L}_h^{-1}\mathbf{l}_h. \tag{2.26}$$

We now split $\mathbf{l}_h$ into two contributions: $\mathbf{l}_h^{(2)}$ includes all nodes where (2.25) is actually the second order accurate five point stencil and $\mathbf{l}_h^{(1)}$ includes those nodes close to the boundary where one of the factors $\sigma_e,\sigma_w,\sigma_n,\sigma_s$ is different from one (i.e. $\mathbf{l}_h^{(2)}=0$ on $\Omega_h^{(1)}$ and $\mathbf{l}_h^{(1)}=0$ on $\Omega_h^{(2)}$). Consequently (if the solution of the original problem is smooth enough)

$$\|\mathbf{l}_h^{(1)}\|_{h,\infty}=\mathcal{O}(h),\qquad \|\mathbf{l}_h^{(2)}\|_{h,\infty}=\mathcal{O}(h^2).$$

Due to (2.26), we get a related splitting of $\mathbf{e}_h$ so that stability of the scheme implies

$$\|\mathbf{e}_h^{(2)}\|_{h,\infty}=\mathcal{O}(h^2).$$

For the remaining part $\mathbf{e}_h^{(1)}=\mathcal{L}_h^{-1}\mathbf{l}_h^{(1)}$ we have to apply a more careful estimate instead of just using stability $\|\mathcal{L}_h^{-1}\|_{h\infty}\le C$. We consider the discrete function $\mathbf{v}_h$ which is equal to Kh^3 in Ω_h and zero on Γ_h where $K>0$ is some constant.

Applying $\mathcal{L}_h$ to $\mathbf{v}_h$ we obtain a data vector $\tilde{\mathbf{d}}_h$ which is zero on Γ_h and $\Omega_h^{(2)}$. On $\Omega_h^{(1)}$, we find

$$\tilde{d}_{ij} \geq \tilde{K} K h$$

where $\tilde{K} > 0$ just depends on the factors σ_*. The reason is that in (2.25), the weight $2/\sigma_w\sigma_e + 2/\sigma_n\sigma_s$ is just the sum of all other weights but on $\Omega_h^{(1)}$, at least one of the neighboring values is zero (due to the zero boundary values of $\mathbf{v}_h$). Hence, if K is big enough, we find

$$-\tilde{d}_{ij} \leq l_{ij}^{(1)} \leq \tilde{d}_{ij}.$$

(A simple but tedious analysis of the error terms shows that $K = \|u\|_{C^3(\bar{\Omega})}/3$ is sufficient [**6**].) Due to the positivity of $\mathcal{L}_h^{-1}$, we get

$$-\mathbf{v}_h \leq -\mathcal{L}_h^{-1}\tilde{\mathbf{d}}_h \leq \mathcal{L}_h^{-1}\mathbf{l}_h^{(1)} = \mathbf{e}_h^{(1)} = \mathcal{L}_h^{-1}\mathbf{l}_h^{(1)} \leq \mathcal{L}_h^{-1}\tilde{\mathbf{d}}_h = \mathbf{v}_h$$

so that $\|\mathbf{e}_h^{(1)}\|_{h,\infty} \leq \|\mathbf{v}_h\|_{h,\infty} \leq Kh^3$ and thus

$$\|\mathbf{e}_h\|_{h,\infty} \leq C^{(1)}h^3 + C^{(2)}h^2.$$

Altogether, we have

Lemma 2.36. *The finite difference scheme based on the Shortley–Weller approximation* (2.25) *is first order convergent (for solutions which are sufficiently smooth). If the grid is irregular only at the boundary, the order of convergence is two.*

Instead of changing the regular structure of the grid at the boundary, it is also possible to use interpolation of the boundary values to construct the values of the approximate solution on points close to the boundary. For more details in that direction we refer to [**6**].

Although we did not consider the theory of classical solutions for Neumann problems, we will nevertheless briefly indicate the numerical treatment. For simplicity, we consider

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega \\ \frac{\partial u}{\partial n} &= g \qquad \text{on } \partial\Omega \end{aligned}$$

where Ω is the unit square. Again, we can introduce a boundary operator $\tilde{B}$ which now operates on $C^1(\bar{\Omega})$ and maps functions into $C(\partial\Omega)$ according to the rule

$$\tilde{B}u = \left.\frac{\partial u}{\partial n}\right|_{\partial\Omega}.$$

Then, the operator formulation of the Neumann Poisson problem is obtained with $\tilde{\mathcal{L}} = \left(\begin{smallmatrix} -\Delta \\ \tilde{B} \end{smallmatrix}\right)$ in the form

$$\tilde{\mathcal{L}}u = \begin{pmatrix} f \\ g \end{pmatrix}.$$

In contrast to the Dirichlet problem, now also the boundary operator $\tilde{B}$ needs a discretization. If $\Omega_h \cup \Gamma_h$ is the regular grid on the unit square, we can use the standard five point discretization for the Laplacian in all nodes $(x_i, y_j) \in \Omega_h$. At points on Γ_h, however, the entries of $\mathbf{u}_h$ have to be determined from the boundary condition. A simple approach is to use half sided differences to approximate $\tilde{B}u$. For example, at a point (x_n, y_j) on the east boundary, the normal derivative is just an x–derivative and we set

$$\tilde{B}u(x_n, y_j) \approx \frac{u(x_n, y_j) - u(x_n - h, y_j)}{h}, \qquad j = 1, \ldots, n-1.$$

Thus, $\tilde{B}u = g$ can be used to express the values of $\mathbf{u}_h$ on Γ_h according to

$$u_{nj} = u_{n-1,j} + h g_{nj}, \qquad j = 1, \ldots, n-1.$$

The corresponding matrix $\tilde{\mathcal{L}}_h$ which results from this discretization is investigated in Exercise 2.15. In particular, we see that the strategy to show stability cannot be taken over form the Dirichlet case (see [**6**] for a stability proof).

Another approximation method is based on central differences for the Neumann condition. In this case, we use the discrete Laplacian also on boundary nodes. The missing neighbor outside the domain is then determined with the Neumann condition which is approximated by a central difference stencil at the boundary point. For a point (x_n, y_j) on the east face this yields

$$\tilde{B}u(x_n, y_j) \approx \frac{u(x_n + h, y_j) - u(x_n - h, y_j)}{2h}, \qquad j = 0, \ldots, n$$

so that the missing information for the Laplace stencil at (x_n, y_j) is obtained with

$$u_{n+1,j} = u_{n-1,j} + 2h g_{nj}, \qquad j = 1, \ldots, n-1.$$