Analytical and Numerical Methods for Elliptic Partial Differential Equations

WS 1998/99

Dr. Michael Junk
Universität Kaiserslautern

Contents

CHAPTER 1

Introduction

1.1. Classification of PDEs – some aspects

Generally speaking, partial differential equations (PDEs) characterize functions by relating their partial derivatives. To denote partial derivatives in a concise way, we introduce multi–index notation. A multi–index is a vector of nonnegative integers $\alpha \in \mathbb{N}_0^n$. If $x \in \mathbb{R}^n$ is some vector, the power x^α abbreviates

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n}.$$

Applying this rule formally to the gradient operator

$$\nabla_x = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n} \right)^T$$

we obtain the general differential operator (the variable x as subscript can be omitted in unambiguous cases)

$$\nabla_x^\alpha = \frac{\partial^{\alpha_1}}{\partial {x_1}^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial {x_n}^{\alpha_n}} = \frac{\partial^{|\alpha|}}{\partial {x_1}^{\alpha_1} \dots \partial {x_n}^{\alpha_n}}$$

Here, we have set

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n.$$

By definition, the zero index gives rise to the identity operator $\nabla^{(0,\dots,0)} = \mathrm{id}$.

Using this notation we can now specify PDEs in the following way: Let $\Omega \subset \mathbb{R}^n$ be nonempty and open and let $\alpha^{(1)}, \dots, \alpha^{(M)}$ be a list of multi–indices. Given a function $F : \Omega \times \mathbb{R}^M \mapsto \mathbb{R}$, the expression

$$F(x, \nabla^{\alpha^{(1)}} u(x), \dots, \nabla^{\alpha^{(M)}} u(x)) = 0, \qquad x \in \Omega \tag{1.1}$$

is called *partial differential equation* for the function $u : \Omega \mapsto \mathbb{R}$. (Observe: the expression is an *ordinary differential equation* if x is a scalar variable, i.e. $n = 1$).

If u is vector valued, say in $\mathbb{R}^d$, the domain of F has to be changed to $\Omega \times (\mathbb{R}^d)^M$. Obviously, a system of PDEs is obtained if F is vector valued.

Our first classification of problem (1.1) is based on the linearity of the function F. If $F(x, v_1, \dots, v_M)$ is nonlinear in the v–variables, (1.1) is called a nonlinear PDE. If F is linear in v, we obtain a linear PDE which can be written in the form

$$L(x,\nabla)u = \sum_{|\alpha|\le m} a_\alpha(x)\nabla^\alpha u = 0. \tag{1.2}$$

Note that the coefficients $a_\alpha(x)$ need not be linear. By introducing zero coefficients, it is always possible to write the equation as sum over all multi-indices up to the highest appearing order.

Inbetween the fully linear and the nonlinear case one often introduces quasi–linear cases. They are characterized by F being linear at least in the variables corresponding to the highest u–derivatives.

Apart from the linearity of F, the choice of the multi-indices $\alpha^{(i)}$ also characterizes equation (1.1). In particular, the *order* of the equation

$$m = \max\{\, |\alpha^{(i)}| \;:\; i = 1,\dots,M \,\}$$

generally influences the number of boundary conditions that have to be provided to make (1.1) uniquely solvable.

It turns out, however, that this crude classification with respect to linearity and order is not sufficient to describe the qualitative behavior of solutions to PDEs. Already within the class of second order, linear equations with constant coefficients, solutions differ considerably for example in their smoothness properties. A more refined classification scheme should identify classes of equations with common properties such that the type of an equation determines the nature of

- boundary and initial conditions which may be imposed,
- singularities which solutions may have,
- methods which can be used to approximate a solution.

In the following we briefly introduce the criterion which singles out *elliptic* equations.

Let us start with a second–order PDE in n space dimensions,

$$Lu = \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = 0.$$

It turns out, that the operator L is effectively characterized by algebraic properties of the matrix $A(x) = (a_{ij}(x))$ (we can assume w.l.o.g. that $A(x)$ is symmetric). In particular, L (respectively $Lu = 0$) is called elliptic in $x \in \Omega$ if all eigenvalues of $A(x)$ have the same sign. In other words, ellipticity of L is related to definiteness of the matrix $A(x)$. (Observe that the curve $\{\, \xi \in \mathbb{R}^2 \;:\; \xi^T A\xi = 1 \,\}$ describes an *ellipse* if $A \in \mathbb{R}^{2\times 2}$ is positive definite.)

To generalize this characterization to equations of higher order we take a slightly different point of view. Given the general operator (1.2) of order m we define its *principal part* as the sum of the m^{th} order derivatives

$$L^p(x,\nabla) = \sum_{|\alpha|=m} a_\alpha(x)\nabla^\alpha.$$

The *symbol* of the principal part is then obtained by replacing ∇ with $i\xi$ where $\xi \in \mathbb{R}^n$ is a variable

$$L^p(x, i\xi) = \sum_{|\alpha|=m} i^m a_\alpha(x) \xi^\alpha, \qquad \xi \in \mathbb{R}^n.$$

For the second order operator we get in particular

$$L^p(x, i\xi) = -\xi^T A(x) \xi$$

which is just the quadratic form determined by the symmetric matrix $A(x)$. In accordance with the previous definition, we call the general operator $L(x, \nabla)$ elliptic in x iff

$$L^p(x, i\xi) = 0 \quad \Leftrightarrow \quad \xi = 0.$$

Finally, L is called elliptic if it is elliptic at each point $x \in \Omega$.

The definition of ellipticity for systems of equations is slightly more complicated and we refer to [**14**] for more details. The classification of nonlinear equations is typically reduced to the linear case by going over to the linearized equation.

In the remaining chapters we will see the importance of the ellipticity condition for both theoretical and numerical investigations. It gives rise to typical methods and thus really characterizes a coherent group of partial differential equations.

1.2. Examples of elliptic PDEs

The Laplacian

$$\Delta := \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$$

is a paradigmatic elliptic operator (ellipticity is easily checked since the symbol of the principal part is just $\xi \mapsto -|\xi|^2$, which obviously satisfies the required condition). Using a suitable transformation of variables, the principal part of any second order linear elliptic operator with constant coefficients turns into a Laplacian. In this sense, the Laplacian is a *normal form* for these operators and its investigation yields results which apply to the whole family.

In the following section, we present some exemplary physical problems where the Laplacian plays an important role. After that, we list some classical properties of harmonic functions, i.e. solutions of the equation $\Delta u = 0$.

1.2.1. Examples arising in physics. If we consider Maxwell's equations for the electric field E and the magnetic field B in vacuum ($x \in \mathbb{R}^3$)

$$\begin{aligned} \operatorname{curl} E &= -\frac{\partial}{\partial t} B, & \operatorname{div} E &= \rho, \\ \operatorname{curl} B &= \frac{\partial}{\partial t} E + j, & \operatorname{div} B &= 0, \end{aligned}$$

we find in the steady case (i.e vanishing time derivatives)

$$\operatorname{curl} E = 0, \quad \operatorname{div} E = \rho, \qquad x \in \mathbb{R}^3.$$

Here, the charge density ρ is the source of the electric field. Since $\operatorname{curl} E = 0$, we can introduce an electric potential Φ such that $E = \nabla\Phi$. Inserting this relation in the divergence condition yields the so called *Poisson equation*

$$\Delta\Phi = \rho.$$

In the special case $\rho = 0$, we obtain Laplace's equation

$$\Delta\Phi = 0$$

for the electric potential. We stress that both equations are obtained in a steady situation. In fact, elliptic equations typically appear as time independent problems (although time independence does not imply ellipticity in general).

Another source for elliptic equations are conservation laws, i.e equations of the form

$$\frac{\partial u}{\partial t} + \operatorname{div} q(u) = f.$$

In steady state, the equation turns into

$$\operatorname{div} q(u) = f$$

and it is elliptic if the flux $q(u)$ is proportional to the gradient of u, i.e. $q(u) = \lambda\nabla u$. If $\lambda \neq 0$ is a constant, we get again Poisson equation

$$\lambda\Delta u = f.$$

As example, we mention heat transfer problems. Here, u is the heat density which can be expressed in terms of temperature T, specific heat c and density ρ of the material

$$u = c\rho T.$$

According to Fourier's law, the heat flux vector q points in the opposite direction of the temperature gradient

$$q = -k\nabla T$$

where k denotes the heat conductivity. If there are no heat sources ($f = 0$), the conservation law for u yields in the steady case

$$\operatorname{div}(k\nabla T) = 0.$$

If k is a constant, this is just Laplace's equation. If k is a differentiable function of position, we get by product rule

$$\operatorname{div}(k\nabla T) = k\Delta T + \nabla T \cdot \nabla k.$$

Again, the principal part of this operator is $k\Delta$, which is elliptic if $k \neq 0$ everywhere.

A similar situation occurs in diffusion processes. If c is the concentration of some material contained in a medium, the diffusion of c is described by Fick's law which relates the mass flux q of the concentration c by the relation

$$q = -D\nabla c.$$

Due to conservation of mass, c obeys a conservation law, so that in steady state

$$\operatorname{div}(D\nabla c) = 0.$$

Our everyday experience with heat conduction or diffusion processes (switching on a heater, putting milk in coffee, etc.) already gives us some insight into the qualitative behavior of solutions to Laplace's equation. In particular, a room which has been heated for a long time (steady state) does not show jumps in temperature – instead, the temperature distribution is smooth. The same is true in diffusion processes where jumps in concentration are smeared out so that in steady state the concentration varies smoothly.

1.2.2. Some properties of harmonic functions. In this section we present properties of solutions to Laplace's equation $\Delta u = 0$.

Definition 1.1. *Let $\Omega \subset \mathbb{R}^n$ be open. Any function $u \in C^2(\Omega)$ which satisfies $\Delta u = 0$ in Ω is called harmonic.*

Simple examples of harmonic functions are affine linear maps. A more instructive example is obtained with the constraint that the harmonic function should be radially symmetric, i.e. $u(x) = \bar{g}(|x|)$. Using the derivatives of $r(x) = |x|$

$$\frac{\partial r}{\partial x_i} = \frac{x_i}{r}, \qquad \frac{\partial^2 r}{\partial x_i \partial x_j} = \frac{1}{r}\left(\delta_{ij} - \frac{x_i x_j}{r^2}\right)$$

this implies

$$\frac{d^2\bar{g}}{dr^2} + \frac{n-1}{r}\frac{d\bar{g}}{dr} = 0.$$

The general solution of this equation is given by

$$\bar{g}(r) = c_1 + c_2 g(r), \tag{1.3}$$

where

$$g(r) = \begin{cases} \frac{1}{\omega_1} r & n = 1 \\ \frac{1}{\omega_2} \ln r & n = 2 \\ \frac{1}{(2-n)\omega_n} r^{2-n} & n > 2. \end{cases} \tag{1.4}$$

Although g has a different structure in different dimensions, the derivative is always of the form

$$g'(r) = \frac{1}{\omega_n}\frac{1}{r^{n-1}}.$$

The scaling constant ω_n is chosen in view of later applications. It represent the surface measure of the unit sphere S^{n-1} in $\mathbb{R}^n$ which is in terms of the gamma function

$$\omega_n = \frac{2\Gamma\left(\frac{1}{2}\right)^n}{\Gamma\left(\frac{n}{2}\right)}, \quad \omega_1 = 2, \quad \omega_2 = 2\pi, \quad \omega_3 = 4\pi.$$

Note that in the one dimensional case the unit ball is just the interval $(-1, 1)$ so that the unit sphere S^0 consists of the two points $x = \pm 1$. If we use the surface measure

$$\sigma = \delta_{-1} + \delta_{+1},$$

surface integrals reduce to sums

$$\int_{S^0} f(x)\, d\sigma = f(-1) + f(+1).$$

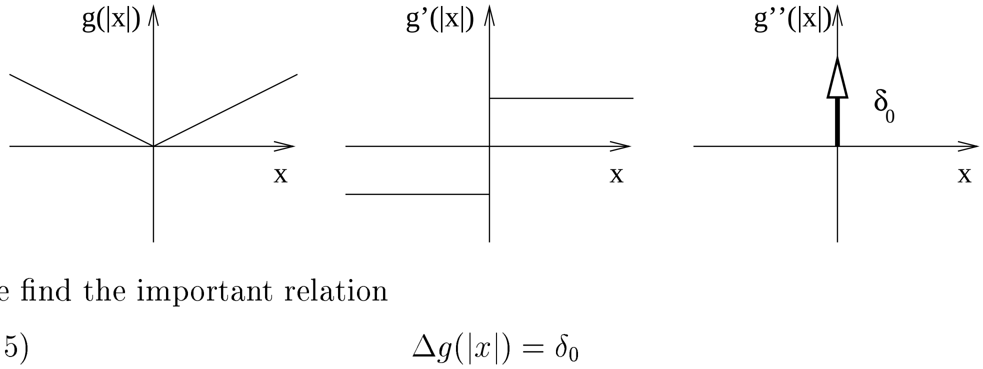
In particular, the surface volume ω_1 is given by

$$\omega_1 = \int_{S^0} d\sigma = 2.$$

We also remark that Gauss' theorem reduces to the fundamental theorem of calculus if we adopt the above convention for the surface measure. More precisely, let $\Omega = (a, b)$ be some interval with boundary $\partial\Omega = \{a, b\}$ and surface measure $\sigma = \delta_a + \delta_b$. Since the divergence of a function $f \in C^1(\bar{\Omega})$ is just the derivative $f'(x)$ and the outward normal field on $\partial\Omega$ is given by $n(a) = -1$ and $n(b) = +1$, we find

$$\int_a^b f'(x)\, dx = \int_\Omega \operatorname{div} f(x)\, dx = \int_{\partial\Omega} f(x) n(x)\, d\sigma = f(b) - f(a).$$

Going back to (1.4), we stress that $g(|x|)$ is not harmonic in $\mathbb{R}^n$ due to the singularity in $x = 0$. However, away from the origin, g is C^∞ smooth and satisfies $\Delta g(|x|) = 0$. The behavior of g at the singularity is most easily seen in the one dimensional case.



We find the important relation

$$\Delta g(|x|) = \delta_0 \tag{1.5}$$

which will be proved rigorously in Lemma 1.4 with the help of Green's formula. These formulas play an important role in the investigation of harmonic functions. They will also be frequently used in later chapters.

Lemma 1.2. *Let $\Omega \subset \mathbb{R}^n$ be a bounded open set with a piecewise C^1–boundary whose unit outward normal field is n. For $u, v \in C^1(\bar{\Omega}) \cap C^2(\Omega)$, Green's formula holds*

$$\int_\Omega u\Delta v - v\Delta u \, dx = \int_{\partial\Omega} u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\, d\sigma.$$

Proof: Using the identity

$$\operatorname{div}(u\nabla v) = u\Delta v + \nabla u \cdot \nabla v$$

and Gauss' theorem, we get immediately

$$\int_\Omega u\Delta v\, dx = -\int_\Omega \nabla u \cdot \nabla v\, dx + \int_{\partial\Omega} u\frac{\partial v}{\partial n}\, d\sigma.$$

Changing the role of u and v and subtracting the formulas leads to Green's formula since the symmetric term $\int_\Omega \nabla u \cdot \nabla v\, dx$ disappears. ■

As immediate consequence we obtain

Corollary 1.3. *Let $\Omega \subset \mathbb{R}^n$ be a bounded open set with a piecewise C^1–boundary whose unit outward normal field is n. If $u \in C^1(\bar{\Omega}) \cap C^2(\Omega)$ is harmonic then*

$$\int_{\partial\Omega} \frac{\partial u}{\partial n}\, d\sigma = 0.$$

Proof: Just set $v \equiv 1$ in Green's formula. ■

With the help of Green's formula, we can prove an auxiliary lemma which allows us to deduce another property of harmonic functions. The lemma is a reformulation of (1.5) which does not refer explicitly to the Dirac delta distribution.

Lemma 1.4. *Let $\Omega \subset \mathbb{R}^n$ be a bounded open set with a piecewise C^1–boundary. If $u \in C^1(\bar{\Omega}) \cap C^2(\Omega)$ then for any $x \in \Omega$*

$$u(x) = \int_{\partial\Omega} u(y)\frac{\partial}{\partial n_y}G(x,y) - G(x,y)\frac{\partial}{\partial n}u(y)\, d\sigma + \int_\Omega G(x,y)\Delta u(y)\, dy$$

where

$$G(x,y) = g(|x-y|) = \begin{cases} \frac{1}{\omega_1}|x-y| & n = 1 \\ \frac{1}{\omega_2}\ln|x-y| & n = 2 \\ \frac{1}{(2-n)\omega_n}|x-y|^{2-n} & n > 2. \end{cases}$$

Proof: The basic idea of the proof is as follows (note that this is just a formal argument): Granting that $x \mapsto G(x,0) = g(|x|)$ solves

$$\Delta G(x,0) = \delta_0(x),$$

we get

$$\Delta_y G(x,y) = \Delta_y g(|x-y|) = \delta_0(y-x)$$

so that

$$u(x) = \int_\Omega u(y)\text{"}\Delta_y G(x,y)\text{"}\,dy.$$

Applying Green's formula (only formally (!) since G is not in $C^2(\Omega)$) we get

$$u(x) = \int_{\partial\Omega} u(y)\frac{\partial}{\partial n_y}G(x,y) - G(x,y)\frac{\partial}{\partial n}u(y)\,d\sigma + \int_\Omega G(x,y)\Delta u(y)\,dy$$

which is the desired equality. A rigorous proof of the statement essentially avoids direct integration of the singularity of $\Delta_y G(x,y)$ at $y = x$. This is achieved by cutting out a ball B_ϵ of radius $\epsilon > 0$ around $x \in \Omega$. Since $y \mapsto G(x,y)$ is obviously contained in $C^1(\bar{\Omega}_\epsilon) \cap C^2(\Omega_\epsilon)$ with $\Omega_\epsilon = \Omega\backslash\bar{B}_\epsilon$, we can apply Green's formula in the modified domain Ω_ϵ

$$\begin{aligned}\int_{\Omega_\epsilon} & u(y)\Delta_y G(x,y) - G(x,y)\Delta u(y)\,dy \\ &= \int_{\partial\Omega_\epsilon} u(y)\frac{\partial}{\partial n_y}G(x,y) - G(x,y)\frac{\partial}{\partial n}u(y)\,d\sigma.\end{aligned}$$

Splitting $\partial\Omega_\epsilon$ into $\partial\Omega$ and ∂B_ϵ and using $\Delta_y G(x,y) = 0$ in Ω_ϵ we get

$$\begin{aligned}\int_{\partial B_\epsilon} & u(y)\frac{\partial}{\partial n_y}G(x,y) - G(x,y)\frac{\partial}{\partial n}u(y)\,d\sigma \\ &= \int_{\partial\Omega} u(y)\frac{\partial}{\partial n_y}G(x,y) - G(x,y)\frac{\partial}{\partial n}u(y)\,d\sigma + \int_{\Omega_\epsilon} G(x,y)\Delta u(y)\,dy.\end{aligned}$$

Observe that the outer normal of $\partial\Omega_\epsilon$ on the boundary ∂B_ϵ points into the ball B_ϵ. By changing the sign of the integral, however, we have used $\frac{\partial}{\partial n}$ to denote the derivative with respect to the outer normal of ∂B_ϵ. Using the definition of $G(x,y)$, we can easily calculate the normal derivative

$$\begin{aligned}\frac{\partial}{\partial n_y} & G(x,y) = n(y)\cdot\nabla_y G(x,y) \\ &= \frac{y-x}{|y-x|}\cdot\left(g'(|x-y|)\frac{y-x}{|y-x|}\right) = \frac{1}{\omega_n}\frac{1}{|x-y|^{n-1}}.\end{aligned}$$

Due to continuity of u we can thus infer

$$\begin{aligned}\int_{\partial B_\epsilon} & u(y)\frac{\partial}{\partial n_y}G(x,y)\,d\sigma \\ &= \frac{1}{\omega_n}\frac{1}{\epsilon^{n-1}}\int_{\partial B_\epsilon} u(y)\,d\sigma = \frac{1}{|\partial B_\epsilon|}\int_{\partial B_\epsilon} u(y)\,d\sigma \underset{\epsilon\to 0}{\longrightarrow} u(x).\end{aligned}$$

Since $y \mapsto G(x,y)$ is locally integrable in $\mathbb{R}^n$ (the singularity in x is *weak*), we also get

$$\int_{\Omega_\epsilon} G(x,y)\Delta u(y)\,dy \underset{\epsilon\to 0}{\longrightarrow} \int_\Omega G(x,y)\Delta u(y)\,dy$$

(by Lebesgue's dominated convergence theorem). Since the singularity of G is even integrable on the sphere in $\mathbb{R}^n$, the remaining integral also vanishes. Transformation to the unit sphere S^{n-1} yields

$$\int_{\partial B_\epsilon} G(x,y)\frac{\partial}{\partial n}u(y)\,d\sigma = g(\epsilon)\epsilon^{n-1}\frac{1}{\omega_n}\int_{S^{n-1}}\frac{\partial}{\partial n_\xi}u(x+\epsilon\xi)\,d\sigma.$$

While the integral is bounded, the leading factor tends to zero as $\epsilon \to 0$. ■

An important conclusion which we draw from Lemma 1.4 is

Theorem 1.5. (Mean Value Property)
Let u be harmonic in $\Omega \subset \mathbb{R}^N$ and let $\overline{B_R(x_0)} \subset \Omega$ be a ball around some point $x_0 \in \Omega$. Then $u(x_0)$ is equal to its spherical mean

$$u(x_0) = \frac{1}{|\partial B_R(x_0)|}\int_{\partial B_R(x_0)} u\,d\sigma = \frac{1}{\omega_n}\int_{S^{n-1}} u(x_0+R\xi)\,d\sigma(\xi)$$

and also to its mean value in $B_R(x_0)$

$$u(x_0) = \frac{1}{|B_R(x_0)|}\int_{B_R(x_0)} u\,dx = \frac{n}{\omega_n}\int_{B_1(0)} u(x_0+Ry)\,dy.$$

Proof: Since u is harmonic, we obtain from Lemma 1.4

$$u(x_0) = \int_{\partial B_R(x_0)} u(y)\frac{\partial}{\partial n_y}g(|x_0-y|) - g(|x_0-y|)\frac{\partial}{\partial n}u(y)\,d\sigma(y).$$

On the surface of the ball $B_R(x_0)$ we have $|x_0-y| = R$ so that with Corollary 1.3

$$\int_{\partial B_R(x_0)} g(R)\frac{\partial}{\partial n}u(y)\,d\sigma(y) = 0.$$

The normal derivative $\frac{\partial}{\partial n_y}g(|x_0-y|)$ is just

$$\frac{\partial}{\partial n_y}g(|x_0-y|) = g'(R)\frac{y-x_0}{|y-x_0|}\cdot\frac{y-x_0}{|y-x_0|} = g'(R) = \frac{1}{\omega_n R^{n-1}}.$$

Since $\omega_n R^{n-1}$ is the surface volume of $\partial B_R(x_0)$ we obtain

$$u(x_0) = \frac{1}{|\partial B_R(x_0)|}\int_{\partial B_R(x_0)} u\,d\sigma.$$

Finally, a transformation of variables yields

$$u(x_0) = \frac{1}{\omega_n}\int_{S^{n-1}} u(x_0+R\xi)\,d\sigma(\xi).$$

Since there has been no restriction on the radius R other than the ball being contained in Ω, we also get for any $0 < \rho < 1$

$$u(x_0) = \frac{1}{\omega_n}\int_{S^{n-1}} u(x_0+\rho R\xi)\,d\sigma(\xi). \tag{1.6}$$

Multiplying (1.6) by ρ^{n-1} and integrating over $(0,1)$ leads to

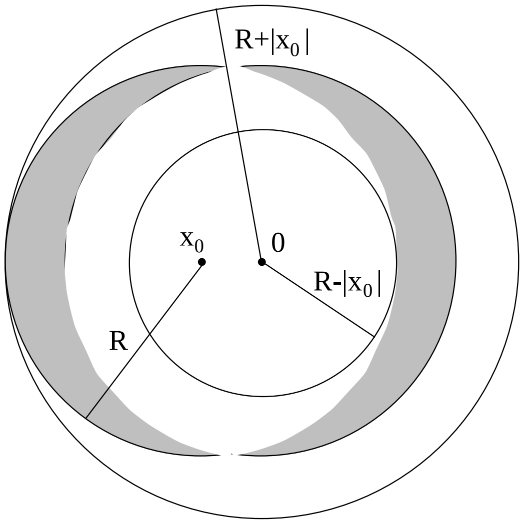
$$\frac{1}{n}u(x_0) = \frac{1}{\omega_n}\int_{B_1(0)} u(x_0+Ry)\,dy.$$

■

Applying the mean value theorem to $\Omega = \mathbb{R}^n$ we obtain the interesting result that the only bounded harmonic functions on the whole space are the constant functions. To see this, we compare the values of a harmonic function at two points, say at $x = x_0$ and the origin $x = 0$. Choosing a ball of very large radius R around these two points, we have

$$u(0) = \frac{1}{|B_R(0)|}\int_{B_R(0)} u(x)\,dx \qquad \text{and} \qquad u(x_0) = \frac{1}{|B_R(0)|}\int_{B_R(x_0)} u(x)\,dx.$$

If $R > |x_0|$, the difference Σ in the domains of integration has a shape as depicted in the figure below (the shaded part).



Clearly, the volume on which the domains differ is confined inbetween the balls $B_{R-|x_0|}(0)$ and $B_{R+|x_0|}(0)$. It can be estimated by

$$|\Sigma| \leq |B_{R+|x_0|}(0)\backslash B_{R-|x_0|}(0)| = \frac{\omega_n}{n}\left(|R+|x_0||^n - |R-|x_0||^n\right) = \mathcal{O}(R^{n-1}).$$

Taking into account the scaling by $|B_R(0)| \sim R^n$ and the boundedness of u, we get

$$|u(x_0) - u(0)| \leq \frac{1}{|B_R(0)|}\int_\Sigma |u(x)|\,dx = \mathcal{O}\left(\frac{1}{R}\right).$$

In the limit $R \to \infty$ we find $u(x_0) = u(0)$ for any $x_0 \in \mathbb{R}^n$. The result is summarized in the following theorem.

Theorem 1.6. *Let u be harmonic on $\mathbb{R}^n$ and bounded. Then u is constant.*

The next conclusion which we draw from the mean value theorem is

Theorem 1.7. (Maximum Principle)
Let u be harmonic on the open, connected set $\Omega \subset \mathbb{R}^n$ with $\sup_{x\in\Omega} u(x) = A < \infty$. Then either $u(x) < A$ for all $x \in \Omega$ or $u(x) = A$ on Ω.

Proof: The set $\Lambda := \{x \in \Omega : u(x) = A\}$ is relatively closed in Ω since u is continuous. Our aim is to show that Λ is at the same time open. Since Ω is assumed to be connected, this implies that $\Lambda = \Omega$ or $\Lambda = \emptyset$ and thus yields the result. To show that Λ is open, we pick $x_0 \in \Lambda$. For $R > 0$ sufficiently small, we get $B_R(x_0) \subset \Omega$ and by the mean value theorem

$$A = u(x_0) = \frac{1}{|B_R(x_0)|} \int_{B_R(x_0)} u(y)\, dy. \tag{1.7}$$

If, somewhere in the ball $B_R(x_0)$, $u(y)$ is less than the maximum A of u, then also the average is less than A so that (1.7) implies $B_R(x_0) \subset \Lambda$ which completes the proof. ■

In particular, if Ω is bounded and u is continuous on $\bar{\Omega}$, then u attains its maximum somewhere on the compact set $\bar{\Omega}$. According to the maximum principle, u is either constant or does not attain its maximum in Ω. Altogether, the maximum is attained on the boundary of Ω.

Corollary 1.8. (weak maximum principle) *If Ω is bounded and $u \in C^2(\Omega) \cap C(\bar{\Omega})$ is harmonic in Ω then the maximum and the minimum of u is attained on the boundary $\partial\Omega$.*

Proof: If u is harmonic, also $-u$ is harmonic and we find with the above argument that $\max_{x\in\bar{\Omega}}(-u(x))$ is attained at some point $x_0 \in \partial\Omega$. Since $-\max(-u) = \min u$, the function u is minimal in x_0 which shows the result concerning the minimum. We note that also Theorem 1.7 can be formulated for the minimum by going over from u to $-u$. ■

In Chapter 2 we will investigate maximum principles for more general elliptic operators. They turn out to be useful to show uniqueness of solutions as well as continuous dependence of the solution on the boundary data.

To complete this section, we will now prove an inverse version of the mean value theorem.

Theorem 1.9. *Let $\Omega \subset \mathbb{R}^n$ be open and $u \in C(\Omega)$. If for all $x_0 \in \Omega$ and $R > 0$ such that $\overline{B_R(x_0)} \subset \Omega$, the spherical mean value property*

$$u(x_0) = \frac{1}{|\partial B_R(x_0)|} \int_{\partial B_R(x_0)} u(x)\, d\sigma(x)$$

is satisfied, then u is harmonic and contained in $C^\infty(\Omega)$.

Proof: We pick a rotationally symmetric function $\phi(x) = \psi(|x|) \geq 0$ which is supported on the unit ball $B_1(0) \subset \mathbb{R}^n$ and is C^∞ smooth (a so called *mollifier*). The integral over ϕ is supposed to be normalized to one. After rescaling

$$\phi_\epsilon(x) := \frac{1}{\epsilon^n}\phi\left(\frac{x}{\epsilon}\right)$$

the latter condition remains true but the support shrinks to the ball $B_\epsilon(0)$. For some given $\epsilon > 0$ we now define the set Ω_ϵ as those points in Ω whose distance to the boundary is at least ϵ. On Ω_ϵ we can then convolute the given function u with ϕ_ϵ and obtain

$$\tilde{u}(x) := \int_{\mathbb{R}^n} u(y)\phi_\epsilon(x-y)\,dy, \qquad x \in \Omega_\epsilon.$$

Since $\phi_\epsilon \in C^\infty$, also $\tilde{u} \in C^\infty(\Omega_\epsilon)$. We will show that $u = \tilde{u}$ on Ω_ϵ if u satisfies the mean value property. Since ϵ can be chosen arbitrarily small, this implies $u \in C^\infty(\Omega)$. Taking $x \in \Omega_\epsilon$, we get

$$\begin{aligned}
\tilde{u}(x) &= \int_{\mathbb{R}^n} u(x-y)\phi_\epsilon(y)\,dy = \int_{B_1(0)} u(x-\epsilon z)\psi\left(|z|\right)\,dz \\
&= \int_0^1 \left(\int_{S^{n-1}} u(x-\epsilon r\xi)\,d\sigma(\xi)\right)\psi(r)r^{n-1}\,dr \\
&= \omega_n u(x)\int_0^1 \psi(r)r^{n-1}\,dr.
\end{aligned}$$

In the last step, the mean value property has been used. Doing all transformations backwards without u in the integral, we get

$$\tilde{u}(x) = u(x)\int_{\mathbb{R}^n}\phi_\epsilon(x)\,dx = u(x).$$

To show that u is also harmonic, we apply Green's formula with $v \equiv 1$ on the domain $B_R(x_0)$

$$\int_{B_R(x_0)} \Delta u\,dy = \int_{\partial B_R(x_0)} \frac{\partial u}{\partial n}\,d\sigma = \int_{S^{n-1}} \frac{\partial u}{\partial n}(x_0 + R\xi)\,d\sigma(\xi).$$

Since

$$\left.\frac{d}{dr}u(x_0 + r\xi)\right|_{r=R} = \nabla_x u(x_0 + R\xi)\cdot\xi = \frac{\partial u}{\partial n}(x_0 + R\xi)$$

we get again with the mean value property

$$\int_{B_R(x_0)} \Delta u\,dy = \left.\frac{d}{dr}\int_{S^{n-1}} u(x_0 + r\xi)\,d\sigma(\xi)\right|_{r=R} = \frac{d}{dr}(\omega_n u(x_0)) = 0.$$

After dividing by the volume of $B_R(x_0)$, we obtain in the limit $R \to 0$ that $\Delta u(x_0) = 0$ and hence $\Delta u = 0$ in Ω. ∎

Instead of requiring the spherical mean value property, the same result follows if we assume

$$u(x_0) = \frac{1}{|B_R(x_0)|} \int_{B_R(x_0)} u(y)\, dy.$$

Indeed, this implies the spherical mean value property (Exercise 1.8). Finally, we note that any harmonic function is automatically C^∞ smooth. Since harmonic functions have the mean value property (Theorem 1.5), the smoothness follows from Theorem 1.9. Also, the limit of a sequence of harmonic functions which converges uniformly on $\bar{\Omega}$ is again harmonic.

Corollary 1.10. *Let $(u_n)_{n\in\mathbb{N}}$ be a sequence of harmonic functions on Ω which converges locally uniformly on Ω. Then, the limit function is again harmonic.*

Proof: Locally, the limit u satisfies for any $x \in \Omega$ and $R > 0$ such that $\overline{B_R(x_0)} \subset \Omega$.

$$\begin{aligned} u(x) = \lim_{n\to\infty} u_n(x) = \lim_{n\to\infty} \frac{1}{|\partial B_R(x)|} \int_{\partial B_R(x)} u_n(y)\, d\sigma(y) \\ = \frac{1}{|\partial B_R(x)|} \int_{\partial B_R(x)} u(y)\, d\sigma(y) \end{aligned}$$

Hence, u has the mean value property and thus is harmonic according to Theorem 1.9. ■

One can even show that harmonic functions are analytic [**10**].

1.3. Elementary solution concepts

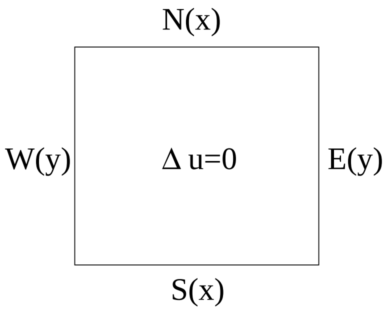
1.3.1. Separation of variables. The idea of this method is to reduce a PDE to a system of ODEs and expressing the solution of the PDE as a sum or infinite series. To illustrate the method, we consider Laplace's equation on the unit square

$$\Omega = \{ (x, y) \in \mathbb{R}^2 \ : \ 0 < x < 1,\, 0 < y < 1 \}.$$

We wish to find a function $u : \bar{\Omega} \mapsto \mathbb{R}$ satisfying

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 0 \tag{1.8}$$

at each point in Ω, having the values $N(x), E(y), S(x), W(y)$ at north, east, south and west face respectively. Moreover, the boundary values are assumed to be continuous including the corners, i.e. $W(1) = N(0)$, $N(1) = E(1)$, etc.



The key to separation of variables is to look for solutions of (1.8) of the form

$$u(x,y) = X(x)Y(y).$$

When we put a function of this form into (1.8), the partial derivatives in the differential equation appear as ordinary derivatives of the functions X and Y, i.e. (1.8) becomes

$$X''(x)Y(y) + X(x)Y''(y) = 0.$$

At any point (x,y) at which u is nonzero, we can divide this equation by u and rearrange to get

$$\frac{X''(x)}{X(x)} = -\frac{Y''(x)}{Y(x)}.$$

Since the right hand side of the equation does not depend on x, neither can the left hand side. Using the same argument for y, we find that both quotients are independent of x and y, i.e. they are constant on Ω. Denoting this constant λ, we obtain

$$X'' = \lambda X,$$
$$Y'' = -\lambda Y.$$

Solving these equations, we get the following four parameter family of solutions of (1.8)

$$u(x,y) = A\left(e^{\sqrt{\lambda}x} + Be^{-\sqrt{\lambda}x}\right)\left(e^{\sqrt{-\lambda}y} + Ce^{-\sqrt{-\lambda}y}\right).$$

The parameters now have to be chosen in such a way that the prescribed boundary conditions are met. Before we go into that, we simplify the problem by splitting it up into four sub problems of the same structure but with the assumption that only one of the functions N, E, S, W is not identically zero. If the corresponding solutions are denoted u_N, u_E, u_S, u_W, the overall solution is simply the sum

$$u = u_N + u_E + u_S + u_W$$

due to the linearity of the problem, To calculate u_N, we note that the zero boundary conditions on eastern, southern and western face are obtained by setting $\lambda = -(m\pi)^2$, $B = -1$, $C = -1$, i.e.

$$u_N(x,y) = A\sin(m\pi x)\sinh(m\pi y), \qquad m \in \mathbb{N}.$$

Now, if $N(x)$ is of the form

$$N(x) = \sum_{m=1}^{M} \alpha_m \sin(m\pi x)$$

we can simply take

$$A_m := \frac{\alpha_m}{\sinh(m\pi)}, \qquad m = 1, \dots, M$$

and define

$$u_N(x,y) = \sum_{m=1}^{M} A_m \sin(m\pi x) \sinh(m\pi y).$$

From the theory of Fourier series, we know that any function in $\mathbb{L}^2(0,1)$ can be expressed in that way if we allow $M = \infty$. However, the series converges in general not point wise which is most easily seen at the endpoints $x = 0$ and $x = 1$ (e.g. $N(x) \equiv 1$ can not be approximated in these points). On the other hand, one can show that the convergence is uniform if N is sufficiently smooth and vanishes at the endpoints. Note that we can assume without loss of generality that $N(0) = N(1) = 0$. The trick is to subtract a function which satisfies Laplace's equation and satisfies the boundary conditions in the corners. As example we mention

$$\hat{u}(x,y) = S(0)(1-x)(1-y) + S(1)x(1-y) + N(0)(1-x)y + N(1)xy.$$

Altogether, we can write

$$u(x,y) = \hat{u}(x,y) + \sum_{m=1}^{\infty} \big\{ \hat{N}_m \sin(m\pi x)\sinh(m\pi y) + \hat{E}_m \sinh(m\pi x)\sin(m\pi y)$$
$$+ \hat{S}_m \sin(m\pi x)\sinh(m\pi(1-y)) + \hat{W}_m \sinh(m\pi(1-x))\sin(m\pi y) \big\}$$

where

$$\hat{N}_m = \frac{2}{\sinh(m\pi)} \int_0^1 (N(x) - \hat{u}(x,1)) \sin(m\pi x)\, dx,$$
$$\hat{E}_m = \frac{2}{\sinh(m\pi)} \int_0^1 (E(y) - \hat{u}(1,y)) \sin(m\pi y)\, dy,$$
$$\hat{S}_m = \frac{2}{\sinh(m\pi)} \int_0^1 (S(x) - \hat{u}(x,0)) \sin(m\pi x)\, dx,$$
$$\hat{W}_m = \frac{2}{\sinh(m\pi)} \int_0^1 (W(y) - \hat{u}(0,y)) \sin(m\pi y)\, dy.$$

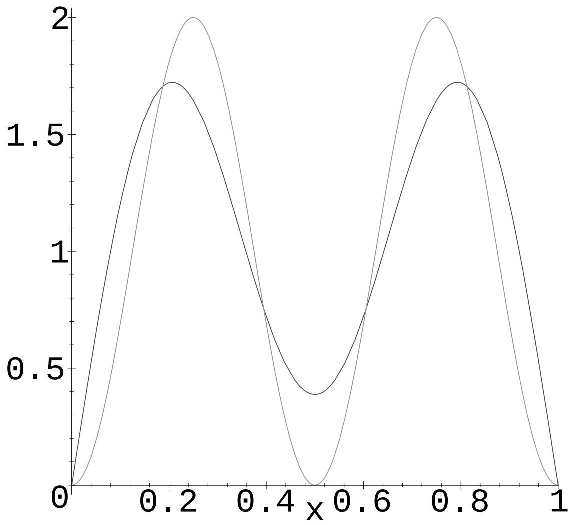
As example, we consider the case where N, E, S, W are restrictions of the function

$$b(x,y) = \cos(4\pi x)\cos(3\pi y)$$

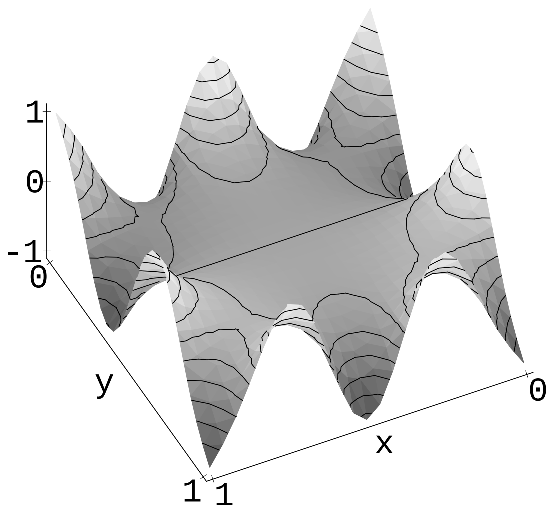
to the boundary. Truncating the Fourier series at $m = 3$, the approximation of the boundary values is still poor. After subtraction of

$$\hat{u}(x, y) = (1 - x)(1 - y) + x(1 - y) - (1 - x)y - xy$$

we get for example at the northern face



where the solid curve is the approximation. Truncating at $m = 5$ the graphs almost coincide and we get a good approximation of the solution of Laplace's equation.



1.3.2. Solution by integral operators. In order to find a solution of Poisson equation

$$\Delta\Phi(x) = \rho(x), \qquad x \in \mathbb{R}^n,\ n \geq 2, \tag{1.9}$$

we try the integral representation

$$\Phi(x) = \int_{\mathbb{R}^n} K(x, y)\rho(y)\, dy.$$

If such a function K exists (the so called *kernel* of the integral operator), it must be of the form

$$K(x, y) = k(|x - y|), \qquad x, y \in \mathbb{R}^n \tag{1.10}$$

with a suitable function k. This property follows from the fact that both the underlying space $\mathbb{R}^n$ and the Laplacian are invariant under translations and rotations. Translation invariance of Δ means nothing else but

$$\Delta_x\big(\Psi(x - z)\big) = (\Delta_x \Psi)(x - z), \qquad \forall x, z \in \mathbb{R}^n.$$

Now, if Φ solves (1.9) then $x \mapsto \Phi(x-z)$ solves (1.9) with respect to the shifted source $\rho(x - z)$. Consequently,

$$\Phi(x - z) = \int_{\mathbb{R}^n} K(x - z, y)\rho(y)\, dy = \int_{\mathbb{R}^n} K(x, y)\rho(y - z)\, dy,$$

so that

$$\int_{\mathbb{R}^n} \big(K(x - z, y) - K(x, y + z)\big)\rho(y)\, dy = 0.$$

Since ρ can be chosen arbitrarily from a sufficiently large class of functions (say $C_0^\infty(\mathbb{R}^n)$), we get

$$K(x, y + z) = K(x - z, y).$$

Setting $y = 0$, we finally obtain

$$K(x, z) = K(x - z, 0).$$

As far as rotational invariance is concerned, we note that for any rotation matrix $R \in SO(n)$ with Einstein's summation convention on repeated indices

$$\frac{\partial^2}{\partial x_i^2}\big(\Psi(Rx)\big) = \frac{\partial}{\partial x_i}\left(\frac{\partial \Psi}{\partial x_k}(Rx)R_{ki}\right) = \frac{\partial^2 \Psi}{\partial x_k \partial x_l}(Rx)R_{ki}R_{li} = \frac{\partial^2 \Psi}{\partial x_k^2}(Rx).$$

Again, if Φ solves (1.9) then $x \mapsto \Phi(Rx)$ is a solution with respect to the rotated source $\rho(Rx)$. With a similar argument as for the translation, we get

$$K(x, R^T y) = K(Rx, y).$$

Setting $y = 0$, we find for any $x \in \mathbb{R}^n$

$$K(x, 0) = K(|x|e_1, 0)$$

by choosing a rotation matrix which maps x onto the first unit vector e_1. A combination of translational and rotational invariance thus leads to (1.10) by setting

$$k(r) = K(re_1, 0).$$

To find out more about the structure of K, we formally replace the regular function $\rho(x)$ by a Dirac delta distribution $\delta_0(x)$. Due to our ansatz, the resulting function

$$\Phi(x) = \int_{\mathbb{R}^n} k(|x-y|)\text{"}\delta_0(y)\text{"}\, dy = k(|x|)$$

then solves

$$\Delta k(|x|) = \delta_0(x).$$

(All the above can be made precise in the framework of generalized functions which we introduce in the appendix.) The solution of this problem has already been determined in Section 1.2.2 where it is denoted $g(|x|)$. Instead of validating each step in the above derivation, we just take it as a motivation and check directly that the constructed integral operator yields a solution of (1.9).

Theorem 1.11. (Existence)
Let $\rho \in C_0^2(\mathbb{R}^n)$. Then

$$\Phi(x) = \int_{\mathbb{R}^n} G(x,y)\rho(y)\, dy$$

is a solution of the Poisson problem

$$\Delta\Phi(x) = \rho(x), \qquad x \in \mathbb{R}^n. \tag{1.11}$$

The function G is defined by

$$G(x,y) = g(|x-y|) = \begin{cases} \frac{1}{\omega_1}|x-y| & n=1 \\ \frac{1}{\omega_2}\ln|x-y| & n=2 \\ \frac{1}{(2-n)\omega_n}|x-y|^{2-n} & n>2 \end{cases}$$

Proof: First let us remark that the integral is well defined since ρ is compactly supported and bounded. Rewriting the expression as

$$\Phi(x) = \int_{\mathbb{R}^n} g(|z|)\rho(x-z)\, dz$$

we can obviously differentiate under the integral

$$\begin{aligned}\Delta\Phi(x) &= \int_{\mathbb{R}^n} g(|z|)\Delta_x\rho(x-z)\, dz \\ &= \int_{\mathbb{R}^n} g(|z|)(\Delta_z\rho)(x-z)\, dz = \int_{\mathbb{R}^n} G(x,y)\Delta\rho(y)\, dy.\end{aligned}$$

Since ρ is supported inside a ball $B_R(0)$ of sufficiently large radius $R>0$, we get

$$\Delta\Phi(x) = \int_{B_R(0)} G(x,y)\Delta\rho(y)\, dy.$$

Applying Lemma 1.4 then yields

$$\Delta\Phi(x) = \rho(x) - \int_{\partial B_R(0)} \rho(y)\frac{\partial}{\partial n_y}G(x,y) - G(x,y)\frac{\partial}{\partial n}\rho(y)\, d\sigma.$$

Now, both ρ and $\frac{\partial}{\partial n}\rho$ are zero on the boundary of B_R so that indeed $\Delta\Phi(x) = \rho(x)$. ■

We remark that the solution presented in Theorem 1.11 is not unique. Indeed, by adding any harmonic function Ψ, we get

$$\Delta(\Phi + \Psi)(x) = \rho(x), \qquad x \in \mathbb{R}^n.$$

Nevertheless, we get uniqueness under the additional condition that the solution of the Poisson problem has a certain behavior for large $|x|$. The solution in Theorem 1.11 satisfies in the case supp $\rho \subset B_R(0)$

$$|\Phi(x)| \leq \int_{\mathbb{R}^n} |g(|x-y|)|\, dy \|\rho\|_\infty$$

where $\|\cdot\|_\infty$ is the supremum norm on $\mathbb{R}^n$. Since

$$|x-y| = |x| \left| \frac{x}{|x|} - \frac{y}{|x|} \right| \leq \left(1 + \frac{R}{|x|}\right) |x|, \qquad \forall y \in B_R(0)$$

we get

$$\frac{g(|x-y|)}{g(|x|)} = \mathcal{O}(1), \qquad \text{for } |x| \to \infty$$

so that also

$$\left| \frac{\Phi(x)}{g(|x|)} \right| = \mathcal{O}(1), \qquad \text{for } |x| \to \infty.$$

If $n > 2$ this implies in particular that $\Phi(x) \to 0$ for $|x| \to \infty$.

Theorem 1.12. **(Uniqueness)** *For $n > 2$, the solution Φ from Theorem 1.11 is unique under the condition*

$$\Phi(x) \to 0, \qquad \textit{for } |x| \to \infty. \tag{1.12}$$

Proof: Assume Φ and $\tilde{\Phi}$ are two solutions of the Poisson problem which satisfy (1.12). Due to linearity, the difference $w = \Phi - \tilde{\Phi}$ is then a harmonic function ($\Delta w = 0$) which also satisfies (1.12). Consequently, there exists $R > 0$ such that $w(x)$ is bounded uniformly for $|x| > R$. If the support of the source ρ is contained in the ball $B_{\tilde{R}}(0)$, we get on $|x| \leq R$

$$|w(x)| = \|\rho\|_\infty \sup_{|x| \leq R} \int_{B_{\tilde{R}}(0)} G(x,y)\, dy < \infty$$

so that w is a bounded harmonic function on $\mathbb{R}^n$. According to Theorem 1.6, w is constant and since w vanishes for $|x| \to \infty$ this constant must be zero which yields uniqueness. ■

Another important property of the solution Φ given in Theorem 1.11 is its continuous dependence on ρ (the data of the problem).

Theorem 1.13. (Continuous dependence)
Let $\rho, \hat{\rho} \in C_0^2(\mathbb{R}^n)$ with supports contained in a ball of radius $R > 0$ around the origin. If $\|\cdot\|_\infty$ denotes the supremum norm in $\mathbb{R}^n$, we have for any compact subset K of $\mathbb{R}^n$

$$\sup_{x \in K} \|\Phi[\rho](x) - \Phi[\hat{\rho}](x)\| \leq C(R, K)\|\rho - \hat{\rho}\|_\infty$$

where $\Phi[\rho]$ is the solution of the problem $\Delta\Phi = \rho$ given in Theorem 1.11.

Proof: We choose $\tilde{R} \geq R$ such that $K + B_R(0)$ is contained in a ball of radius $\tilde{R}$ around the origin. Then for $x \in K$, we have

$$\begin{aligned} |\Phi[\rho](x) - \Phi[\hat{\rho}](x)| &\leq \int_{\mathbb{R}^n} |G(x,y)|\,|\rho(y) - \hat{\rho}(y)|\,dy \\ &\leq \int_{B_{\tilde{R}}(0)} |G(x,y)|\,dy \|\rho - \hat{\rho}\|_\infty \leq \int_0^{\tilde{R}} |g(r)| r^{n-1}\,dr \|\rho - \hat{\rho}\|_\infty \end{aligned}$$

■

For a general linear elliptic differential operator L with constant coefficients, a function (resp. distribution N) which satisfies the equation

$$LN = \delta_0, \qquad \text{in } \mathbb{R}^n$$

is called *fundamental solution* of L. Hence,

$$N(x) = \begin{cases} \frac{1}{\omega_1}|x| & n = 1 \\ \frac{1}{\omega_2}\ln|x| & n = 2 \\ \frac{1}{(2-n)\omega_n}|x|^{2-n} & n > 2 \end{cases}$$

is the fundamental solution of the Laplace operator. We have seen that the Poisson problem $\Delta\phi = \rho$ can be solved by *convolution* of the source ρ with the fundamental solution N

$$\phi(x) = \int_{\mathbb{R}^n} N(x - y)\rho(y)\,dy.$$

This is also true for general operators L as we shall see in Section 2.4. If we consider the Poisson problem on a more general domain

$$\begin{aligned} \Delta\phi &= \rho \quad \text{in } \Omega \\ \phi &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

it would still be convenient if the solution could be represented by an integral operator acting on ρ

$$\phi(x) = \int_\Omega G(x,y)\rho(y)\,dy.$$

The kernel G of the solution operator, which can be viewed as a generalization of the fundamental solution, is called *Green's function.*

Finally, we want to give an example where the general Dirichlet problem on a ball of radius R around the origin is solved by integral operators.

Theorem 1.14. *Let $\Omega = B_R(0)$ and let g be continuous on $\partial\Omega$. Then, the function*

$$u(x) = \frac{R^2 - |x|^2}{\omega_n R} \int_{\partial\Omega} \frac{g(y)}{|x-y|^n} \, d\sigma(y), \qquad x \in \Omega \tag{1.13}$$

satisfies $\Delta u = 0$ in Ω and for every $y \in \partial\Omega$, we have

$$\lim_{x \to y} u(x) = g(y).$$

Proof: For abbreviation, we introduce

$$K(x,y) = \frac{R^2 - |x|^2}{\omega_n R} \frac{1}{|x-y|^n}, \qquad x \in \Omega, \ y \in \partial\Omega$$

so that

$$u(x) = \int_{\partial\Omega} K(x,y) g(y) \, d\sigma(y).$$

If $x \in \Omega$, we can differentiate under the integral and find $u \in C^\infty(\Omega)$ and $\Delta u = 0$ in Ω. It remains to check the boundary values of u. To this end, we introduce

$$\psi(x) = \int_{\partial\Omega} K(x,y) \, d\sigma(y)$$

which is harmonic in Ω due to the remarks above. Moreover, ψ is radially symmetric. At the beginning of Section 1.2.2, we have already determined the radially symmetric solutions of Laplace's equation. Since ψ is smooth, the constant c_2 in (1.3) has to be zero which shows that ψ is, in fact, constant. More precisely, with

$$K(0,y) = \frac{R}{\omega_n} \frac{1}{|y|^n}$$

we have

$$\psi(x) = \psi(0) = \frac{R}{\omega_n} \frac{\omega_n R^{n-1}}{R^n} = 1.$$

Now, let $x_0 \in \partial\Omega$ and let $\epsilon > 0$ be given. Choose $\delta > 0$ so that $|g(y) - g(x_0)| < \epsilon$ whenever $|y - x_0| < \delta$. Then

$$\begin{aligned}
|u(x) - g(x_0)| &= \left| \int_{\partial\Omega} K(x,y)(g(y) - g(x_0)) \, d\sigma(y) \right| \\
&\leq \int_{\{y \in \partial\Omega : |y - x_0| \leq \delta\}} K(x,y) |g(y) - g(x_0)| \, d\sigma(y) \\
&\quad + \int_{\{y \in \partial\Omega : |y - x_0| > \delta\}} K(x,y) |g(y) - g(x_0)| \, d\sigma(y) \\
&\leq \epsilon + \int_{\{y \in \partial\Omega : |y - x_0| > \delta\}} K(x,y) |g(y) - g(x_0)| \, d\sigma(y)
\end{aligned}$$

If x is sufficiently close to x_0 (e.g. $|x - x_0| < \delta/2$) we have $|x - y| > \delta/2$ if $y \in \partial\Omega$ and $|y - x_0| > \delta$. Then

$$K(x,y) \le \frac{R^2 - |x|^2}{\omega_n R(\delta/2)^n}$$

on that set and if M is an upper bound for g on $\partial\Omega$ we find

$$|u(x) - g(x_0)| \le \epsilon + \frac{2M(R^2 - |x|^2)R^{n-2}}{(\delta/2)^n}.$$

For $x \to x_0$, we have in particular $|x| \to R$ so that also the second term is less than ϵ for x close enough to x_0. ■

Equation (1.13) is known as *Poisson formula*. The case $x = 0$ leads to the mean-value property presented in Theorem 1.5.

1.3.3. Variational formulation. The variational approach to elliptic PDEs is an extremely powerful technique. Although it is not as elementary as the previous examples (in the sense that quite some heavy mathematical theory is required for a rigorous treatment), we still present it at this point to stress the main ideas.

The general principle is found in many fields of physics where the unique solution of some differential equation can be interpreted as optimizer of a functional (typically an energy functional according to the principle that processes in nature use energy in an optimal way). We will see in the following, that an *energy functional* corresponding to Laplace's equation on a bounded domain Ω is given by

$$\mathscr{E}(u) := \frac{1}{2}\int_\Omega |\nabla u(x)|^2\,dx.$$

Our aim is to characterize the minimizer $\bar{u}$ of $\mathscr{E}$ under the constraint that $\bar{u} = f$ on the boundary $\partial\Omega$. Hence, our class of admissible functions is given by

$$\mathcal{A}_f := \{\, u : \Omega \mapsto \mathbb{R} \,:\, u(x) = f(x) \text{ for } x \in \partial\Omega,\ \mathscr{E}(u) < \infty \,\}.$$

Similar to functionals on a finite dimensional space, a necessary condition in an optimum is that all directional derivatives vanish. A valid *direction* in the set $\mathcal{A}_f$ is a function $v : \Omega \mapsto \mathbb{R}$ which vanishes on the boundary and satisfies $\mathscr{E}(v) < \infty$, i.e. $v \in \mathcal{A}_0$. Now, if $\bar{u} \in \mathcal{A}$ is the minimum of the functional $\mathscr{E}$, then

$$\bar{u} + tv \in \mathcal{A} \qquad \forall t \in \mathbb{R}, \quad v \in \mathcal{A}_0.$$

The necessary condition for an optimum in $\bar{u}$ can be formulated as

$$\frac{d}{dt}\mathscr{E}(\bar{u} + tv)\Big|_{t=0} = 0, \qquad \forall v \in \mathcal{A}_0. \tag{1.14}$$

Using the definition of $\mathscr{E}$, we get

$$\mathscr{E}(\bar{u}+tv) = \mathscr{E}(\bar{u}) + t\int_\Omega \nabla\bar{u}\cdot\nabla v\,dx + t^2\mathscr{E}(v)$$

so that (1.14) implies

$$\int_\Omega \nabla\bar{u}\cdot\nabla v\,dx = 0, \qquad \forall v\in\mathcal{A}_0. \tag{1.15}$$

With partial integration (under the assumption of sufficient smoothness), we deduce further

$$0 = \int_\Omega \nabla\bar{u}\cdot\nabla v\,dx = -\int_\Omega v\Delta\bar{u}\,dx + \int_{\partial\Omega} v\frac{\partial\bar{u}}{\partial n}\,d\sigma$$

and since v vanishes on $\partial\Omega$, we get

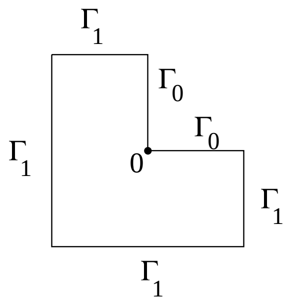
$$\int_\Omega v\Delta\bar{u}\,dx = 0 \qquad \forall v\in\mathcal{A}_0. \tag{1.16}$$

The result which allows us to use (1.16) to deduce that $\bar{u}$ satisfies Laplace's equation is referred to as *fundamental lemma of the calculus of variation.* Assuming that $\Delta\bar{u}$ is continuous and not identically zero, we just pick a point $\bar{x}\in\Omega$ such that, for example, $\Delta\bar{u}(\bar{x})>0$. Due to continuity, $\Delta\bar{u}$ is then positive on a whole neighborhood of $\bar{x}$. Picking a function $\bar{v}\geq 0$ which is compactly supported in that neighborhood, we get the contradiction

$$\int_\Omega \bar{v}\Delta\bar{u}\,dx > 0$$

so that indeed $\Delta\bar{u}=0$.

We stress that all arguments presented here are more or less formal. In particular, the conditions on the boundary value f in conjunction with the shape of the boundary $\partial\Omega$ could imply $\mathcal{A}=\emptyset$ (if no function u of finite energy $\mathscr{E}(u)$ can have these boundary values). Consequently, restrictions have to be put on f and $\partial\Omega$. Also, the question of existence of a minimizer $\bar{u}$ in $\mathcal{A}$ has yet to be answered. Finally, most arguments in the outline above required some smoothness of the minimizer (like $C^2(\bar{\Omega})$) – again a topic which has to be clarified. In fact, it turns out that $C^2(\bar{\Omega})$ cannot always be achieved. As example we mention the following geometry.



Using polar coordinates

$$x = r\cos\phi, \qquad y = r\sin\phi$$

we find that

$$u(r, \phi) = r^{\frac{2}{3}} \sin(2\phi - \pi)$$

is a solution of Laplace's equation with smooth boundary data (in fact, $u \equiv 0$ on Γ_0). Nevertheless, the first derivatives of u are unbounded in $r = 0$ so that $u \notin C^1(\bar{\Omega})$.

Due to such problems, it turns out to be useful to start with a weaker solution concept than in the classical approach. We have already seen that (1.16) implies $\Delta\bar{u} = 0$ if $\bar{u}$ is smooth enough. In that sense, (1.16) is a generalized solution concept since it requires less regularity of $\bar{u}$ but reduces to the classical solution if enough regularity is there. In view of these considerations, relation (1.15) is called *weak form* of Laplace's equation and a solution $\bar{u}$ of the weak form is called *weak solution* (as opposed to *classical solutions* which obey $\Delta\bar{u} = 0$ point wise). The problems touched here are common in the modern theory of PDEs. It is often easy to find some sort of weak solution to an equation but relatively hard to show that the weak solution is in fact a classical solution.

1.4. Fundamental questions

Given some elliptic PDE, we try to answer the following questions:

1) Does a solution exist?
2) Is the solution unique?
3) Does the solution depend continuously on the data of the problem?
4) How can the solution be approximated numerically?

A problem for which questions one to three have positive answers are called *well posed (in the sense of Hadamard)*. Otherwise the problem is said to be *ill posed*. Note, that well–posedness is equivalent to the existence of a continuous solution operator (which maps the given data to the solution).

To answer the above questions, we will first adopt a classical point of view (smooth functions, finite difference approximations). Then, we will switch over to a weak formulation (Sobolev spaces, finite element methods) which is strongly related to the variational approach presented in the previous section.

1.5. Exercises

Exercise 1.1. Show that the *biharmonic operator* $\Delta^2 = \Delta\Delta$ is elliptic.

Exercise 1.2. Show that the Laplace operator in general curvilinear coordinates $y_1, \ldots, y_n$ has the form

$$\Delta_x u = \frac{1}{\sqrt{g}} \sum_{i,j=1}^{n} \frac{\partial}{\partial y_i} \left(g^{ij} \sqrt{g} \frac{\partial u}{\partial y_j} \right)$$

where g^{ij} is the metric tensor

$$g^{ij} = \sum_{k=1}^{n} \frac{\partial y_i}{\partial x_k} \frac{\partial y_j}{\partial x_k}, \qquad g = 1/\det(g^{ij}).$$

Exercise 1.3. Verify that the Laplacian takes the following form in polar coordinates

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

Exercise 1.4. Consider a linear, second order elliptic equation with constant coefficients

$$\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial y_i \partial y_j} + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial y_i} + cu = 0, \qquad y \in \Omega.$$

Show that by some suitable transformation of variables $y = y(x)$, the equation can be reduced to

$$\Delta_x u + \sum_{i=1}^{n} \tilde{b}_i \frac{\partial u}{\partial x_i} + cu = 0.$$

Is that also possible, if a_{ij} are continuous functions of y?

Exercise 1.5. Let Ω be the unit disk in $\mathbb{R}^2$ without the origin (i.e. the boundary of Ω is the unit circle S^1 and the origin). Show that the following boundary value problem has *no* solution $u \in C^2(\Omega) \cap C(\bar{\Omega})$.

$$\begin{aligned} \Delta u &= 0 \qquad \text{in } \Omega \\ u &= 1 \qquad \text{on } S^1 \\ u &= 0 \qquad \text{in } x = 0 \end{aligned}$$

(Hint: Show first that if there is a solution, then there is also a radially symmetric solution.)

Exercise 1.6. Let Ω be a domain in $\mathbb{R}^n$. For abbreviation we set $L := -\Delta$ for the negative Laplacian. A function $u \in C^2(\Omega) \cap C(\bar{\Omega})$ is called *subharmonic* if $Lu \leq 0$ in Ω and *superharmonic* if $Lu \geq 0$.

a) Let u be subharmonic in Ω. Show that for any $x \in \Omega$ and any $R > 0$ such that $\bar{B}_R(x) \subset \Omega$, we have

$$u(x) \leq \frac{1}{\mathrm{vol}(B_R(x))} \int_{B_R(x)} u(y)\, dy.$$

(Hint: Similar to the proof of Theorem 1.5.)

b) Let u be subharmonic in Ω and let $A = \sup_{x \in \bar{\Omega}} u(x)$. If Ω is connected, then either $u = A$ throughout Ω or $u < A$ in Ω.

Exercise 1.7.

a) Derive the formula for the surface area ω_n of the unit sphere in $\mathbb{R}^n$. (Hint: integrate $\exp(-|x|^2)$ in Cartesian and spherical coordinates.)

b) Show that the volume of a ball of radius R is $\omega_n R^n/n$.

Exercise 1.8. Let u be a continuous function on $B_{\tilde{R}}(0)$.

a) By considering the difference quotient, calculate the derivative

$$\frac{d}{dr}\int_{B_r(0)} u(y)\,dy\Big|_{r=R}, \qquad 0 < R < \tilde{R}.$$

b) Show that the mean value property over balls

$$u(0) = \frac{1}{|B_R(0)|}\int_{B_R(0)} u(y)\,dy, \qquad \forall\ 0 < R < \tilde{R}$$

implies the mean value property over spheres

$$u(0) = \frac{1}{|\partial B_R(0)|}\int_{\partial B_R(0)} u(y)\,d\sigma(y), \qquad \forall\ 0 < R < \tilde{R}.$$

Exercise 1.9. Show that, if $n \geq 3$, subharmonic functions on $\Omega = \mathbb{R}^n$ (as defined in Exercise 1.6) can be bounded without being constant. What about the case $n = 2$?

Exercise 1.10. Let $\Omega \subset \mathbb{R}^2$ be the unit square and let $b(x,y) = (1 - |x - 4/7|)(1 - |y - 3/8|)$. Use separation of variables (Section 1.3.1) and any math–package (maple, matlab, etc.) to find an approximate solution of $\Delta u = 0$ in Ω and $u = b$ on $\partial\Omega$.

a) Compare the Fourier–sine approximations with the exact boundary values for $M = 5$ and $M = 10$ at the northern and eastern boundary (plots!).
b) Give an isoline plot of the solution for $M = 10$.
c) Give an estimate for the maximal error of the approximation ($M = 10$) in Ω.

Exercise 1.11. Let $u \in C^2(\Omega) \cap C(\bar{\Omega})$ be harmonic in the domain $\Omega \subset \mathbb{R}^n$. Show that for any $x \in \Omega$ with $\operatorname{dist}(x, \partial\Omega) \geq d$, the following estimate holds for the derivative

$$\left|\frac{\partial u}{\partial x_i}(x)\right| \leq \frac{n}{d} \sup_{y \in \partial B_d(x)} |u(y)|.$$

Hint: First, show a mean value property for the derivatives of u.

Exercise 1.12. Investigate the relation between harmonic and holomorphic (i.e. complex differentiable) functions.

a) Show: if f is holomorphic, its real and imaginary part is harmonic.
b) What are the counterparts of the mean value property (Theorem 1.5), the maximum principle and Theorem 1.6 ? Find other similarities...

CHAPTER 2

Classical solutions

In this chapter we investigate second–order elliptic operators of the form

$$Lu = -\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + cu.$$

The following assumptions are made throughout:

- Ω is an open subset of $\mathbb{R}^n$,
- a_{ij}, b_i and c are continuous on $\bar{\Omega}$,
- the matrix $A = (a_{ij})$ is symmetric and positive definite on $\bar{\Omega}$.

The problems we want to solve are of the form

$$Lu = f \qquad \text{in } \Omega \tag{2.1}$$

with $f \in C(\bar{\Omega})$. On the boundary $\partial\Omega$ we require the so called Dirichlet condition

$$u = g \qquad \text{on } \partial\Omega \tag{2.2}$$

where g is a continuous function on the boundary. In the theory of classical solutions, all derivatives in (2.1) are assumed to be defined in the classical sense. Also, for (2.2) to be meaningful, u should not have jumps at the boundary (otherwise any solution of (2.1) could be converted into a function satisfying the Dirichlet condition just by defining $u = g$ on $\partial\Omega$). Thus, our second requirement is $u \in C(\bar{\Omega})$.

Definition 2.1. *Any function $u \in C^2(\Omega) \cap C(\bar{\Omega})$ which satisfies* (2.1) *and* (2.2) *is called* classical solution *of the Dirichlet problem.*

Beside the Dirichlet condition, other possible boundary conditions are the so called *Neumann conditions*

$$n \cdot (A\nabla u) = g, \qquad \text{on } \partial\Omega \tag{2.3}$$

or mixtures between Dirichlet and Neumann conditions

$$\alpha u + \beta n \cdot (A\nabla u) = g, \qquad \text{on } \partial\Omega. \tag{2.4}$$

where n denotes the outer unit normal vector to Ω. The theory of classical solutions for the Neumann problem will not be considered here in detail. We remark that the case $A = \text{id}$ gives rise to the negative Laplacian $L = -\Delta$ for which the Neumann condition is a condition on the normal derivative $\frac{\partial u}{\partial n} = g$.

To answer the fundamental questions raised in Section 1.4 the most important tool will be the *maximum principle*. It can easily be used to show convergence and continuous dependence on the data. Also, the existence proof for the Laplace Dirichlet problem uses arguments which are based on the maximum principle. It even appears in a discrete form in stability proofs for finite difference schemes which are introduced as a numerical solution technique in Section 2.7.2.

2.1. Maximum Principles

2.1.1. The weak maximum principle. We recall that a necessary condition for $u \in C^2(\Omega)$ to have a maximum in $x_0 \in \Omega$ is

$$\frac{\partial u}{\partial x_i}(x_0) = 0, \qquad i = 1, \dots, n.$$

Moreover, in a maximum the matrix of second derivatives is negative semi definite, i.e.

$$\sum_{i,j=1}^{n} \frac{\partial^2 u}{\partial x_i \partial x_j}(x_0)\xi_i\xi_j \le 0, \qquad \forall \xi \in \mathbb{R}^n. \tag{2.5}$$

Since $A(x_0) = (a_{ij}(x_0))$ is symmetric and positive definite, the square root $S = \sqrt{A(x_0)}$ can be defined. It is a positive definite, symmetric matrix which satisfies $A(x_0) = S^2$. Due to (2.5), we get

$$-\sum_{i,j=1}^{n} a_{ij}(x_0)\frac{\partial^2 u}{\partial x_i \partial x_j}(x_0) = -\sum_{k=1}^{n}\sum_{i,j=1}^{n} \frac{\partial^2 u}{\partial x_i \partial x_j}(x_0) S_{ik} S_{jk} \ge 0.$$

Hence, if the zeroth-order coefficient c of L vanishes on Ω, then u can only have a maximum in $x_0 \in \Omega$ if $(Lu)(x_0) \ge 0$. This observation is the basis for the weak maximum principle.

Theorem 2.2. (weak maximum principle)
Assume that $Lu \le 0$ in a bounded domain Ω and that $c \equiv 0$ in Ω. Then, the maximum of u is achieved on the boundary of Ω. If $Lu < 0$, the maximum cannot be achieved in Ω.

Proof: The case $Lu < 0$ contradicts the existence of a maximum in Ω due to our considerations above. Nevertheless, u achieves its maximum on $\bar{\Omega}$ due to continuity. Correspondingly, the maximum must be located on $\partial\Omega$.

The more general case $Lu \le 0$ is reduced to the previous one by a perturbation argument. Introducing the function

$$u_\epsilon(x) = u(x) + \epsilon \exp(\gamma x_1)$$

we find

$$Lu_\epsilon(x) = Lu(x) - \epsilon\left(\gamma^2 a_{11}(x) - \gamma b_1(x)\right)\exp(\gamma x_1).$$

Choosing γ large enough so that $\gamma^2 a_{11} - \gamma b_1 > 0$ throughout Ω (which is possible since $a_{11} = e_1 \cdot (Ae_1) > 0$ is continuous in $\bar{\Omega}$), we get $Lu_\epsilon < 0$ on Ω for any $\epsilon > 0$. Thus

$$\max_{\bar{\Omega}} u_\epsilon = \max_{\partial\Omega} u_\epsilon.$$

Since $u_\epsilon \to u$ for $\epsilon \to 0$, we also have

$$\max_{\bar{\Omega}} u = \max_{\partial\Omega} u.$$

■

We note that the adjective *weak* refers to the fact that the principle does not exclude that the maximum is also attained inside of Ω. In the *strong maximum principle* this possibility is essentially removed. As usual, the maximum principle 2.2 (and also the following results) can always be formulated as minimum principle by changing the sign of u. More precisely, if $Lu \geq 0$ throughout Ω and $c \equiv 0$ then the minimum of u is attained on the boundary. We also remark that the restriction $c \equiv 0$ can be relaxed if u has a definite sign on Ω. If, for example, $u \geq 0$ in Ω then we still find $Lu(x_0) \geq 0$ in maxima provided $c \geq 0$. Consequently, $Lu \leq 0$ implies that maxima are achieved on the boundary also in that case. If no information on the sign of u is available, we can still use these considerations to get the following weaker result

Corollary 2.3. *Let Ω be bounded and assume $c \geq 0$ in Ω and $Lu \leq 0$. Then*

$$\max_{\bar{\Omega}} u \leq \max_{\partial\Omega} u^+$$

where $u^+ = \max(u, 0)$. In particular, if $Lu = 0$ in Ω, then

$$\max_{\bar{\Omega}} |u| = \max_{\partial\Omega} |u|$$

Proof: If $u \leq 0$ everywhere, the corollary is trivially true since $u^+ = 0$ in that case. Let us therefore turn to the case where

$$\Omega^+ := \{\, x \in \Omega \,:\, u(x) > 0 \,\}$$

is not empty. Clearly, on Ω^+ we have

$$\tilde{L}u := Lu - cu \leq 0.$$

Applying the weak maximum principle to $\tilde{L}$ on Ω^+, we get

$$\max_{\bar{\Omega}} u = \max_{\bar{\Omega}^+} u = \max_{\bar{\Omega}^+} u^+ = \max_{\partial\Omega^+} u^+.$$

The maximum on $\partial\Omega^+$ cannot be achieved at a point $x_0 \in \partial\Omega^+ \cap \Omega$ since $u(x_0) = 0$ but $u > 0$ on Ω^+. Hence,

$$\max_{\partial\Omega^+} u^+ = \max_{\partial\Omega} u^+$$

which completes the first part of the proof. If $Lu = 0$, we simultaneously have $Lu \le 0$ and $Lu \ge 0$ throughout Ω. Consequently,

$$\max_{\bar\Omega} u \le \max_{\partial\Omega} u^+ \le \max_{\partial\Omega} |u|.$$

With $u^- = \min(u, 0)$ we obtain in the same way

$$\min_{\bar\Omega} u \ge \min_{\partial\Omega} u^- \ge -\max_{\partial\Omega} |u|$$

so that the result follows. ∎

With the previous result, we get uniqueness for the Dirichlet problem as a simple consequence of the maximum principle.

Corollary 2.4. (Uniqueness)
Let Ω be bounded and $c \ge 0$. Then, the Dirichlet problem

$$\begin{aligned} Lu &= f \qquad \textit{in } \Omega \\ u &= g \qquad \textit{on } \partial\Omega \end{aligned}$$

admits at most one classical solution.

Proof: Assume u, v are two solutions and consider $w = u - v$. Then, $Lw = 0$ in Ω and $w = 0$ on the boundary. Using Corollary 2.3, we get

$$\max_{\bar\Omega} |w| = \max_{\partial\Omega} |w| = 0$$

so that really $u = v$ in $\bar\Omega$. ∎

Finally, let us mention an important monotonicity property of the solution operator of the Dirichlet problem. By combining L with the boundary operator

$$B : C(\bar\Omega) \mapsto C(\partial\Omega), \qquad Bu := u|_{\partial\Omega}$$

into the operator $\mathcal{L} = \binom{L}{B}$, we can formulate the Dirichlet problem as

$$\mathcal{L}u = d \tag{2.6}$$

if d denotes the *data* vector of the problem $d = \binom{f}{g}$. To get a concise formulation of the following result, we introduce the order relation $\ge$ on pairs by component wise application, i.e.

$$d_1 = \begin{pmatrix} f_1 \\ g_1 \end{pmatrix} \ge \begin{pmatrix} f_2 \\ g_2 \end{pmatrix} = d_2 \quad \Leftrightarrow \quad f_1 \ge f_2 \quad \text{and} \quad g_1 \ge g_2.$$

Corollary 2.5. *Let Ω be bounded and $c \ge 0$. Then*

$$\mathcal{L}u \ge \mathcal{L}v \quad \Longrightarrow \quad u \ge v,$$

or equivalently,

$$\mathcal{L}u \ge 0 \quad \Longrightarrow \quad u \ge 0.$$

Proof: Again, we set $w = v - u$ and find $Lw \leq 0$ in Ω and $w \leq 0$ on the boundary. Hence, $w^+ = 0$ and Corollary 2.3 yields $w \leq 0$ throughout Ω which gives $u \geq v$. ■

For a moment, let us assume the existence of the inverse operator $\mathcal{L}^{-1}$ which maps the vector $d = \binom{f}{g}$ to the solution of the Dirichlet problem with the corresponding data (thus LB^{-1} is the *solution operator* of the Dirichlet problem). Plugging the relation $u = \mathcal{L}^{-1}d$ into the result of Corollary 2.5, we find

$$d \geq 0 \qquad \Longrightarrow \qquad \mathcal{L}^{-1}d \geq 0$$

so that $\mathcal{L}^{-1}$ is a *positive operator.* We remark that positive operators are interesting objects in their own right. They show such a rich structure that investigations independent of elliptic PDEs are rewarding and instructive. In view of this important property of the solution operator, it is not surprising, that in the numerics section elements form the theory of positive matrices are frequently used (positive matrices are the finite dimensional versions of positive operators).

2.1.2. The strong maximum principle. According to Theorem 2.2, the maximum of u is not achieved in Ω if $Lu < 0$. For the case $Lu \leq 0$, the weak maximum principle just assures that the maximum is taken on the boundary but it does not rule out the possibility that maximal values are also found inside Ω. The strong maximum principle clarifies this topic.

Theorem 2.6. (strong maximum principle)
Let Ω be a connected open subset of $\mathbb{R}^n$, assume $Lu \leq 0$ in Ω and assume that u is not constant. If $c \equiv 0$, then u does not achieve its maximum in the interior of Ω. If $c \geq 0$, u cannot achieve a non–positive maximum in the interior. Regardless of the sign of c, u cannot be zero at an interior maximum.

Before we proof the theorem, we first consider a preparatory result which is interesting in its own right.

Lemma 2.7. *Assume $Lu \leq 0$ and let x_0 be a point on $\partial\Omega$ such that $u(x_0) > u(x)$ for every $x \in \Omega$. Also assume that there exists $y \in \Omega$, $R > 0$ such that $B_R(y) \subset \Omega$ and $x_0 \in \partial B_R(y)$. Finally, the derivative of u in the direction of $n(x_0) = (x_0 - y)/R$ is supposed to exist in x_0. If either*

1) $c = 0$, *or*
2) $c \geq 0$ *and* $u(x_0) \geq 0$, *or*
3) $u(x_0) = 0$

then the directional derivative $\frac{\partial u}{\partial n}(x_0)$ is strictly positive.

Proof: Since $u(x_0) > u(x)$ for all $x \in \Omega$, we already know

$$\frac{\partial u}{\partial n}(x_0) \geq 0.$$

To get strict inequality, we again use a perturbation argument. The perturbation function

$$v(x) = \exp(-\alpha|x-y|^2) - \exp(-\alpha R^2), \qquad \alpha > 0$$

is chosen in such a way that it vanishes on $\partial B_R(y)$ and is positive inside the shell

$$S := \left\{ x \,:\, \frac{R}{2} < |x-y| < R \right\}.$$

Applying the differential operator, we get

$$\begin{aligned} Lv(x) = &-\exp(-\alpha|x-y|^2)\Bigg(4\alpha^2 \sum_{i,j=1}^n a_{ij}(x)(x_i-y_i)(x_j-y_j) \\ &+ 2\alpha \sum_{i=1}^n (a_{ii}(x) + b_i(x)(x_i-y_i)) - c(x)\left(1-\exp(-\alpha(R^2-|x-y|^2))\right)\Bigg). \end{aligned}$$

For large enough α and $x \in S$, the term in the bracket will be positive so that we can assume

$$Lv \le 0 \qquad \text{in } S. \tag{2.7}$$

Due to continuity of u and compactness, we get

$$\max_{|x-y|=R/2} (u(x) - u(x_0)) < 0.$$

In particular, we can choose $\epsilon > 0$ so that even

$$\max_{|x-y|=R/2} (u(x) - u(x_0) + \epsilon v(x)) < 0.$$

Since, by construction, $v(x) = 0$ for $x \in \partial B_R(y)$, we get altogether

$$u(x) - u(x_0) + \epsilon v(x) \le 0, \qquad x \in \partial S.$$

Using (2.7) and the assumptions 1), 2) or 3), we find

$$L(u(x) - u(x_0) + \epsilon v(x)) \le -u(x_0)c \le 0.$$

For $c \ge 0$, the weak maximum principle (Theorem 2.2) yields

$$u(x) - u(x_0) + \epsilon v(x) \le 0, \qquad x \in S.$$

Consequently,

$$\begin{aligned} \frac{\partial u}{\partial n}(x_0) &= \lim_{h \searrow 0} \frac{u(x_0) - u(x_0 - hn(x_0))}{h} \\ &\ge \epsilon \lim_{h\searrow 0} \frac{v(x_0 - hn(x_0)) - v(x_0)}{h} = -\epsilon \frac{\partial v}{\partial n}(x_0) \end{aligned}$$

and since

$$\nabla v(x_0) = -2\alpha \exp(-\alpha R^2)(x_0 - y), \qquad n(x_0) = (x_0 - y)/R$$

we get

$$-\epsilon \frac{\partial v}{\partial n}(x_0) = 2\epsilon\alpha R \exp(-\alpha R^2) > 0$$

which completes cases 1) and 2). The case $u(x_0) = 0$ is reduced to the previous ones by considering the modified differential operator

$$\tilde{L} = L - c^-, \qquad c^- = \min(c, 0).$$

Since $u(x) < u(x_0) = 0$ and $c^- \leq 0$, we get

$$\tilde{L}u = Lu - c^- u \leq Lu \leq 0.$$

Moreover, the zeroth order part of $\tilde{L}$ is given by the continuous function $c^+ = c - c^- \geq 0$ so that our previous result can be applied to $\tilde{L}$. ■

Now, we are in the position to prove the strong maximum principle.

Proof: (of Theorem 2.6)
We assume that u is not constant and assumes its maximum M at some point in Ω. Then, $A := \{x : u(x) = M\}$ is a non empty set which is relatively closed in Ω. If we show, that A is at the same time open, we get that A is the whole set Ω (due to connectedness), which contradicts the assumption that u is not constant. To show that A is open, we use a contradiction argument. Assume there is a point $z \in A$, so that each ball around z contains points where u is less than M. In particular, we find a point y with $u(y) < M$ whose distance to the boundary is larger than its distance to z. consequently, the largest ball $B_R(y)$ which is completely contained in $\{x : u(x) < M\}$ touches A in some point x_0. Applying the previous lemma to $B_R(y)$, we conclude that ∇u is non zero at x_0 contradicting the assumption that u assumes its maximum there. ■

2.2. Continuous dependence on the data

The maximum principle can be used to obtain point wise estimates for solutions of $Lu = f$ in bounded domains. To state these bounds, we introduce the following quantities

$$\lambda(x) = \min_{\xi \in S^{n-1}} \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j, \qquad \beta = \max_{\bar{\Omega}} \frac{|b|}{\lambda}.$$

We have the following result.

Theorem 2.8. *Assume that Ω is bounded and contained in the strip between two parallel planes of distance d. Assume also that $c \geq 0$. If $Lu = f$ in Ω and $u = g$ on $\partial\Omega$, then*

$$\max_{\bar{\Omega}} |u| \leq \max_{\partial\Omega} |g| + C \max_{\bar{\Omega}} \frac{|f|}{\lambda}.$$

Here, $C = \exp((\beta + 1)d) - 1$.

Proof: We shall assume $Lu \le f$ and prove that

$$\max_{\bar\Omega} |u| \le \max_{\partial\Omega} u^+ + C \max_{\bar\Omega} \frac{f^+}{\lambda}.$$

The theorem follows by applying the result both to u and $-u$. Since λ and β do not change under rotations of the coordinate system, we may assume that Ω is contained in the strip $0 < x_1 < d$. We set $L_0 = L - c$. Then

$$L_0 \exp(\alpha x_1) = (\alpha^2 a_{11}(x) + \alpha b_1(x)) \exp(\alpha x_1).$$

Since $a_{11} = e_1 \cdot (Ae_1) \ge \lambda$ and $b_1 \ge -\lambda\beta$, we obtain

$$L_0 \exp(\alpha x_1) \ge \lambda(x)\alpha(\alpha - \beta) \exp(\alpha x_1).$$

If we choose α such that $\alpha - \beta > 1$, then $\alpha \ge 1$ and $\exp(\alpha x_1) \ge 1$ so that $L_0 \exp(\alpha x_1) \ge \lambda(x)$. Setting

$$v = \max_{\partial\Omega} u^+ + (\exp(\alpha x_1) - 1) \max_{\bar\Omega} \frac{f^+}{\lambda}$$

we find $v \ge 0$ and thus also $cv \ge 0$ in Ω and $u \le v$ on $\partial\Omega$. Using our previous result $L_0 \exp(\alpha x_1) \ge \lambda(x)$, we get

$$Lv = L_0 v + cv \ge \lambda \max_{\bar\Omega} \frac{f^+}{\lambda} \ge f^+$$

so that $L(u - v) \le f - f^+ = f^- \le 0$. An application of Corollary 2.5 yields

$$u \le v \le \max_{\partial\Omega} u^+ + (\exp(\alpha d) - 1) \max_{\bar\Omega} \frac{f^+}{\lambda}$$

completes the proof of the theorem. ■

We remark that the uniqueness result (Corollary 2.3) can be viewed as a special case of continuous dependence. Indeed, if u, v are two solutions of the Dirichlet problem, then the difference w solves the same problem with zero data and hence $w = 0$ according to the theorem above.

In terms of the combined interior and boundary operator $\mathcal{L} = \binom{L}{B}$, the result is expressed by

$$\|u\|_\infty \le \|Bu\|_\infty + C \left\| \frac{Lu}{\lambda} \right\| \le \left(1 + \frac{C}{\min_{\bar\Omega} \lambda}\right) \max\{\|Bu\|_\infty, \|Lu\|_\infty\}.$$

With the norm

$$\|d\|_\infty = \max\{\|Bu\|_\infty, \|Lu\|_\infty\}, \qquad d = \begin{pmatrix} f \\ g \end{pmatrix}$$

for data vectors, we get the concise formulation

$$\|u\|_\infty \le \tilde{C} \|\mathcal{L}u\|_\infty. \tag{2.8}$$

If the inverse operator exists, this implies the boundedness property

$$\|\mathcal{L}^{-1} d\|_\infty \le \tilde{C} \|d\|_\infty$$

(just set $u = \mathcal{L}^{-1}d$ in (2.8)). If this holds for all data vectors, we get for the operator norm $\|\mathcal{L}^{-1}\|_\infty \le \tilde{C}$. Consequently, continuous dependence of the solution just describes continuity of the inverse operator. Please note that the maximum norm $\|\cdot\|_\infty$ comes into play through the maximum principle and is very natural in this context.

2.3. Existence for the Dirichlet Laplace problem

We will show the existence of a classical solution to the Laplace problem with Dirichlet boundary conditions $g \in C(\partial\Omega)$.

$$\begin{aligned} \Delta u &= 0 && \text{in } \Omega \\ u &= g && \text{on } \partial\Omega \end{aligned} \tag{2.9}$$

The admissible class of domains is quite general.

Definition 2.9. *Let Ω be an open subset of $\mathbb{R}^n$. We say that Ω has the* exterior sphere property *if to each point $x \in \partial\Omega$ there exists a ball $B \subset \bar{\Omega}^c$ such that $\bar{\Omega} \cap \bar{B} = \{x\}$.*

As examples we mention convex sets or sets with C^2 boundaries.

Theorem 2.10. *Assume $\Omega \subset \mathbb{R}^n$ is a bounded domain with the exterior sphere property. Then, the problem* (2.9) *admits a unique solution.*

In order to prove Theorem (2.10), we need the concept of subharmonic functions.

Definition 2.11. *A function $u \in C(\bar{\Omega})$ is called* subharmonic *if for every ball B with $\bar{B} \subset \Omega$ and every function $h \in C(\bar{B})$ which is harmonic in B and satisfies $h \ge u$ on ∂B, we have $h \ge u$ throughout B.*
A subsolution *of the Dirichlet problem is a subharmonic function on Ω which satisfies $u \le g$ on $\partial\Omega$.*
Exchanging $\le$ by $\ge$, we similarly define the notions superhamonic *and* supersolution.

Simple examples of super- and subsolutions are the constant functions

$$\text{Max} := \max_{\partial\Omega} g, \qquad \text{Min} := \min_{\partial\Omega} g.$$

As already indicated by the names, we have that any subsolution for the Dirichlet problem is less than or equal to every supersolution. Indeed, if v is a subsolution and w a supersolution, then $v - w$ is subharmonic (Exercise 2.1) and $v - w \le 0$ on $\partial\Omega$. Using the maximum principle for subharmonic functions (Exercise 2.1), we get either $v = w$ throughout Ω or $v < w$ in Ω. Consequently, subsolutions are bounded from above by any supersolution, and thus in particular by the constant $\text{Max} = \max_{\partial\Omega} g$. Also, we can assume that subsolutions are bounded from below by going over form v to $\max(v, \text{Min})$ which is again a subsolution (see Exercise 2.1).

Our candidate for the solution of the Dirichlet problem is now the supremum of all subsolutions.

$$u(x) = \sup_{v \in \mathrm{Sb}(g)} v(x), \qquad x \in \Omega \tag{2.10}$$

where $\mathrm{Sb}(g)$ denotes the set of all subsolutions to the Dirichlet problem with boundary values g. Due to our considerations above, u is well defined and we have

$$\mathrm{Min} \le u \le \mathrm{Max}.$$

The advantage in the approach (2.10) is that differentiability considerations do not enter in the construction which simplifies the further treatment.

An important tool to prove that u is really the solution is the *harmonic lifting* of a subharmonic function v with respect to some ball $B \subset \Omega$. The idea is to enlarge v on the ball B by going over to the harmonic function $\bar{v}$ on B which has the same values at the boundary ∂B as the original function v. This function exists due to the Poisson formula (1.13)

$$V(x) = \begin{cases} \bar{v}(x) & x \in B \\ v(x) & x \in \Omega \backslash B \end{cases}$$

Since v is subharmonic and $v = \bar{v}$ on ∂B, we have $v \le \bar{v}$ in B so that $v \le V$ in Ω, i.e. the harmonic lifting V is closer to the supremum over all subsolutions than v, provided V is again a subsolution.

Lemma 2.12. *Let v be a subharmonic function on Ω and B a ball whose closure is contained in Ω. Then, the harmonic lifting V of v with respect to B is again subharmonic.*

Let $\Theta \subset \Omega$ be any ball and let h be a harmonic function in Θ which satisfies $V \le h$ on $\partial\Theta$. We have to show that then also $V \le h$ throughout Θ. Since $v \le V$ and v is subharmonic, we get from the definition that $v \le h$ in B. Hence, at least for the points in $\bar{\Theta} \backslash B$ we have $V = v \le h$. To finish the proof, we just have to consider $B \cap \Theta$. Here, V is, by construction, harmonic so that $V \le h$ follows from the corresponding relation at the boundary of $B \cap \Theta$ (due to the maximum principle for harmonic functions). By definition, we have $V \le h$ on $\partial\Theta$ and, because of the previous argument, $V \le h$ on $\partial B \cap \bar{\Theta}$. Since the boundary of $B \cap \Theta$ is contained in the union of $\partial B \cap \bar{\Theta}$ and $\partial\Theta \cap \bar{B}$, the proof is complete. ∎

Finally, we need a result on convergence of sequences of harmonic functions.

Theorem 2.13. *Let $\Omega \subset \mathbb{R}^n$ be a domain. Let $(f_m)_{m \in \mathbb{N}}$ be a sequence of harmonic functions on Ω which is uniformly bounded, i.e. $|f_m(x)| \le M$ for every $x \in \Omega$ and $m \in \mathbb{N}$. Then, $(f_m)_{m \in \mathbb{N}}$ has a subsequence which converges to a harmonic function on Ω, uniformly on compact subsets of Ω.*

Proof: Let

$$\Omega_k := \left\{ x \in \Omega \,:\, |x| \le k, \qquad \operatorname{dist}(x, \partial\Omega) \ge \frac{1}{k} \right\}.$$

Then, Ω_k is compact and $\Omega = \cup_{k=1}^{\infty}\Omega_k$. Using Exercise 1.11, we find locally uniform bounds for the derivatives of all members of the sequence

$$|\nabla f_m(x)| \le nkM, \qquad \forall x \in \Omega_k.$$

Thus, $(f_m)_{m\in\mathbb{N}}$ is equicontinuous on Ω_k. By the Arzela-Ascoli theorem, we can extract a uniformly convergent subsequence on each Ω_k and using a diagonal argument, we obtain a subsequence which converges on Ω uniformly on each Ω_k. Due to Corollary 1.10, the limit of a uniformly convergent sequence of harmonic functions is again harmonic, which completes the proof. ∎

Now, we have all tools at hand to show that (2.10) is the solution of the Dirichlet problem (2.9).

Proof: (of Theorem 2.10)
First, we check that u defined in (2.10) is harmonic. To this end, we pick $x \in \Omega$ and $(v_m)_m \in \mathbb{N} \subset \mathrm{Sb}(g)$ such that $v_m(x) \to u(x)$ for $m \to \infty$. If B is a ball around x whose closure is contained in Ω, we can go over to the lifted sequence $(V_m)_{m\in\mathbb{N}}$ with respect to B. Since $v_m \le V_m$ we also have $V_m(x) \to u(x)$ since $u(x)$ is the supremum over all subsolutions at that point. Using Theorem 2.13, V_m has a convergent subsequence on B (again denoted $(V_m)_{m\in\mathbb{N}}$ to avoid sub indices), which converges uniformly on B to some harmonic function V. Since $V_m \le u$ for all m, we also have $V \le u$. We now show, that even $V = u$, which proves that u is equal to a harmonic function in a neighborhood of x and thus is harmonic.

To get a contradiction to $V = u$, we assume $V(y) < u(y)$ for some $y \in B$. Then, there exists (by definition of u as the supremum over subharmonic functions) a function $\phi \in \mathrm{Sb}(g)$, such that $V(y) < \phi(y) \le u(y)$. Let then V_m^{ϕ} be the harmonic lifting of $\max\{\phi, V_m\}$ with respect to B. As before, a subsequence of V_m^{ϕ} converges uniformly to some harmonic function V^{ϕ} in B. By construction, we have

$$V \le V^{\phi} \quad \text{in } B, \qquad V(y) < V^{\phi}(y)$$

so that certainly $\sup_{\bar{B}}(V - V^{\phi}) \le 0$. Since at x all our sequences converge to the supremum $u(x)$, we even have

$$V(x) = V^{\phi}(x), \qquad \sup_{\bar{B}}(V - V^{\phi}) = 0.$$

Since the difference $V - V^{\phi}$ is harmonic in B, we can apply the strong maximum principle which gives either

- $V - V^{\phi} = \sup_{\bar{B}}(V - V^{\phi}) = 0$, which contradicts $V(y) < V^{\phi}(y)$, or
- $V - V^{\phi} < \sup_{\bar{B}}(V - V^{\phi}) = 0$, which contradicts $V(x) = V^{\phi}(x)$.

It remains to show, that u is actually continuous up to the boundary $\partial\Omega$ and assumes the given boundary data. Let ξ be any point on $\partial\Omega$. Due to the exterior ball property, we can find a ball $B = B_R(y)$ in $(\bar{\Omega})^c$ such that $\bar{B} \cap \bar{\Omega} = \{\xi\}$. We now define

$$w(x) := \begin{cases} \ln \frac{|x-y|}{R} & \text{if } n = 2 \\ R^{2-n} - |x-y|^{2-n} & \text{if } n \neq 2 \end{cases}$$

Then

- w is (super)harmonic in Ω,
- $w > 0$ on $\bar{\Omega}\backslash\{\xi\}$ and
- $w(\xi) = 0$.

(Any continuous function with these properties is called a *barrier* at ξ relative to Ω.) Using the barrier w at ξ, we show that $u(x) \to g(\xi)$ as $x \to \xi$. First, we choose $\epsilon > 0$ and set $M := \sup_{\partial\Omega} |g|$. Then, $\delta > 0$ and $k > 0$ are selected such that

$$|g(x) - g(\xi)| < \epsilon \qquad \text{whenever } x \in \partial\Omega \text{ and } |x-y| < \delta$$

and

$$kw(x) > 2M \qquad \text{whenever } x \in \partial\Omega \text{ and } |x-y| \geq \delta.$$

The function

$$s(x) = g(\xi) - \epsilon - kw(x)$$

then satisfiers for $x \in \partial\Omega$ the relation $s(x) < g(x)$ if $|x - \xi| < \delta$, and $s(x) < -M < g(x)$ for $|x-y| \geq \delta$. Also, s is subharmonic (since w is superharmonic), so that s is, in fact, a subsolution of the Dirichlet problem. Correspondingly,

$$S(x) = g(\xi) + \epsilon + kw(x)$$

is a supersolution and we find $s(x) \leq u(x) \leq S(x)$. Since $w(x) \to 0$ for $x \to \xi$, we find due to continuity of w a neighborhood of ξ such that $|kw| \leq \epsilon$ for $|x - \xi| \leq \delta'$. Thus, for $|x - \xi| \leq \delta'$

$$g(\xi) - 2\epsilon \leq g(\xi) - \epsilon - kw(x) \leq u(x) \leq g(\xi) + \epsilon + kw(x) \leq g(\xi) + 2\epsilon.$$

This shows that u can be continuously extended to $\partial\Omega$ and that the continuation is the prescribed function g. ■

2.4. Existence for the Dirichlet Poisson problem

The Dirichlet Laplace problem (2.9) is now extended to more general sources.

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= g && \text{on } \partial\Omega \end{aligned} \tag{2.11}$$

The idea how to solve (2.11) is the following: Assume, we have some function $w \in C^2(\Omega) \cap C(\bar{\Omega})$ such that $-\Delta w = f$. We then solve the auxiliary problem

$$\begin{aligned} \Delta\phi &= 0 && \text{in } \Omega \\ \phi &= g - w && \text{on } \partial\Omega \end{aligned}$$

which we can do (under suitable assumptions on Ω and g) according to Theorem 2.10. The solution of (2.11) is then obtained with $u = \phi + w$. Thus, we only need to construct a suitable function w to solve the Dirichlet Poisson problem. For $f \in C_0^2(\Omega)$ we have already found the solution in Theorem 1.11

$$w(x) = \int_\Omega g(|x-y|)(-f(y))\,dy. \tag{2.12}$$

However, the required smoothness assumptions and the restrictions on the support seem to be too strong. In fact, the integral (2.12) makes sense already for bounded and integrable functions. We even have

Lemma 2.14. *Let Ω be a bounded domain and let f be measurable and bounded on Ω. Then, the Newtonian potential*

$$w(x) = \int_\Omega g(|x-y|)f(y)\,dy$$

satisfies $w \in C^1(\bar{\Omega})$ and for any $x \in \Omega$

$$\frac{\partial w}{\partial x_i}(x) = \int_\Omega \frac{\partial}{\partial x_i} g(|x-y|)f(y)\,dy, \qquad i = 1, \dots, n.$$

Proof: Since

$$\left|\frac{\partial}{\partial x_i} g(|x-y|)\right| = \left|\frac{1}{\omega_n}\frac{1}{|x-y|^{n-1}}\frac{x_i-y_i}{|x-y|}\right| \le \frac{1}{\omega_n}\frac{1}{|x-y|^{n-1}},$$

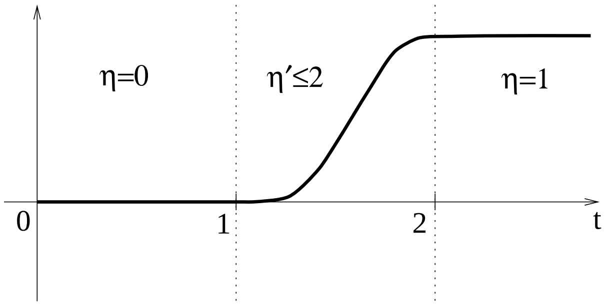
the restriction of $\partial g/\partial x_i$ on $B_1(x)$ is integrable and the one on $\Omega\backslash B_1(x)$ is bounded. Consequently,

$$\begin{aligned} \int_\Omega \left|\frac{\partial}{\partial x_i} g(|x-y|)f(y)\right|\,dy \le{}& \|f\|_{\mathbb{L}^\infty} \|\frac{\partial}{\partial x_i} g(|x-\cdot|)\|_{\mathbb{L}^1, B_1(x)} \\ &+ \|\frac{\partial}{\partial x_i} g(|x-\cdot|)\|_{\mathbb{L}^\infty, \Omega\backslash B_1(x)} \|f\|_{\mathbb{L}^1} \end{aligned}$$

so that

$$v(x) = \int_\Omega \frac{\partial}{\partial x_i} g(|x-y|)f(y)\,dy$$

is well defined for $x \in \bar{\Omega}$. Choosing a smooth function $\eta \in C^1(\mathbb{R})$ as in the figure below



we define $\eta_\epsilon(t) = \eta(t/\epsilon)$ so that the nonzero part starts at $t = \epsilon$ with a slope of at most $2/\epsilon$ and $\eta_\epsilon = 1$ for $t > 2\epsilon$. We use $\eta_\epsilon(|x-y|)$ to cut out the singular point in the Newtonian potential which leads to a smooth function $w_\epsilon \in C^1(\bar{\Omega})$

$$w_\epsilon(x) = \int_\Omega g(|x-y|)\eta_\epsilon(|x-y|) f(y)\, dy.$$

The difference between $\partial w_\epsilon/\partial x_i$ and v defined above is

$$v(x) - \frac{\partial w_\epsilon}{\partial x_i}(x) = \int_{|x-y|\leq 2\epsilon} \frac{\partial}{\partial x_i} \left((1-\eta_\epsilon(|x-y|))g(|x-y|)\right) f(y)\, dy$$

so that

$$\left| v(x) - \frac{\partial w_\epsilon}{\partial x_i}(x) \right| \leq \|f\|_{\mathbb{L}^\infty} \int_{|x-y|\leq 2\epsilon} \left| \frac{\partial}{\partial x_i} g(|x-y|) \right| + \frac{2}{\epsilon} |g(|x-y|)|\, dy$$

where the bound on η_ϵ and its slope has been used. Writing the integral in spherical coordinates, we see that the Jacobian factor r^{n-1} cures the singularity of the first derivative of g, so that only a term of order ϵ remains because the range of integration (in r) is $[0, 2\epsilon]$. The second term in the integral is divided by ϵ, but also its singularity is weaker. Altogether, the integral vanishes with $\epsilon \to 0$, so that $\partial w_\epsilon/\partial x_i$ converges uniformly to v. Since also w_ϵ converges uniformly to w, the result follows. ■

If we want to calculate the second derivatives in the same way, we face the problem, that

$$\frac{\partial^2}{\partial x_i \partial x_j} g(|x-y|) = \frac{1}{\omega_n} \frac{1}{|x-y|^n} \left(\delta_{ij} - \frac{(x_i - y_i)(x_j - y_j)}{|x|^2} \right)$$

is no longer locally integrable around x. If, however, $f(y) - f(x)$ vanishes sufficiently fast for $y \to x$, we find that

$$\left(\frac{\partial^2}{\partial x_i \partial x_j} g(|x-y|) \right) (f(y) - f(x))$$

is again integrable. More precisely, if f is locally Hölder continuous to the exponent $\alpha > 0$, i.e. for any compact $K \subset \Omega$

$$|f(y) - f(x)| < C_K |x-y|^\alpha, \qquad \forall x, y \in K$$

then

$$\left|\frac{\partial^2}{\partial x_i \partial x_j} g(|x-y|)\right| \, |f(y)-f(x)| \le \frac{2C_K}{\omega_n} \frac{1}{|x-y|^{n-\alpha}}$$

which again allows a treatment as for the first derivative.

Lemma 2.15. *Let Ω be a bounded domain and let f be bounded and locally Hölder continuous on Ω. Then*

$$w(x) = \int_\Omega g(|x-y|) f(y) \, dy$$

is contained in $C^2(\Omega) \cap C^1(\bar{\Omega})$ and satisfies $\Delta w = f$.

Proof: See [**9**]. ■

Finally, we can summarize our results on the Dirichlet Poisson problem in

Theorem 2.16. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain which satisfies the exterior sphere property. If f is a bounded, locally Hölder continuous function in Ω and $g \in C(\partial\Omega)$, the Dirichlet Poisson problem* (2.11) *is uniquely solvable.*

2.5. The general Dirichlet problem

The existence proof for the Dirichlet Laplace problem was essentially based on Poisson's formula and the strong maximum principle. While the maximum principle for general linear, elliptic second order operators L with $c \ge 0$ has been proven in Theorem 2.6, an extension of Poisson's formula to general cases is quite involved. Nevertheless, the basic idea is to use results for the Laplacian and extend them by functional analytic methods to other operators. The crucial point in this approach are a priori estimates similar to those obtained in Section 2.2. However, not only estimates on the function itself are required but also bounds for the derivatives in terms of the data of the problem. It turns out, in accordance with the previous section, that the coefficients of the operator and the source should be locally Hölder continuous. For a detailed investigation, we refer to [**9**]. The main result of this so called *Schauder approach* is valid for *strictly elliptic* operators L with $c \ge 0$. To explain this concept, we use our notation from Section 2.2

$$\lambda(x) = \min_{\xi \in S^{n-1}} \sum_{i,j=1}^n a_{ij}(x) \xi_i \xi_j.$$

The operator L is called elliptic, if $\lambda > 0$ in Ω. If λ is even bounded away from zero, i.e. $\lambda \ge \lambda_0 > 0$ in Ω, then L is called strictly elliptic. In our original setup, we assumed that $A(x) = (a_{ij}(x))$ is continuous and positive definite in $\bar{\Omega}$ so that on bounded domains the operator is automatically strictly elliptic (due to continuity and compactness). However, the continuity and definiteness up to the boundary is only a sufficient condition for strict ellipticity.

Theorem 2.17. *Let L be strictly elliptic in a bounded domain Ω with $c \geq 0$, and let f and the coefficients of L be bounded and locally Hölder continuous with exponent $\alpha > 0$. Suppose that Ω satisfies the exterior sphere condition. Then, if g is continuous on $\partial\Omega$, the Dirichlet problem*

$$Lu = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega,$$

has a unique solution $u \in C(\bar{\Omega}) \cap C^{2,\alpha}(\Omega)$.

Here, $C^{2,\alpha}(\Omega)$ is the set of functions which are twice continuously differentiable in Ω and whose second derivatives are locally Hölder continuous to the exponent α. With the stronger assumption that the boundary of Ω is locally the graph of a $C^{2,\alpha}$ function (i.e. Ω is a $C^{2,\alpha}$ domain), that the boundary data are contained in $C^{2,\alpha}(\partial\Omega)$ and that all other prescribed functions are Hölder continuous up to the boundary, we get the stronger result

Theorem 2.18. *Let L be strictly elliptic in a bounded $C^{2,\alpha}$ domain Ω with $c \geq 0$, and let f and the coefficients of L belong to $C^{\alpha}(\bar{\Omega})$. If $g \in C^{2,\alpha}(\partial\Omega)$ then the Dirichlet problem*

$$Lu = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega,$$

has a unique solution $u \in C^{2,\alpha}(\bar{\Omega})$.

The space $C^{2,\alpha}(\bar{\Omega})$ is a Banach space with the norm

$$\|u\|_{2,\alpha;\Omega} := \|u\|_{\infty,\Omega} + \max_{i=1}^{n} \left\| \frac{\partial u}{\partial x_i} \right\|_{\infty,\Omega} + \max_{i,j=1}^{n} \left\| \frac{\partial^2 u}{\partial x_i \partial x_j} \right\|_{\infty,\Omega} + \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{\left| \frac{\partial^2 u}{\partial x_i \partial x_j}(x) - \frac{\partial^2 u}{\partial x_i \partial x_j}(y) \right|}{|x-y|^{\alpha}}.$$

The following result extends the continuous dependence form Section 2.2 to this stronger space.

Theorem 2.19. *Under the assumptions of Theorem 2.18 we find for the solution u of the Dirichlet problem*

$$\|u\|_{2,\alpha;\Omega} \leq C\left(\|u\|_{\infty,\Omega} + \|g\|_{2,\alpha;\partial\Omega} + \|f\|_{0,\alpha;\Omega}\right).$$

Finally, we state that higher regularity of $\partial\Omega$, coefficients and data implies higher regularity of the solution.

Theorem 2.20. *Let Ω be a $C^{2+k,\alpha}$ domain ($k \geq 0, \alpha > 0$) and let $g \in C^{2+k,\alpha}(\partial\Omega)$. Suppose that $u \in C(\bar{\Omega}) \cap C^2(\Omega)$ satisfies $Lu = f$ in Ω, $u = g$ on $\partial\Omega$ where f and the coefficients of the strictly elliptic operator L belong to $C^{k,\alpha}(\bar{\Omega})$. Then $u \in C^{k+2,\alpha}(\bar{\Omega})$.*

2.6. Results in terms of the solution operator

We summarize the previous results as properties of the Dirichlet solution operator $\mathcal{L}$. The general assumptions are that L is a linear, second order, strictly elliptic operator on a bounded domain Ω with α–Hölder continuous coefficients and $c \geq 0$. As usual, $B : C(\bar{\Omega}) \mapsto C(\partial\Omega)$ denotes the trace operator $u \mapsto u|_{\partial\Omega}$. In the previous sections we have investigated several aspects of the Dirichlet problem $\mathcal{L}u = \binom{f}{g}$. In particular, we have studied the fundamental questions concerning existence, uniqueness and continuous dependence of the solution on the data f and g (i.e. the *well-posedness* of the problem).

In terms of the operator $\mathcal{L}$, uniqueness means injectivity and existence leads to surjectivity. Consequently, existence and uniqueness together are equivalent to the invertibility of $\mathcal{L}$. In our case,we have seen that both questions are answered positively, if $g \in C(\partial\Omega)$ and f is locally Hölder continuous, i.e. $f \in C^\alpha(\Omega)$. (We have proven this result for $L = -\Delta$ in Theorem 2.16 and stated it for general L in Theorem 2.17.) Hence,

$$\mathcal{L}^{-1} : C^\alpha(\Omega) \times C(\partial\Omega) \mapsto C^{2,\alpha}(\Omega) \cap C(\bar{\Omega}).$$

For more general data (on $C^{k+2,\alpha}$ domains with $C^{k,\alpha}(\bar{\Omega})$ coefficients and $k \geq 0$) we even have according to Theorem 2.20

$$\mathcal{L}^{-1} : C^{k,\alpha}(\bar{\Omega}) \times C^{k+2,\alpha}(\partial\Omega) \mapsto C^{k+2,\alpha}(\bar{\Omega}).$$

The third aspect of well–posedness, the continuous dependence on the data, naturally translates into a continuity property of $\mathcal{L}^{-1}$. We remark that the linearity of the problem is reflected by the linearity of $\mathcal{L}$ and $\mathcal{L}^{-1}$, so that continuity is equivalent to boundedness. As a direct consequence of the maximum principle, we obtain estimates in the maximum norm on Ω. As stated in Section 2.2, we get

$$\|\mathcal{L}^{-1}d\|_{\infty,\Omega} \leq \tilde{C}\|d\|_{\infty,\Omega}, \tag{2.13}$$

where $d = \binom{f}{g}$ is some admissible data vector from $C^\alpha(\Omega) \times C(\partial\Omega)$ (with $\|d\|_{\infty,\Omega} = \max\{\|g\|_{\infty,\partial\Omega}, \|f\|_{\infty,\Omega}\}$). Theorem 2.19 even yields the stronger estimate

$$\|\mathcal{L}^{-1}d\|_{2,\alpha;\Omega} \leq C\left(\|f\|_{0,\alpha;\Omega} + \|g\|_{2,\alpha;\partial\Omega}\right), \qquad d = \begin{pmatrix} f \\ g \end{pmatrix}$$

(where (2.13) has been used to estimate $\|u\|_{\infty,\Omega}$). Finally, the maximum principle reflects the positivity of $\mathcal{L}^{-1}$ as explained in Section 2.1.1. In a nutshell

existence+uniqueness	$\Longrightarrow$	$\mathcal{L}^{-1}$ exists,
linearity of the problem	$\Longrightarrow$	$\mathcal{L}^{-1}$ linear,
continuous dependence	$\Longrightarrow$	$\mathcal{L}^{-1}$ continuous,
maximum principle	$\Longrightarrow$	$\mathcal{L}^{-1}$ positive.

2.7. Finite Difference Schemes for Elliptic problems

First, we consider finite difference schemes for one–dimensional problems and develop a convergence theory. After that, we will focus on multi–dimensional problems. Our assumptions on the smoothness of the domain, the coefficients and the data of the problem are as in Theorem 2.18. More precisely,

- Ω is a bounded $C^{2,\alpha}$ domain,
- coefficients of L are in $C^{\alpha}(\bar{\Omega})$ for some $\alpha > 0$,
- $c \geq 0$ (to ensure existence and uniqueness)
- $g \in C^{2,\alpha}(\partial\Omega)$, $f \in C^{\alpha}(\bar{\Omega})$.

2.7.1. Finite Difference Schemes in 1-D. If we restrict to 1D problems, then the general, linear elliptic differential operator L of second order is defined by

$$Lu = -au'' + bu' + cu$$

with $a > 0$. In particular (see Exercise 2.7), we may take $a = 1$ and consider the Dirichlet problem on the unit interval $(0, 1)$, i.e.

$$\begin{aligned} Lu = -u'' + bu' + cu &= f \qquad \text{in } \Omega = (0,1) \\ u &= g \qquad \text{on } \partial\Omega = \{0,1\} \end{aligned}$$

To formulate a finite difference scheme for this elliptic problem, we proceed in the following way:

1) we introduce a grid on Ω and the boundary $\Gamma = \partial\Omega$ by

$$\begin{aligned} \Omega_h &= \{x_i \, : \, x_i = ih, i = 1, \dots, n-1\}, \qquad h = \frac{1}{n} \\ \Gamma_h &= \{x_0 = 0, x_n = 1\} \end{aligned}$$

   and denote by u_i the discrete values $u(x_i), x_i \in \Omega_h \cup \Gamma_h$,
2) we discretize the differential operator using difference quotients, i.e.

$$\begin{aligned} u'' &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + \mathcal{O}(h^2), \quad u \in C^4 \\ u' &= \frac{u(x+h) - u(x-h)}{2h} + \mathcal{O}(h^2), \quad u \in C^3 \end{aligned}$$

3) we discretize the boundary values (which is trivial in 1D).

After these steps, we obtain a discrete version of our continuous elliptic problem $Lu = f$ at the grid points $x_i \in \Omega_h$ as follows:

(2.14)
$$\begin{gathered} -\frac{(u_{i+1} - 2u_i + u_{i-1})}{h^2} + b_i \frac{u_{i+1} - u_{i-1}}{2h} + c_i u_i = f_i, \qquad i = 1, \dots, n-1 \\ u_0 = g(0), \; u_n = g(1) \end{gathered}$$

The coefficients in (2.14) depend on the grid point x_i, i.e. $b_i = b(x_i)$, $c_i = c(x_i)$ and $f_i = f(x_i)$. The equation is obviously linear in the three unknowns u_{i-1}, u_i and u_{i+1} and can be rewritten in the form

$$l_i u_{i-1} + d_i u_i + r_{i+1} u_{i+1} = f_i$$

with

$$l_i = -\frac{1}{h^2} - \frac{b_i}{2h}, \quad d_i = \frac{2}{h^2} + c_i, \quad r_i = -\frac{1}{h^2} + \frac{b_i}{2h}.$$

Altogether we obtain a linear system of $n+1$ equations in $n+1$ unknowns defined by

$$\begin{pmatrix} 1 & & & & \\ l_1 & d_1 & r_1 & & \\ & \ddots & \ddots & \ddots & \\ & & l_{n-1} & d_{n-1} & r_{n-1} \\ & & & & 1 \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} = \begin{pmatrix} g_0 \\ f_1 \\ \vdots \\ f_{n-1} \\ g_n \end{pmatrix}$$

which we abbreviate in the following as

$$\mathcal{L}_h \mathbf{u}_h = \mathbf{d}_h \tag{2.15}$$

We note that the infinite number of conditions $Lu(x) = f(x)$ for all $x \in \Omega$ is reduced to a finite number by discretization. Since the original conditions are linear in u, our discretization thus leads to a linear system of equations. Also, the operator $\mathcal{L}_h$ includes the treatment of boundary values so that it compares rather to the operator $\mathcal{L}$ introduced in Section 2.1.1 which also combines the interior operator L with the boundary operator B.

We will now investigate the solvability of the reduced system and study the relation between solutions of the original system and the discretized one. More precisely, we focus on the following questions:

1) Is the linear system always uniquely solvable?
2) What happens in the limit $h \to 0$?
3) How good is the discrete approximation?

To answer the questions, we have to perform a convergence analysis. The typical sub steps in this analysis are *consistency* and *stability* checks. In the consistency part, one tries to answer the question, how good the discretization $\mathcal{L}_h$ approximates the original operator $\mathcal{L}$. To compare the two operators, one first has to overcome the problem that that they work on different spaces. Here, we use discretization and compare the actions of the operators in the finite dimensional space. Another possibility is to suitably interpolate the discrete values and compare the operators on a function space. For discretization, we introduce an operator D_h which works on pairs of functions out of $C(\Omega) \times$

$C(\partial\Omega)$

$$D_h \begin{pmatrix} v \\ g \end{pmatrix} = \begin{pmatrix} g(x_0) \\ v(x_1) \\ \vdots \\ v(x_{n-1}) \\ g(x_n) \end{pmatrix}$$

For convenience, we use the same symbol for the discretization of functions in $C(\bar{\Omega})$ according to the rule

$$D_h u = D_h \begin{pmatrix} u \\ u|_{\partial\Omega} \end{pmatrix}$$

where the right hand side is already defined above. To estimate the quality of approximation on the discrete level, we then introduce the *local discretization error* in a function $u \in C^2(\Omega) \cap C(\bar{\Omega})$

$$\mathbf{l}_h(u) = \mathcal{L}_h D_h u - D_h \mathcal{L} u$$

and use it as a measure for the consistency of the approximation.

Definition 2.21.
The finite difference scheme defined by the operator $\mathcal{L}_h$ is consistent of order m in $v \in C^2(\Omega) \cap C(\bar{\Omega})$, if

$$\|\mathbf{l}_h(v)\|_{h,\infty} \leq C h^m \qquad \textit{for } h \to 0$$

where

$$\|w\|_{h,\infty} = \max_{i=0,\ldots,n} |w_i|, \qquad w \in \mathbb{R}^{n+1}.$$

In contrast to the local discretization error, the actual *approximation error* is obtained by comparing the solution of the original problem with the solution of the discretization. Our aim is of course, that this error vanishes for $h \to 0$.

Definition 2.22.
The finite difference scheme defined by $\mathcal{L}_h$ is convergent of order m, if for the error $\mathbf{e}_h$ defined by

$$\mathbf{e}_h = \mathbf{u}_h - D_h u,$$

(where $\mathbf{u}_h = \mathcal{L}_h^{-1}\mathbf{d}_h$ is the approximate solution and $u = \mathcal{L}^{-1}\left(\begin{smallmatrix} f \\ g \end{smallmatrix}\right)$ the exact solution), we have

$$\|\mathbf{e}_h\|_{h,\infty} \leq C h^m \qquad \textit{for } h \to 0$$

The relation between the two notions of errors is given by the following calculation.

$$\begin{aligned} \mathcal{L}_h \mathbf{e}_h &= \mathcal{L}_h \mathbf{u}_h - \mathcal{L}_h D_h u \\ &= \mathbf{d}_h - (\mathbf{l}_h(u) + D_h \mathcal{L} u) \\ &= \mathbf{d}_h - D_h\left(\begin{smallmatrix} f \\ g \end{smallmatrix}\right) - \mathbf{l}_h(u) = -\mathbf{l}_h(u) \end{aligned}$$

Hence, we have $\mathbf{e}_h = -\mathcal{L}_h^{-1}\mathbf{l}_h(u)$ and

$$\|\mathbf{e}_h\|_{h,\infty} \leq \|\mathcal{L}_h^{-1}\|_{h,\infty}\|\mathbf{l}_h(u)\|_{h,\infty}$$

But then, if $\|\mathcal{L}_h^{-1}\|_{h,\infty}$ is bounded independently of h, consistency of the scheme even means convergence.

Definition 2.23.
The scheme defined by $\mathcal{L}_h$ is called stable, if $\|\mathcal{L}_h^{-1}\|_{h,\infty} \leq K$, is bounded independent of h.

Having isolated the two sub steps *consistency* and *stability*, we immediately get

Theorem 2.24. *Let the scheme defined by $\mathcal{L}_h$ be stable and consistent in the true solution $u = \mathcal{L}^{-1}\binom{f}{g}$ of the problem $Lu = f$ and $Bu = g$. Then, the scheme converges in the sense*

$$\|\mathbf{e}_h\|_{h,\infty} \xrightarrow[h\to 0]{} 0,$$

and the order of consistency is equal to the order of convergence.

The definitions and results given above are rather general, i.e. they may be directly generalized to multi–dimensional problems. For the one–dimensional case, we have

Theorem 2.25. *The difference scheme defined by $\mathcal{L}_h$ for the 1-D problem is consistent in every $u \in C^{2+k,\alpha}(\bar{\Omega})$ with $m \in \mathbb{N}_0$ and $\alpha > 0$. The order of consistency is* $\min\{2, k+\alpha\}$.

Proof by Taylor expansion (see Exercise 2.8). ∎

We observe, that strong regularity results on the true solution are needed in order to obtain a high order of consistency for the finite difference scheme.

In the next step, we try to find necessary conditions for the stability of the finite difference scheme. In particular, we have to estimate the norm of $\mathcal{L}_h^{-1}$. We know already, that $\mathcal{L}_h$ approximates $\mathcal{L}$ for $h \to 0$ so that we can assume that the inverse operator $\mathcal{L}_h^{-1}$ also behaves similar to $\mathcal{L}^{-1}$ in that limit. Thus, it is not surprising, that we need detailed knowledge about the exact solution operator to draw conclusions about the discretized one. In Section 2.2, we have seen that $\mathcal{L}^{-1}$ satisfies a bound exactly like the one we ask of $\mathcal{L}_h^{-1}$. In particular, a suitable norm is the maximum norm which is the reason for using that norm also in the discrete case. The derivation of the bound itself was based on the maximum principle which again is just a result of the positivity of the solution operator $\mathcal{L}^{-1}$. Therefore, it is natural, to investigate the positivity properties of the discretization. It turns out, that this investigation yields the desired invertibility and the bound. Since we need a few aspects from the theory of positive operators, we slightly divert from the stability problem and consider *positive matrices* (which are the finite dimensional counterpart of positive operators).

Definition 2.26. *On $\mathbb{R}^N$, we define an order relation $\geq$ by*

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \geq \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \qquad \Longleftrightarrow \qquad y_i \geq x_i \quad i = 1, \ldots, N.$$

If $y \geq x$ and $y \neq x$, the usual notation is $y > x$. In spaces of dimension bigger than one, there is a third notion $y \gg x$ which describes the case $y_i > x_i$ for all $i = 1, \ldots, N$. Finally, a matrix $B \in \mathbb{R}^{N \times N}$ is called positive ($B \geq 0$), if

$$Bx \geq 0 \qquad \forall x \geq 0.$$

Geometrically, the set of all positive vectors forms a cone in $\mathbb{R}^N$ (in $\mathbb{R}^2$ it is the first quadrant). While strictly positive vectors $x > 0$ can still be located on the boundary of that cone, strongly positive vectors $x \gg 0$ point into the interior. By taking the unit vectors e_i, we immediately find that positive matrices are exactly those matrices which have nonnegative entries. We further have

Lemma 2.27. *Positive multiples, sums and products of positive matrices are again positive. Positive matrices are monotone in the sense that if $B \geq 0$ then*

$$y \geq x \Longrightarrow By \geq Bx.$$

Proof: Exercise 2.9. ■

Similar to the scalar case, one defines order intervals $[a, b] \subset \mathbb{R}^N$ for $a, b \in \mathbb{R}^N$ by

$$x \in [a, b] \qquad \Longleftrightarrow \qquad a \leq x \leq b.$$

In particular, if $y \gg 0$, the order interval $[-y, y]$ can be considered as the unit ball of some norm $\| \cdot \|_y$ on $\mathbb{R}^N$. If we take for example the typical strictly positive vector $e = (1, \ldots, 1)^T$, then $\| \cdot \|_e$ is just the maximum norm since $x \in [-e, e]$ implies for all i that $-1 \leq x_i \leq 1$ so that $\|x\|_\infty \leq 1$ which means that x is in the $\| \cdot \|_\infty$ unit ball. For more general $y \gg 0$ we have

Lemma 2.28. *Let $y \gg 0$ and define*

$$\|x\|_y := \max_{i=1}^{N} \frac{|x_i|}{y_i}, \qquad x \in \mathbb{R}^N.$$

Then, the order interval $[-y, y]$ is the unit ball with respect to the $\|\cdot\|_y$. Defining the matrix norm

$$\|B\|_y = \max_{\|x\|_y = 1} \|Bx\|_y$$

we have for any positive matrix $B \geq 0$

$$\|B\|_y = \|By\|_y.$$

In particular, for $y = e = (1, \ldots, 1)^T$, we find

$$\|Bx\|_\infty \leq \|Be\|_\infty \|x\|_\infty \qquad \forall x \in \mathbb{R}^N.$$

Proof: It is easy to check that $\|\cdot\|_y$ defines a norm on $\mathbb{R}^N$. Now, let $-y \leq x \leq y$, i.e. $|x_i| \leq y_i$ for all i. Then, since $y \gg 0$, we have $\frac{|x_i|}{y_i} \leq 1$ and thus $\|x\|_y \leq 1$. Conversely, if $\|x\|_y \leq 1$, we conclude that $-y \leq x \leq y$ so that

$$[-y, y] = \{x \,:\, \|x\|_y \leq 1\,\}.$$

If $B \geq 0$, then $-y \leq x \leq y$ implies $-By \leq Bx \leq By$ and thus also

$$\frac{|(Bx)_i|}{y_i} \leq \frac{(By)_i}{y_i}$$

so that $\|Bx\|_y \leq \|By\|_y$. Using this relation for those $x \in [-y, y]$ which satisfy $\|x\|_y = 1$, we get by definition of the matrix norm $\|B\|_y \leq \|By\|_y$. Since also $\|y\|_y = 1$, we even have $\|B\|_y = \|By\|_y$. ■

An immediate consequence of this Lemma is that positive matrices B which are contracting in the sense $By \ll y$ are also contracting in the $\|\cdot\|_y$–norm because $B[-y, y]$ is strictly contained in $[-y, y]$, or in other words, $\|B\|_y = \|By\|_y < \|y\|_y = 1$. We thus get

Lemma 2.29. *Let $y \gg 0$. If $B \geq 0$ with $By \ll y$, then $(I - B)^{-1}$ exists and is positive. Similarly, if $By \gg 0$ and $(I - B) \geq 0$ then B^{-1} exists and is positive.*

Proof: Since $\|B\|_y < 1$, the von Neumann series

$$(I - B)^{-1} = \sum_{k=0}^{\infty} B^k$$

converges in the $\|\cdot\|_y$ norm (and thus in any matrix norm). Together with B, each power is positive so that the same holds for the series. If $By \gg 0$, we can write

$$(I - B)y = y - By \ll y$$

so that the result follows if we replace B by $(I - B)$ in the argument above. ■

With this result, we have established the connection between positivity properties and invertibility of a matrix. Considering the matrix $\mathcal{L}_h$, we first note that the diagonal terms are always strictly positive. Indeed, the top and bottom entry is equal to one and the other entries are of the form $d_i = 2/h^2 + c_i$ which is certainly positive since $c_i = c(x_i) \geq 0$ and $h > 0$. The off-diagonal terms, on the other hand, are non positive, provided

$$h \leq \frac{2}{\max_{\bar{\Omega}} |b|} =: h_0. \tag{2.16}$$

(A matrix whose off-diagonal elements are non positive is sometimes called an L_0–matrix.) If we collect the diagonal elements of $\mathcal{L}_h$ in the matrix $\mathcal{D}$, then $\mathcal{D}$

is invertible due to the strict positivity. Of course, we also have $\mathcal{D}, \mathcal{D}^{-1} \geq 0$. By subtracting $\mathcal{D}$ from $\mathcal{L}_h$, we isolate the non positive terms, i.e.

$$\mathcal{D} - \mathcal{L}_h \geq 0.$$

Multiplying this inequality by $\mathcal{D}^{-1} \geq 0$, we get

$$I - \mathcal{D}^{-1}\mathcal{L}_h \geq 0.$$

To apply Lemma 2.29, we just need a vector $y \gg 0$ such that $\mathcal{L}_h y \gg 0$ (since then also $\mathcal{D}^{-1}\mathcal{L}_h y \gg 0$). Given such a vector y, we can conclude that $\mathcal{D}^{-1}\mathcal{L}_h$ has a positive inverse so that also $\mathcal{L}_h^{-1}$ exists and is positive. To estimate $\mathcal{L}_h^{-1}$ is the maximum norm, we observe that

$$\mathcal{L}_h y \geq \min_{i=1}^{N} (\mathcal{L}_h y)_i e, \qquad e = (1, \ldots, 1)^T.$$

Multiplying with $\mathcal{L}_h^{-1}$ yields

$$0 \leq \mathcal{L}_h^{-1} e \leq \frac{y}{\min_{i=1}^{N} (\mathcal{L}_h y)_i}$$

which finally gives

$$\|\mathcal{L}_h^{-1}\|_{h,\infty} = \|\mathcal{L}_h^{-1} e\|_{h,\infty} \leq \frac{\|y\|_{h,\infty}}{\min_{i=1}^{N} (\mathcal{L}_h y)_i}. \tag{2.17}$$

The remaining problem is the construction of a suitable vector $y \gg 0$. Of course, y has to be chosen in such a way that its components stay bounded for $h \to 0$ (to bound $\|y\|_{h,\infty}$ in (2.17)). In addition, $\mathcal{L}_h y$ must be bounded away from zero. The latter condition could be obtained by taking for y the solution of

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

While this would guarantee the required lower bounds, we are not sure about the norm of the solution y. The trick is to resort to the corresponding continuous problem which we have much better under control. Since the data vector $d_h = e$ is a discretization of $d = \binom{1}{1}$ (i.e. $f \equiv 1$ and $g \equiv 1$), the continuous problem is

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

Applying the minimum principle (since $Lv \geq 0$), we get $v \geq 1$ and since v is continuous, we also find $\|v\|_\infty = \sup_{\bar{\Omega}} |v| < \infty$. Setting

$$y = D_h v,$$

we get $y \gg e \gg 0$ (independent of h, i.e. the dimension of the discrete space), and also

$$\mathcal{L}_h y = D_h \mathcal{L} v + \mathbf{l}_h(v) = e + \mathbf{l}_h(v) \gg \frac{1}{2} e, \qquad \text{for } h \text{ small enough.}$$

(Here, we need that $\|\mathbf{l}_h(v)\|_{h,\infty} \to 0$ for $h \to 0$ for the solution v of the auxiliary problem.) Using the so defined vector y in connection with (2.17), we end up with the bound

$$\|\mathcal{L}_h^{-1}\|_{h,\infty} \le 2\|v\|_\infty$$

which proves that the scheme is stable. We summarize the result in

Theorem 2.30. *Under the general smoothness assumptions, the finite difference scheme defined by $\mathcal{L}_h$ is convergent. If the solution is contained in $C^{2+k,\alpha}(\bar{\Omega})$, the order of convergence is $\min\{2, k+\alpha\}$.*

We remark that extra smoothness, which leads to higher order of convergence, is typically connected to a corresponding regularity of the data, the coefficients of L and the domain Ω (see Section 2.5).

Before we go over to 2D problems, let us first investigate the matrix $\mathcal{L}_h$ a little closer. To get an idea about the structure of $\mathcal{L}_h$, we consider its eigenvalues (which essentially describe the action of the matrix). To this end, the following result is very useful.

Lemma 2.31. (Gerschgorin)
Let $A \in \mathbb{C}^{N\times N}$ be a complex matrix. Then, the eigenvalues of A are contained in the union of closed disks $\overline{B_{r_i}(A_{ii})}$ in the complex plane, where $r_i = \sum_{j\neq i} |A_{ij}|$.

Proof: Let $\lambda \in \mathbb{C}$ be an eigenvalue of A with a corresponding eigenvector x which has been normalized to one in the norm

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

Hence, there is at least one index i, such that $|x_i| = 1$ and thus

$$(\lambda - A_{ii})x_i = \sum_{j\neq i} A_{ij}x_j.$$

Taking absolute values on both sides and using $|x_i| = 1$, $|x_j| \le 1$ for $i \neq j$ yields

$$|\lambda - A_{ii}| \le \sum_{j\neq i} |A_{ij}| = r_i.$$

■

For the matrix $\mathcal{L}_h$, we find

$$\sigma(\mathcal{L}_h) \subset \bigcup_{i=1}^{n-1} B_{|l_i|+|r_i|}(d_i) \cup \{1\}.$$

For $h < h_0$ (defined in (2.16)), we get

$$\left|\frac{b_i}{2}\right| < \frac{1}{h}$$

so that

$$|l_i| = \frac{1}{h^2} + \frac{b_i}{2h}, \qquad |r_i| = \frac{1}{h^2} - \frac{b_i}{2h}$$

and hence $|l_i| + |r_i| = 2/h^2$. Consequently,

$$\sigma(\mathcal{L}_h) \subset \bigcup_{i=1}^{n-1} B_{2/h^2}\left(\frac{2}{h^2} + c_i\right) \cup \{1\},$$

i.e. all eigenvalues are contained in the positive half space $\{z \in \mathbb{C} : \Re(z) \geq 0\}$. While $c_i > 0$ for all i implies invertibility of $\mathcal{L}_h$ (since $0 \notin \sigma(\mathcal{L}_h)$), the case $c_i = 0$ for some i needs a more involved analysis than the Lemma of Gerschgorin. (We proved it based on a contraction property.) The case $c_i < 0$, on the other hand, can definitely lead to $0 \in \sigma(\mathcal{L}_h)$ for special choices of c.

We conclude the section on 1D problems with the remark that instead of $\mathcal{L}_h$ we can also derive another linear system which just operates on the interior points $x_1, \ldots, x_{n-1}$. In some sense, it is therefore a discretization of the continuous operator L. However, the boundary values enter in the right hand side, which has no direct counterpart in the original problem. To derive that system, we use the fact that u_0 and u_n are already known to be g_0 and g_n. Extracting the system for the remaining components $u_1, \ldots, u_{n-1}$, we find the equivalent set of equations

$$\begin{pmatrix} d_1 & r_1 & & \\ & \ddots & \ddots & \ddots \\ & & l_{n-1} & d_{n-1} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} f_1 - l_1 g_0 \\ \vdots \\ f_{n-1} - r_{n-1} g_n \end{pmatrix}$$

which we abbreviate $\bar{\mathcal{L}}_h \bar{\mathbf{u}}_h = \bar{\mathbf{d}}_h$. In contrast to the original matrix, $\bar{\mathcal{L}}_h$ can be symmetric if $r_i = l_{i+1}$, i.e.

$$-\frac{1}{h^2} + \frac{b_i}{2h} = -\frac{1}{h^2} - \frac{b_{i+1}}{2h} \quad \Longleftrightarrow \quad b_i = -b_{i+1}.$$

In particular, if $b = 0$, the latter condition is satisfied. For the important case of the Laplacian, we have in addition $c = 0$ so that

$$\bar{\mathcal{L}}_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 \end{pmatrix}$$

Since $\bar{\mathcal{L}}_h$ is symmetric, all eigenvalues are real and due to the Lemma of Gerschgorin, they are not negative. Finally, the invertibility of $\bar{\mathcal{L}}_h$ excludes the eigenvalue zero so that $\bar{\mathcal{L}}_h$ is positive definite. In order to solve the approximate system it is therefore advisable to use the form $\bar{\mathcal{L}}_h$ because very efficient numerical algorithms exist for positive definite symmetric matrices.

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.

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

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

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

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 $i,j = 0,\ldots,n$ such that

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

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 $\mathcal{L}_h^{-1}$. First, we introduce the (i,j)–unit vector $\mathbf{e}_h^{(ij)}$ whose components are all zero up to the one corresponding to the grid point (i,j). Using pairs (k,l) to address the components of the vector, we have

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

Multiplying the matrix with $\mathbf{e}_h^{(ij)}$ 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.

$$Lu = -(a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy}) + b_1u_x + b_2u_y + cu$$

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

$$Lu = -a_{11}u_{xx} + b_1u_x + \frac{c}{2}u - a_{22}u_{yy} + b_2u_y + \frac{c}{2}u = (L^x + L^y)u$$

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

$$\text{(2.20)} \qquad \begin{bmatrix} & r^y_{ij} & \\ l^x_{ij} & d^x_{ij} + d^y_{ij} & r^x_{ij} \\ & l^y_{ij} & \end{bmatrix}$$

and

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

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

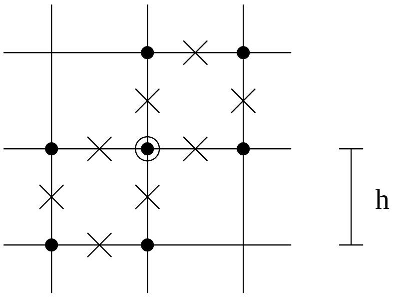
$$m_{ij} = \frac{a_{12}(x_i, y_j)}{2h^2}$$

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

$$
\frac{1}{8h^2}\begin{bmatrix} & & -1 & 0 & 1 \\ & & 0 & & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & & 0 & & \\ 0 & 0 & 0 & & \end{bmatrix} + \frac{1}{8h^2}\begin{bmatrix} & & 0 & 0 & 0 \\ & & 0 & & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & & 0 & & \\ 1 & 0 & -1 & & \end{bmatrix}
= \frac{1}{8h^2}\begin{bmatrix} & & -1 & 0 & 1 \\ & & 0 & & 0 \\ -1 & 0 & 2 & 0 & -1 \\ 0 & & 0 & & \\ 1 & 0 & -1 & & \end{bmatrix}. \tag{2.23}
$$

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 $a_{12} \leq 0$, 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

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

which holds for example if

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

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 $b_1 \leq 0$ 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

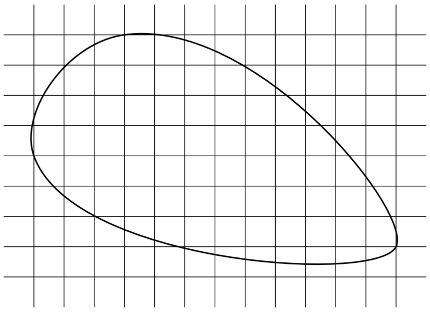
$$\frac{1}{h}\begin{bmatrix} & b_2^- & \\ -b_1^+ & |b_1| + |b_2| & b_1^- \\ & -b_2^+ & \end{bmatrix}.$$

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.

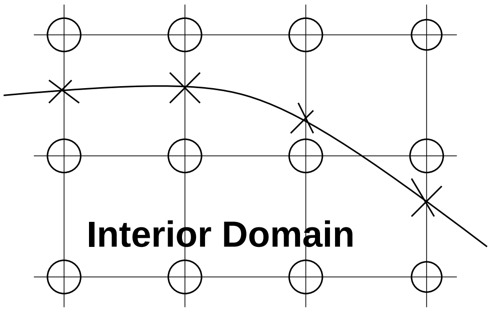
To show that $\mathcal{L}_h^{-1}$ is stable, we need to find again a vector $y \gg 0$ such that $\mathcal{L}_h y \gg 0$ and $\|y\|_{h,\infty} \leq C$, $(\mathcal{L}_h y)_i \geq \mu > 0$ with μ and C independent of h. In Exercise 2.14, such a vector is constructed for the general Dirichlet problem. We 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 [**10**]. 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

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

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 [**10**].) 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 [**10**].

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 $\tilde{\mathcal{L}} = \binom{-\Delta}{\tilde{B}}$ 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 [**10**] 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

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

2.8. Fundamental solutions and Green's functions

Definition 2.37. *Let $L = \sum_{|\alpha| \leq N} c_\alpha \nabla^\alpha$ be a differential operator with constant coefficients on $\mathbb{R}^n$. Then a fundamental solution for L is a distribution $E \in \mathscr{D}'(\mathbb{R}^n)$ satisfying the equation $LE = \delta$.*

Of course, fundamental solutions are unique only up to a solution of the homogeneous equation $Lu = 0$. Typically, one chooses a particular solution with a behavior at infinity which is as nice as possible. The significance of the fundamental solution lies in the fact that it can be used to solve general problems of the form

$$Lu = f, \qquad f \in \mathscr{D}'(\mathbb{R}^n), \quad \text{supp } f \text{ compact}.$$

Indeed, setting $u = E * f$, we find

$$Lu = L(E * f) = (LE) * f = \delta * f = f. \tag{2.27}$$

The *Theorem of Malgrange and Ehrenpreis* states that any differential operator with constant coefficients admits a fundamental solution [**15**]. Here, we show that for the case $L = \Delta$, a fundamental solution is given by

$$N(x) = g(|x|) = \begin{cases} \frac{1}{\omega_1}|x| & n = 1 \\ \frac{1}{\omega_2} \ln |x| & n = 2 \\ \frac{1}{(2-n)\omega_n} |x|^{2-n} & n > 2. \end{cases}$$

(For other examples we refer to [**13**].) Indeed, since $N \in \mathbb{L}^1_{loc}(\mathbb{R}^n)$, we have for any test function ϕ

$$\langle \Delta N, \phi \rangle = \langle N, \Delta\phi \rangle = \int_{\mathbb{R}^n} N(x) \Delta\phi(x)\, dx.$$

Using Lemma 1.4 with $\Omega = B_R(0)$ bigger than the support of ϕ (so that $\phi = 0$ and $\partial\phi/\partial n = 0$ on $\partial\Omega$), we get

$$\int_{\mathbb{R}^n} N(x) \Delta\phi(x)\, dx = \phi(0) = \langle \delta, \phi \rangle$$

which shows that $\Delta N = \delta$. Using the important relation (2.27), we recover the result of Section 1.3.2 that $\Phi = N * \rho$ is a solution of the Poisson problem $\Delta\Phi = \rho$ on $\mathbb{R}^n$. If ρ is a compactly supported, generalized function on $\mathbb{R}^n$, the equation is satisfied in the sense of distributions. If ρ has more regularity (like $C_0^2(\mathbb{R}^n)$), the equation is even satisfied in the classical sense (see Theorem 1.11).

The adjective *fundamental* for the solution N of $\Delta N = \delta$ is justified in the sense that it can be used as building block for solutions of more general equations. In particular, (2.27) shows that a solution operator of the differential equation $\Delta u = f$ is given by the integral operator $u = N * f$. (Note that convolution of two distributions can, in general, not be written as an integral. It only generalizes an integral expression.) Apart from the full space problem, the convolution $N * f$ also yields solutions to problems on bounded domains.

Assume, for example, that $f \in \mathscr{D}'(\Omega)$ is the restriction of a compactly supported distribution $\hat{f}$ on $\mathbb{R}^n$ to the set Ω. (If Ω is bounded, any $f \in \mathbb{L}^\infty(\Omega)$ falls in this class since the extension by zero of f to $\mathbb{R}^n$ is a compactly supported generalized function on $\mathbb{R}^n$.) Using the previous result, we get again that

$$v = N * \hat{f}\Big|_\Omega$$

satisfies $\Delta v = f$ in $\mathscr{D}'(\Omega)$. In order to satisfy also boundary conditions on $\partial\Omega$, however, this approach has to be modified. First of all, we have to require enough regularity on f in order to refer to values of $N * \hat{f}$ on the boundary. Here, we take $f \in \mathbb{L}^\infty(\Omega)$ and Ω bounded. Then, the convolution can be written as

$$v(x) = N * \hat{f}(x) = \int_\Omega N(x-y)f(y)\,dy, \qquad x \in \Omega.$$

(The restricted domain of integration is due to the fact that we convolute N with $\hat{f}$ which is equal to f on Ω and zero in Ω^c.) If we want to solve the Dirichlet problem

$$\begin{aligned} \Delta u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{2.28}$$

we obviously have to modify v (which already satisfies $\Delta v = f$) by adding a harmonic function h with the property $h = -v$ on $\partial\Omega$. More generally, if $H(x, y)$ is harmonic in x for each $y \in \Omega$ and satisfies $H(x, y) = -N(x - y)$ whenever $x \in \partial\Omega$ and $y \in \Omega$, then

$$u(x) = \int_\Omega \left(N(x-y) + H(x,y)\right) f(y)\,dy$$

satisfies (2.28). This idea gives rise to the following

Definition 2.38. *Let Ω be an open subset of $\mathbb{R}^n$. A function $G : \bar{\Omega} \times \Omega \mapsto \mathbb{R}$ is called Green's function, if*

1) $x \mapsto H(x, y) = G(x, y) - N(x - y)$ *is harmonic in Ω and continuous in $\bar{\Omega}$ for all $y \in \Omega$,*
2) $G(x, y) = 0$ *for all $x \in \partial\Omega$ and $y \in \Omega$.*

Note that, on bounded domains, the Green's function is uniquely defined since $x \mapsto H(x, y)$ is the classical solution of a Dirichlet Laplace problem which is unique. Since the singularity of G is exactly the one of the fundamental solution, the following result is not surprising.

Lemma 2.39. *Let G be a Green's function on $\Omega \subset \mathbb{R}^n$. Then*

$$\Delta_x G(x, y) = \delta(x - y), \qquad \forall y \in \Omega$$

holds in $\mathscr{D}'(\Omega)$.

Proof: For given $y \in \Omega$, we can find $\epsilon > 0$ such that $B_{3\epsilon}(y) \subset \Omega$. Thus, we can construct a function $\eta \in \mathscr{D}(\Omega)$ which is equal to one on $B_\epsilon(y)$ and equal to zero outside of $B_{2\epsilon}(y)$. Picking an arbitrary test function $\phi \in \mathscr{D}(\Omega)$, we have

$$\langle \Delta_x G(x,y), \phi(x) \rangle_x = \langle \Delta_x G(x,y), \eta(x)\phi(x) \rangle_x + \langle \Delta_x G(x,y), (1-\eta(x))\phi(x) \rangle_x .$$

Using the fact that $\Delta_x G(x,y) = 0$ for x away from y, we find

$$\langle \Delta_x G(x,y), \phi(x) \rangle_x = \langle \Delta_x G(x,y), \eta(x)\phi(x) \rangle_x .$$

The splitting $G(x,y) = N(x-y) + H(x,y)$ then yields (due to $\Delta_x H(x,y) = 0$)

$$\langle \Delta_x G(x,y), \phi(x) \rangle_x = \langle \Delta_x N(x-y), \eta(x)\phi(x) \rangle_x .$$

Since N is a fundamental solution and $\eta(y) = 1$, we end up with

$$\langle \Delta_x G(x,y), \phi(x) \rangle_x = \langle \delta(x-y), \eta(x)\phi(x) \rangle_x = \langle \delta(x-y), \phi(x) \rangle_x .$$

■

We remark that the equation

$$L_x G(x,y) = \delta(x-y), \qquad \forall y \in \Omega \tag{2.29}$$

is also meaningful for general differential operators (with C^∞ coefficients), giving rise to Green's functions for L. However, the question of suitable boundary conditions (and whether solutions to (2.29) have enough regularity to talk about boundary values) is quite complicated in the general case. For a formal discussion we refer to [**14**]. Due to Definition 2.38, which gives additional structure of G for the special case of the Laplacian, we can, however, infer several results.

Lemma 2.40. *Let Ω be bounded, $f \in \mathbb{L}^\infty(\Omega)$ and G be the Green's function. Then*

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

defines a function in $C(\bar{\Omega})$ which solves $\Delta u = f$ in $\mathscr{D}'(\Omega)$ and satisfies $u = 0$ on $\partial\Omega$. If $x \mapsto H(x,y)$ is contained in $C^1(\bar{\Omega})$, then also $u \in C^1(\bar{\Omega})$ and

$$\nabla u(x) = \int_\Omega \nabla_x G(x,y) f(y)\, dy, \qquad x \in \Omega.$$

Proof: The regularity of u is obtained from Lemma 2.14 which shows that integrals over the fundamental solution yield $C^1(\bar{\Omega})$ functions. If H has also sufficient regularity, the same holds for u. Taking any test function $\phi \in \mathscr{D}(\Omega)$, we get

$$\begin{aligned} \langle \Delta u, \phi \rangle = \langle u, \Delta\phi \rangle = \int_\Omega \int_\Omega G(x,y) f(y) \Delta\phi(x)\, dy\, dx \\ = \int_\Omega \langle G(x,y), \Delta\phi(x) \rangle_x f(y)\, dy. \end{aligned}$$

Using Lemma 2.39, we have

$$\langle \Delta u, \phi \rangle = \int_\Omega \phi(y) f(y)\, dy = \langle f, \phi \rangle$$

which shows that $\Delta u = f$ in $\mathscr{D}'(\Omega)$. The Dirichlet condition follows directly from property 2) in Definition 2.38. ∎

Before we consider the case of more general Dirichlet conditions, let us first investigate the smoothness of the solution u in Lemma 2.40 if the source f has extra smoothness. It turns out that the Laplacian Δ is a *hypoelliptic* operator which means that $\Delta u \in C^\infty$ implies $u \in C^\infty$. In general, we have

Definition 2.41. *The linear partial differential operator L in Ω is said to be hypoelliptic if for any open subset $U \subset \Omega$ and $u \in \mathscr{D}'(\Omega)$, $Lu \in C^\infty(U)$ implies $u \in C^\infty(U)$.*

Suppose that E is any fundamental solution of L in $\mathbb{R}^n$. Consequently, $LE = 0$ in $\mathbb{R}^n \backslash \{0\}$. Hence, if L is hypoelliptic, E must be a C^∞ function in $\mathbb{R}^n \backslash \{0\}$. A classical theorem of L. Schwartz states that the converse is also true.

Theorem 2.42. *If there exists a fundamental solution E of L which is a C^∞–function in $\mathbb{R}^n \backslash \{0\}$ then L is hypoelliptic in $\mathbb{R}^n$.*

Proof: Let $U \subset \mathbb{R}^n$ be an open set and assume $u \in \mathscr{D}'(U)$ satisfies $Lu = f \in C^\infty(U)$. Picking any $x_0 \in U$, we can find $\epsilon > 0$ so that $B_{5\epsilon}(x_0) \subset U$. Selecting a function $\eta \in \mathscr{D}(\mathbb{R}^n)$ which is equal to one on $B_{3\epsilon}(x_0)$ and zero outside of $B_{4\epsilon}(x_0)$, we restrict the distribution u to a neighborhood of x_0 by taking the product ηu. Applying the differential operator, we get

$$L(\eta u) = \eta L u + v = \eta f + v$$

where v is obtained with Leibniz' formula. In particular, v vanishes on $B_{2\epsilon}(x_0)$ and outside of $B_{4\epsilon}(x_0)$ since the derivatives of η are zero in these sets (v is a sum of terms $\nabla^\alpha \eta \nabla^\beta u$ with $|\alpha| \geq 1$). Taking another cut–off function χ which is one on $B_\epsilon(0)$ and vanishes outside $B_{2\epsilon}(0)$, we finally get

$$\eta u = \delta * (\eta u) = (LE) * (\eta u) = E * (L(\eta u)) = E * (\eta f) + E * v$$

and, by introducing χ,

$$\eta u = E * (\eta f) + (\chi E) * v + ((1 - \chi) E) * v.$$

The first term on the right hand side is a convolution of a test function ηf and a distribution E, i.e. $E * (\eta f) \in C^\infty(\mathbb{R}^n)$. The third term is a convolution of a compactly supported distribution v and a C^∞–function $(1 - \chi)E$ and thus also contained in $C^\infty(\mathbb{R}^n)$. Finally, the middle term is a distribution whose support is contained in

$$\operatorname{supp}(\chi E) * v \subset \operatorname{supp}(\chi E) + \operatorname{supp} v \subset B_\epsilon(0) + B_{2\epsilon}(x_0)^c \subset B_\epsilon(x_0)^c.$$

In particular, $(\chi E) * v$ vanishes on $B_\epsilon(x_0)$ and thus ηf is a C^∞–function on this neighborhood. Since $\eta \equiv 1$ on $B_\epsilon(x_0)$ and x_0 was an arbitrary point in U, it follows $u \in C^\infty(U)$ and the proof is complete. ■

The result applied to $L = \Delta$ is known as Weyl's Lemma.

Corollary 2.43. (Weyl's Lemma)
The Laplace operator is hypoelliptic on $\mathbb{R}^n$.

Proof: N is a C^∞ function on $\mathbb{R}^n\backslash\{0\}$. ■

Note that this corollary yields regularity results on $\mathscr{D}'(\Omega)$ solutions to the Poisson problem. If Ω is an open subset of $\mathbb{R}^n$, $f \in C^\infty(\Omega)$ and $\Delta u = f$ in the sense of distributions then, automatically, $u \in C^\infty(\Omega)$. Since N is even analytic in $\mathbb{R}^n\backslash\{0\}$, one can show that Δ is *analytic –hypoelliptic*. The definition is similar to the one of hypoellipticity with C^∞ replaced by analyticity (see [**17**]).

Before we continue with the solution formula for general Dirichlet Laplace problems $\Delta u = 0$ in Ω and $u = g$ on $\partial\Omega$ we need some symmetry result on G.

Lemma 2.44. *Let $\Omega \subset \mathbb{R}^n$ be open and let G be the Green's function on Ω. Then, for any pair of points $x_1, x_2 \in \Omega$ with $x_1 \neq x_2$, we find*

$$G(x_1, x_2) = G(x_2, x_1).$$

Proof: Since x_1 is different from x_2, we can find $\epsilon > 0$ such that $B_{4\epsilon}(x_2) \subset \Omega$ and $x_1 \notin B_{4\epsilon}(x_2)$. We then construct a function $\phi \in \mathscr{D}(\Omega)$ which is equal to one on $B_{2\epsilon}(x_2)$ and $\phi = 0$ outside of $B_{3\epsilon}(x_2)$ so that

$$\begin{aligned} G(x_2, x_1) &= \phi(x_2)G(x_2, x_1) = \langle \delta(x - x_2), \phi(x)G(x, x_1)\rangle \\ &= \langle \Delta G(x, x_2), \phi(x)G(x, x_1)\rangle = \langle G(x, x_2), \Delta(\phi(x)G(x, x_1))\rangle . \end{aligned}$$

By construction, $x \mapsto G(x, x_1)$ is a C^∞–function on the support of ϕ. In particular, $x \mapsto G(x, x_1)$ is harmonic on that set. Thus

$$\begin{aligned} &\Delta(\phi(x)G(x, x_1)) \\ &\quad = \phi(x)\Delta G(x, x_1) + \big(G(x, x_1)\Delta\phi(x) + 2\nabla\phi(x)\cdot\nabla G(x, x_1)\big) \\ &\quad\quad = G(x, x_1)\Delta\phi(x) + 2\nabla\phi(x)\cdot\nabla G(x, x_1). \end{aligned}$$

Since this expression vanishes in $B_{2\epsilon}(x_2)$ (where $\phi \equiv 1$ so that the derivatives vanish), we can multiply it by a function $\psi \in \mathscr{D}(\Omega)$ which vanishes on $B_\epsilon(x_2)$ and is equal to one outside of $B_{2\epsilon}(x_2)$ in a set Ω_δ which contains x_1 and $B_{4\epsilon}(x_2)$. Thus

$$G(x_2, x_1) = \langle \psi(x)G(x, x_2), \Delta(\phi(x)G(x, x_1))\rangle$$

and since both expressions are test functions, we can invert the order

$$\begin{aligned} G(x_2, x_1) &= \langle \Delta(\phi(x)G(x, x_1)), \psi(x)G(x, x_2)\rangle \\ &= \langle \phi(x)G(x, x_1), \Delta(\psi(x)G(x, x_2))\rangle . \end{aligned}$$

If we now take a sequence $(\phi_n)_{n\in\mathbb{N}}$ with $\phi_n \in \mathscr{D}(\Omega\backslash\{x_1\}$ and $0 \le \phi \le \phi_n \le 1$ which converges point wise to one on Ω, we get with Lebesgue's dominated convergence

$$G(x_2, x_1) = \langle \phi_n(x)G(x, x_1), \Delta(\psi(x)G(x, x_2))\rangle \xrightarrow[n\to\infty]{} \langle G(x, x_1), \Delta(\psi(x)G(x, x_2))\rangle\,.$$

Using the property of the Green's function derived in Lemma 2.39, we obtain

$$G(x_2, x_1) = \langle \delta(x - x_1), \psi(x)G(x, x_2)\rangle = \psi(x_1)G(x_1, x_2) = G(x_1, x_2).$$

■

The following Lemma yields a second application of Green's functions.

Lemma 2.45. *Let Ω be a bounded domain in $\mathbb{R}^n$ with a piece-wise C^1 boundary. Let G be the Green's function on Ω which satisfies*

$$x \mapsto G(x, y) - N(x - y) \in C^1(\bar{\Omega}) \qquad \forall y \in \Omega.$$

Then, any harmonic function $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ satisfies

$$u(x) = \int_{\partial\Omega} u(y)\frac{\partial}{\partial n_y}G(x, y)\, d\sigma(y), \qquad x \in \Omega.$$

Proof: Since u is harmonic in Ω, we have

$$\int_\Omega G(x, y)\Delta u(y)\, dy = 0. \tag{2.30}$$

Multiplying (2.30) by a test function $\phi \in \mathscr{D}(\Omega)$ and integrating yields after changing the order of integration

$$0 = \int_\Omega \Delta u(y) \int_\Omega G(x, y)\phi(x)\, dx\, dy.$$

Since $G(x, y) = G(y, x)$, the inner integral

$$v(y) = \int_\Omega G(x, y)\phi(x)\, dx$$

is, according to Lemma 2.40, the solution of the problem

$$\begin{aligned} \Delta v &= \phi \qquad \text{in } \Omega \\ v &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{2.31}$$

Moreover, $v \in C^1(\bar{\Omega}) \cap C^\infty(\Omega)$ so that Green's formula can be applied. We find

$$0 = \int_\Omega v\Delta u\, dy = \int_{\partial\Omega} v\frac{\partial u}{\partial n}\, d\sigma(y) - \int_{\partial\Omega} u\frac{\partial v}{\partial n}\, d\sigma(y) + \int_\Omega u\Delta v\, dy.$$

Due to (2.31), we get

$$0 = -\int_{\partial\Omega} un(y) \cdot \nabla_y \int_\Omega G(x, y)\phi(x)\, dx\, d\sigma(y) + \int_\Omega u\phi\, dy.$$

Due to Lemma 2.40, the derivative ∂_y can be taken under the integral so that

$$0 = \int_\Omega \big(u(x) - \int_{\partial\Omega} u(y) \frac{\partial}{\partial n_y} G(y,x)\, d\sigma(y)\big)\phi(x)\, dx.$$

Using again the symmetry $G(x,y) = G(y,x)$, we obtain

$$u(x) = \int_{\partial\Omega} u(y) \frac{\partial}{\partial n_y} G(x,y)\, d\sigma(y), \qquad x \in \Omega.$$

■

This result shows that the solution to general Dirichlet Laplace problems can be obtained through an integral operator on $\partial\Omega$ with kernel $\frac{\partial}{\partial n_y} G(x,y)$, the so called *Poisson kernel.*

Lemma 2.46. *Let Ω be a bounded $C^{2,\alpha}$ domain in $\mathbb{R}^n$ with $\alpha > 0$. Further let $g \in C(\partial\Omega)$. If G is the Green's function on Ω, then*

$$u(x) = \int_{\partial\Omega} g(y) \frac{\partial}{\partial n_y} G(x,y)\, d\sigma(y), \qquad x \in \Omega$$

is the solution of

$$\begin{aligned} \Delta u &= 0 \qquad \text{in } \Omega \\ u &= g \qquad \text{on } \partial\Omega \end{aligned}$$

Proof: We first note that, for the regular domain under consideration, the Green's function has the smoothness properties required in Lemma 2.45. Indeed, $x \mapsto H(x,y) = G(x,y) - N(x-y)$ is a smooth function on $\partial\Omega$ so that with Theorem 2.18, $x \mapsto H(x,y)$ is contained in $C^{2,\alpha}(\bar\Omega)$. For the same reason, $g \in C^{2,\alpha}(\partial\Omega)$ gives rise to a unique solution u of the Dirichlet Laplace problem which is in $C^{2,\alpha}(\bar\Omega)$. Applying the previous Lemma we find the representation formula

$$u(x) = \int_{\partial\Omega} g(y) \frac{\partial}{\partial n_y} G(x,y)\, d\sigma(y), \qquad x \in \Omega$$

If g is only continuous, we go over to an approximating sequence $(g_n)_{n\in\mathbb{N}}$ of smooth functions which converges uniformly to g on $\partial\Omega$. Using the continuous dependence on the data (Theorem 2.8) we obtain convergence of the corresponding sequence of solutions $(u_n)_{n\in\mathbb{N}}$ to the solution of the limit problem. Applying Lebesgue's dominated convergence yields

$$\begin{aligned} u(x) &= \lim_{n\to\infty} u_n(x) \\ &= \lim_{n\to\infty} \int_{\partial\Omega} g_n(y) \frac{\partial}{\partial n_y} G(x,y)\, d\sigma(y) = \int_{\partial\Omega} g(y) \frac{\partial}{\partial n_y} G(x,y)\, d\sigma(y). \end{aligned}$$

■

For domains Ω with a particularly simple geometry, the Green's function can sometimes be guessed. For the half space $\Omega = \{x \in \mathbb{R}^n : x_1 > 0\}$, the Green's function is obtained by the *method of images.* To construct a harmonic function

$x \mapsto H(x,y)$ which has the values $-N(x-y)$ for x on the boundary $\partial\Omega$ we evaluate the fundamental solution at the *reflected* point

$$\hat{x} = (-x_1, x_2, \dots, x_n)^T.$$

More precisely, for any $y \in \Omega$, we set

$$H(x,y) = -N(\hat{x} - y), \qquad x \in \bar{\Omega}.$$

Then, for $x \in \partial\Omega$, i.e. $x = (0, x_2, \dots, x_n)^T$ we have $x = \hat{x}$ so that

$$G(x,y) = N(x-y) + H(x,y) = N(x-y) - N(\hat{x} - y) = 0$$

which is the second condition in Definition 2.38. The other requirements follow easily from the fact that $|\hat{x} - y| = |x - \hat{y}|$ which implies that $x \mapsto H(x,y) = N(x - \hat{y})$ is harmonic in $\bar{\Omega} \subset \mathbb{R}^n \backslash \{\hat{y}\}$. For the example of a ball in $\mathbb{R}^n$ and the quarter plane in 2D see Exercises 2.18 and 2.19. We remark that the Poisson kernel of the Green's function on the ball has already been investigated in Theorem 1.14.

Instead of the Dirichlet problem one can also consider other boundary conditions with Green's functions. For the Laplace operator and Neumann conditions only 2) in Definition 2.38 has to be changed to $\frac{\partial}{\partial n_x} G(x,y) = 0$ for $x \in \partial\Omega$ and $y \in \Omega$ (plus some extra smoothness on $G(x,y) - N(x-y)$). The construction in particular cases is done similar to the reflection principle above by placing a reflected fundamental solution outside the domain without changing the sign.

We conclude our considerations with the remark that Green's functions can be helpful in special situations where they are easily constructed. Otherwise, the construction of G involves the solution of infinitely many Dirichlet (or Neumann) Laplace problems (for each $y \in \Omega$). In these general cases, showing existence of the Green's function (including smoothness requirements) is as complicated as showing these properties directly for specific Dirichlet (or Neumann) Poisson problems. Thus, Green's functions are, in general, not a means to show existence or regularity of solutions to Dirichlet or Neumann problems. They nicely show, however, that the solution operators to general Dirichlet or Neumann Poisson problems are, in fact, integral operators with weakly singular kernels.

2.9. Exercises

Exercise 2.1. A function $u \in C^0(\bar{\Omega})$ is called *subharmonic*, if for every ball B with $\bar{B} \subset \Omega$ and every function $h \in C(\bar{B})$ with h harmonic in B and $u \le h$ on ∂B, we have $u \le h$ in B. Replacing $\le$ by $\ge$, we also define *superharmonic* functions.

a) Show that this definition generalizes the one from Exercise 1.6.
b) Show that sum and maximum of two subharmonic functions are subharmonic.

c) Show the maximum principle for subharmonic functions: if u is subharmonic with $u \leq 0$ on $\partial\Omega$, then either $u \equiv 0$ or $u < 0$ in Ω, if Ω is connected.

Exercise 2.2. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. On its complement, we consider the (exterior) Dirichlet problem

$$Lu = f \quad \text{in } \Omega^c, \qquad\qquad u = g \quad \text{on } \partial\Omega$$

with continuous functions f and g and a general linear, second order operator L whose zeroth order part is non negative. The boundary condition at infinity is

$$|u(x)| \to 0 \qquad \text{for } |x| \to \infty.$$

Show that the solution to this problem is unique if it exists.

Exercise 2.3. Let Ω be open, bounded and connected with a piecewise C^1 boundary. Further, let $f \in C(\bar{\Omega})$ and $g, g_0, g_1, a \in C(\partial\Omega)$ with $a > 0$. Show that the following problems have at most one solution $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$:

$$\begin{cases} \Delta u = f & \text{in } \Omega \\ \dfrac{\partial u}{\partial n} + au = g & \text{on } \partial\Omega \end{cases}$$

$$\begin{cases} \Delta u = f & \text{in } \Omega \\ u = g_0 & \text{on } \emptyset \neq \Gamma_0 \subset \partial\Omega \\ \dfrac{\partial u}{\partial n} = g_1 & \text{on } \Gamma_1 = \partial\Omega \backslash \Gamma_0 \end{cases}$$

What about the case $\Gamma_0 = \emptyset$?

Exercise 2.4. Show that, in contrast to Δ, the biharmonic operator Δ^2 does not satisfy a maximum principle.

Exercise 2.5. Let u be a classical solution of $\Delta u = u^3 - u$ on a bounded domain Ω. Assume that $u = 0$ on $\partial\Omega$. Show that $u \in [-1, 1]$ throughout Ω. Can the values ± 1 be achieved?

Exercise 2.6. Show that convex sets and sets with a C^2 boundary satisfy the exterior sphere property. A set $\Omega \subset \mathbb{R}^n$ has a C^2–boundary, if the boundary can be locally described by C^2 mappings. More precisely, for any point $x \in \partial\Omega$ there should be some $r > 0$, a rotation matrix R and $\phi \in C^2(B_r(0; \mathbb{R}^{n-1}))$ such that the rotated and translated set $R(\Omega - x)$ can be written as

$$R(\Omega - x) \cap B_r(0) = \{\, (y, z) \in B_r(0) \,:\, y \in \mathbb{R}^{n-1},\ |y| < r,\ z < \phi(y) \,\}.$$

Exercise 2.7. Consider the general linear, second order, elliptic Dirichlet problem

$$\begin{aligned} \tilde{L}u &= \tilde{f} \qquad \text{in } [\alpha, \beta] \\ u &= \tilde{g} \qquad \text{on } \{\alpha, \beta\} \end{aligned}$$

where $\tilde{L} = -au'' + \tilde{b}u' + \tilde{c}u$. Show that under the assumption $a \in C^0([\alpha, \beta])$ with $\min_{[\alpha,\beta]} a > 0$, the problem can be reduced to

$$\begin{aligned} -v'' + bv' + cv &= f \qquad \text{in } [0,1] \\ v &= g \qquad \text{on } \{0,1\} \end{aligned}$$

Exercise 2.8. Show that the difference scheme, defined by $\mathcal{L}_h$ in the lecture, is consistent in every $u \in C^{2+k,\alpha}(\bar{\Omega})$ with $k \in \mathbb{N}_0$ and $\alpha > 0$. The order of consistency is $\min\{2, k+\alpha\}$. ($C^{2+k,\alpha}(\bar{\Omega})$ means that the derivative $2+k$ is Hölder continuous of order α up to the boundary.)

Exercise 2.9.

1) Show that positive multiples, sums and products of positive matrices are again positive.
2) Show that for any positive matrix $B \geq 0$, we have $By \geq Bx$ whenever $y \geq x$.
3) How does the set of positive vectors in $\mathbb{R}^3$ look like?
4) What is the shape of order intervals in $\mathbb{R}^3$?

Exercise 2.10. Show that the assumption $c \geq 0$ can, in general, not be removed from the assumptions for the maximum principle. (Consider the operator $Lu = -u'' - u$ on $(0, \pi)$.)

Exercise 2.11. Let $\Omega = (0, \pi)$, $Lu = -u''$, $Bu = u|_{\partial\Omega}$ and $\mathcal{L} = \binom{L}{B}$.

a) Show that the eigensolutions $u^{(m)}(x) = \sin(mx)$, $m \in \mathbb{N}$ of $\mathcal{L}$ turn into eigenvectors of $\mathcal{L}_h$ under discretization.
b) What is the relation between the discretizations of $u^{(m)}$ and $u^{(m+n)}$ if $h = \pi/n$?
c) Give a list of all eigenvalues of $\mathcal{L}_h$.
d) How do the eigenvalues behave for $h \to 0$?
e) Estimate the condition number $\|\mathcal{L}_h\| \cdot \|\mathcal{L}_h^{-1}\|$ of $\mathcal{L}_h$ for small h.
f) The problem $Lu = u$ on Ω and $u = 0$ on $\partial\Omega$ is not uniquely solvable. What about the corresponding discretized problem?

Exercise 2.12. The matrix $A \in \mathbb{R}^{N \times N}$ is called

1) L_0–matrix, if $a_{ij} \leq 0$ for all $i \neq j$,
2) M–matrix, if A is an L_0–matrix, A^{-1} exists, and A^{-1} is positive.

Show that the L_0–matrix A is an M–matrix if and only if there exists $y \gg 0$ such that $Ay \gg 0$. Moreover

$$\|A^{-1}\|_\infty \leq \frac{\|y\|_\infty}{\min_k (Ay)_k}.$$

Hint: Consider the stability proof for the 1D finite difference scheme.

Exercise 2.13. Construct a fourth order consistent approximation for the second derivative in 1D.

Exercise 2.14. Let $\Omega_h \cup \Gamma_h$ be the discretization of the unit square $\bar{\Omega}$ and let $\mathcal{L}_h$ be a consistent discretization of the continuous operator $\mathcal{L} = \binom{L}{B}$ (with general elliptic operator $L = -\sum a_{ij}\partial^2/(\partial x_i \partial x_j) + \sum b_i \partial/\partial x_i + c$ where a_{ij}, b, c are continuous, $c \geq 0$, (a_{ij}) positive definite on $\bar{\Omega}$ and B being the trace operator on $\partial\Omega$). Show that, for h sufficiently small, there exists a vector $y \gg 0$ such that $\mathcal{L}_h y \gg 0$ and that there are h–independent constants C, μ so that

$$\|y\|_{h,\infty} \leq C < \infty, \qquad \min_i (\mathcal{L}_h y)_i \geq \mu > 0.$$

Hint: Use $v(x) = \beta - \exp(-\alpha |x - x_0|^2)$.

Exercise 2.15. Describe the matrix $\mathcal{L}_h$ which results from the standard discretization of the Poisson problem on the unit square with Neumann boundary conditions, if half–sided (respectively central) differences are used for the normal derivatives. Show that there is no vector $y \gg 0$ such that $\mathcal{L}_h y \gg 0$.

Exercise 2.16. Implement the finite difference scheme presented in the lecture for one–dimensional equations using the SOR algorithm.

a) Apply it to $-u''(x) = \sin(x)$ on $(0, \pi)$ with $u(0) = u(\pi) = 0$ for several discretizations and compare with the exact solution. Give a doubly logarithmic plot of the error versus h to numerically validate second order convergence.
b) Apply it to $-u''(x) + bu'(x) = 1$ on $(0, 1)$ with $u(0) = u(1) = 0$ and b constant. Check, whether dominating convection $b \gg 1$ influences the speed of convergence of the SOR algorithm. Plot the number of cycles necessary to reach a prescribed error compared with the exact solution for different values of b. Repeat this for different values of h.

Exercise 2.17. Construct fundamental solutions for the operators $\Delta + 1$ and $\Delta - 1$ on $\mathbb{R}^3$. (Hint: Look for a spherically symmetric function of the form $u(x) = f(|x|) = \hat{f}(|x|)/|x|$ with a nice behavior at infinity.)

a) Extend the result to the case $\Delta \pm \omega^2$, $\omega > 0$ by a simple transformation.
b) Show that the so called *Helmholtz operator* $\Delta + \omega^2$ naturally appears if the wave equation $\partial^2 u/\partial^2 t = \Delta u$ is solved by Fourier transformation (in the time variable t).
c) Show that a fundamental solution of the Laplacian is recovered in the case $\omega = 0$.

Exercise 2.18. Construct a Green's function for the Dirichlet Poisson problem

$$\begin{aligned} \Delta u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned}$$

with the help of the method of images if Ω is the quarter plane

$$\Omega = \{x \in \mathbb{R}^2 \, : \, x_1, x_2 > 0\}.$$

How does the Green's function for the corresponding problem with Neumann conditions ($\partial u/\partial n = 0$ on $\partial\Omega$) look like?

Exercise 2.19. Construct a Green's function for the Dirichlet Poisson problem

$$\begin{aligned} \Delta u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned}$$

with the help of the method of images if Ω is the unit ball in $\mathbb{R}^n$. (Hint: use the reflection $\hat{x} = x/\|x\|^2$ and a suitable scaling of the reflected fundamental solution.) Show that the Poisson kernel $P(x, y) = n(y) \cdot \nabla_y G(x, y)$ for $y \in \partial B_1(0)$ is the one used in Theorem 1.14.

CHAPTER 3

Weak solutions

By generalizing the classical solution concept it is possible to treat a larger class of problems (i.e. step functions as data, complicated domains). Moreover, since less regularity is required of the solution, existence proofs are simplified. Finally, numerical methods can be constructed which converge with less assumptions on the regularity of the solution as required, for example, for finite difference schemes. Of course, we impose the natural condition that generalized solutions coincide with classical solutions whenever the latter exist (otherwise, the adjective *generalized* would not be justified). In many cases, however, the generalized solution does not have enough regularity to be classical.

We remark that weak formulations of elliptic problems have already been encountered in Section 2.8 where we considered, for example, $\Delta u = f$ in the sense of distributions on Ω. The idea is to multiply the equation with a test function $\phi \in C_0^\infty(\Omega)$ and integrate by parts (or rather to take the dual product and apply the definition of distributional derivatives). This gives rise to the weak formulation of the Poisson equation

$$\langle u, \Delta\phi\rangle = \langle f, \phi\rangle \,, \qquad \forall \phi \in \mathscr{D}(\Omega).$$

The formulation is really a generalization because it reduces to the classical one if the function u is sufficiently smooth. However, it is not at all clear, how boundary conditions like $u = 0$ on $\partial\Omega$ can be interpreted in the distributional setup. A possible way out is to restrict the class of distributions which are considered as possible solutions. A very much investigated family of subclasses of $\mathscr{D}'(\Omega)$ are the so called *Sobolev spaces* which turn out to be natural function spaces to be used for the weak formulation of elliptic problems. In the appendix, we briefly introduce these spaces and present a few concepts which are of importance when dealing with boundary conditions.

3.1. Introductory examples

3.1.1. The Dirichlet Helmholtz problem. We start our considerations with the *Dirichlet Helmholtz problem*

$$\begin{aligned} -\Delta u + u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{3.1}$$

To obtain a weak formulation of (3.1) we first use a formal argument, i.e. we assume sufficient regularity of all participating objects so that all steps of the derivation are justified.

Multiplying the Helmholtz equation by a function v and integrating over Ω yields

$$-\int_\Omega v\Delta u\,dx + \int_\Omega vu\,dx = \int_\Omega vf\,dx.$$

Using the Gauß theorem for the first integral then leads to

$$\int_\Omega \nabla v\cdot\nabla u\,dx + \int_\Omega vu\,dx = \int_\Omega vf\,dx + \int_{\partial\Omega} v\frac{\partial u}{\partial n}\,d\sigma.$$

Now, we observe that the left hand side is nothing but the $H^1(\Omega)$ scalar product and that the integral involving f is the $\mathbb{L}^2$–product

$$\langle u,v\rangle_1 = \langle f,v\rangle_0 + \int_{\partial\Omega} v\frac{\partial u}{\partial n}\,d\sigma. \tag{3.2}$$

To get rid of the boundary integral, we assume that v vanishes on $\partial\Omega$. Also, the Dirichlet boundary condition on u is incorporated by the requirement $Bu = 0$. By choosing u and v from spaces for which (3.2) makes sense, we end up with the following generalization of (3.1):

Definition 3.1. *Let $\Omega\subset\mathbb{R}^n$ be open and $f\in\mathbb{L}^2(\Omega)$. A function $u\in H_0^1(\Omega)$ is called* weak solution *of* (3.1) *if*

$$\langle u,v\rangle_1 = \langle f,v\rangle_0 \qquad \forall v\in H_0^1(\Omega).$$

If we assume that Ω is a bounded $C^{0,1}$ domain, then the derivation already indicates that classical solutions of the Helmholtz problem are automatically weak solutions. It is important to note, however, that the converse is also true if the weak solution is regular enough. To see this, we assume that $u\in C^2(\Omega)\cap C(\bar\Omega)$ is a weak solution. Since, by definition, $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$ we have

$$\langle -\Delta u + u - f, \phi\rangle = 0, \qquad \forall\phi\in C_0^\infty(\Omega)$$

which means $-\Delta u + u = f$ in the sense of distributions. Due to the continuity of $-\Delta u + u$ this implies $f\in C(\Omega)$ and $-\Delta u + u = f$ point wise. The Dirichlet boundary condition is recovered since u is, by assumption, an element of $H_0^1(\Omega)\cap C(\bar\Omega)$ (see Theorem C.4).

If we want to treat more general Dirichlet conditions $u = g$ on $\partial\Omega$ we use a standard trick: we assume the existence of some function u_0 on Ω which has the correct boundary values $u_0 = g$ on $\partial\Omega$. Then, $u - u_0$ has again zero boundary values and satisfies the equation

$$(\Delta+1)(u-u_0) = f - (\Delta+1)u_0 = \hat f$$

which is again of Helmholtz type. For the weak solutions, we proceed exactly in the same way. To guarantee that there is some function u_0 on Ω which has g as boundary values we require that $g\in B(H^1(\Omega))$ (which is $H^{\frac12}(\partial\Omega)$

under suitable assumptions on $\partial\Omega$). Then, there exists $u_0 \in H^1(\Omega)$ such that $Bu_0 = g$ and $u - u_0$ should satisfy

$$\langle u - u_0, v\rangle_1 = \langle f, v\rangle_0 - \langle u_0, v\rangle_1 \qquad \forall v \in H_0^1(\Omega).$$

Definition 3.2. *Let $\Omega \subset \mathbb{R}^n$ be a bounded $C^{0,1}$ domain, $f \in \mathbb{L}^2(\Omega)$ and let $g \in \mathbb{L}^2(\partial\Omega)$ be the trace of some function $u_0 \in H^1(\Omega)$. Then, $u \in H^1(\Omega)$ is called weak solution of $-\Delta u + u = f$ in Ω and $u = g$ on $\partial\Omega$ if $\hat{u} = u - u_0$ satisfies*

$$\langle \hat{u}, v\rangle_1 = \langle f, v\rangle_0 - \langle u_0, v\rangle_1 \qquad \forall v \in H_0^1(\Omega).$$

3.1.2. The Neumann Helmholtz problem. We again consider the Helmholtz equation but this time with a Neumann condition

$$\begin{aligned} -\Delta u + u &= f \qquad \text{in } \Omega \\ \frac{\partial u}{\partial n} &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{3.3}$$

Again, multiplying with a test function and integration by parts yields (3.2) where now the boundary integral vanishes regardless of the trace of v. This leads to the following generalization of (3.3):

Definition 3.3. *Let $\Omega \subset \mathbb{R}^n$ be open and $f \in \mathbb{L}^2(\Omega)$. A function $u \in H^1(\Omega)$ is called* weak solution *of* (3.3) *if*

$$\langle u, v\rangle_1 = \langle f, v\rangle_0 \qquad \forall v \in H^1(\Omega).$$

Again, if Ω has a C^1 boundary (so that normal derivatives can be defined), any classical solution of (3.3) is also a weak solution. If $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ is a weak solution, we find again $-\Delta u + u = f$ point wise by taking test functions $v \in C_0^\infty(\Omega) \subset H^1(\Omega)$. For $\phi \in C_0^\infty(\mathbb{R}^n)$ we get, since $\phi|_\Omega \in H^1(\Omega)$,

$$\begin{aligned} 0 &= \langle u, \phi\rangle_1 - \langle f, \phi\rangle_0 \\ &= \int_\Omega \nabla u \cdot \nabla \phi \, dx + \int_\Omega u\phi \, dx - \int_\Omega f\phi \, dx \\ &= \int_\Omega (-\Delta u + u - f)\phi \, dx + \int_{\partial\Omega} \phi \frac{\partial u}{\partial n} \, d\sigma. \end{aligned}$$

Since $-\Delta u + u - f = 0$ point wise, we obtain

$$\int_{\partial\Omega} \phi \frac{\partial u}{\partial n} \, d\sigma = 0 \qquad \forall \phi \in C_0^\infty(\mathbb{R}^n). \tag{3.4}$$

Since the restriction of $C_0^\infty(\mathbb{R}^n)$ functions to Ω is dense in $H^1(\Omega)$ (see Theorem C.3) and since the boundary values of $H^1(\Omega)$ functions are dense in $\mathbb{L}^2(\partial\Omega)$ (Theorem C.4), we conclude that $\partial u/\partial n$ is orthogonal to a dense subset of $\mathbb{L}^2(\partial\Omega)$ and thus has to be zero.

We remark that the weak formulation in Definition 3.3 does neither require the existence of an outward normal on $\partial\Omega$ nor $u \in H^2(\Omega)$ so that $\partial u/\partial n$ is, in general, not well defined. Nevertheless, if boundary regularity of the weak solution can be shown, our previous argument proves that the Neumann condition is satisfied.

As for the Dirichlet problem, let us also consider the case of non–homogeneous Neumann conditions $\partial u/\partial n = g$ on $\partial\Omega$. Repeating the derivation, we again arrive at (3.2) where now the Neumann condition is inserted giving rise to the boundary integral

$$\int_{\partial\Omega} v\frac{\partial u}{\partial n}\,d\sigma = \int_{\partial\Omega} vg\,d\sigma = \langle g, Bv\rangle_{0,\partial\Omega}\,.$$

Definition 3.4. *Let $\Omega \subset \mathbb{R}^n$ be a bounded $C^{0,1}$ domain, $f \in \mathbb{L}^2(\Omega)$ and let $g \in \mathbb{L}^2(\partial\Omega)$. Then, $u \in H^1(\Omega)$ is called weak solution of $-\Delta u + u = f$ in Ω and $\partial u/\partial n = g$ on $\partial\Omega$ if*

$$\langle \hat{u}, v\rangle_1 = \langle f, v\rangle_0 + \langle g, Bv\rangle_{0,\partial\Omega} \qquad \forall v \in H^1(\Omega).$$

3.1.3. Existence, Uniqueness, Continuous Dependence – the Riesz Lemma. Since the weak formulations of the Helmholtz problem with Dirichlet and Neumann conditions are very similar, we consider the following abstract problem:

Given a Hilbert space V and an element of the dual $l \in V'$, find $u \in V$ such that

$$\langle u, v\rangle_V = l(v) \qquad \forall v \in V. \tag{3.5}$$

We have the important result

Theorem 3.5. *Problem (3.5) admits a unique solution $u \in V$ which depends continuously on the data, i.e. $\|u\|_V = \|l\|_{V'}$.*

Proof: The proof is a simple application of the Riesz representation Theorem (which will be proven below for completeness). ■

By taking $V = H^1_0(\Omega)$, we recover the Dirichlet and with $V = H^1(\Omega)$ the Neumann problem. In the homogeneous cases (Definitions 3.1 and 3.3) the linear functional l is given by

$$l(v) = \langle f, v\rangle_0\,, \qquad f \in \mathbb{L}^2(\Omega).$$

We see here, that we have been unnecessarily restrictive in deriving a weak formulation. Instead of requiring $f \in \mathbb{L}^2(\Omega)$ we could as well have assumed $f \in H^{-1}(\Omega)$ (the dual of $H^1_0(\Omega)$) for the Dirichlet and $f \in H^1(\Omega)'$ for the Neumann problem. Then, the integral $\langle f, v\rangle_0$ is just interpreted as a linear functional acting on v (as in the theory of generalized functions). As far as f is concerned, the same holds also for the non–homogeneous boundary conditions. In the Dirichlet case, the linear functional is

$$l(v) = \langle f, v\rangle_0 - \langle u_0, v\rangle_1$$

and for the Neumann problem

$$l(v) = \langle f, v\rangle_0 + \langle g, Bv\rangle_{0,\partial\Omega}\,. \tag{3.6}$$

Note that $v \mapsto \langle g, Bv\rangle_{0,\partial\Omega}$ is a continuous linear functional on $H^1(\Omega)$ since $B : H^1(\Omega) \mapsto \mathbb{L}^2(\partial\Omega)$ is a bounded linear mapping. We also remark that the condition $g \in \mathbb{L}^2(\partial\Omega)$ in the Neumann case could be relaxed if we used that $B(H^1(\Omega)) = H^{\frac{1}{2}}(\partial\Omega)$ has higher regularity than $\mathbb{L}^2(\partial\Omega)$. Indeed, $\langle g, Bv\rangle_{0,\partial\Omega}$ could be replaced by a linear functional $\hat{l}$ from the dual of $H^{\frac{1}{2}}(\partial\Omega)$ acting on Bv without loosing the property that $v \mapsto \hat{l}(Bv)$ is a continuous linear functional on $H^1(\Omega)$.

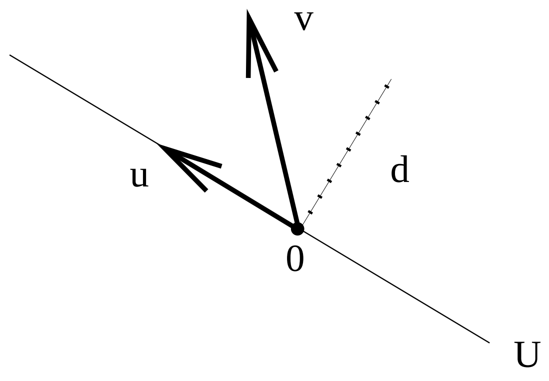
For completeness, we prove the Riesz representation theorem. It will also give a geometrical interpretation of the weak solution concept in Hilbert spaces. We begin with the important *Projection Theorem.*

Theorem 3.6. *Let V be a Hilbert space and let U be a closed subspace of V. Then every $v \in V$ has a unique decomposition $v = u + w$, where $u \in U$ and $w \in U^\perp$. ($U^\perp$ is the orthogonal complement of U, i.e. the set of all vectors which are orthogonal to U.)*

Proof: At the core of the proof is the so called *parallelogram law* which relates the length of the edges a, b of a parallelogram with those of the diagonals

$$\frac{1}{2}\left(\|a+b\|^2 + \|a-b\|^2\right) = \|a\|^2 + \|b\|^2.$$

(We skip the subscript V at the norm and the scalar product.) Motivated by finite dimensional Euclidean geometry, we expect that the orthogonal projection of $v \in V$ on to the subspace U is given by the element with minimal distance to v.



We define

$$d := \inf_{x \in U} \|x - v\|$$

as minimal distance and pick a minimizing sequence $(u_n)_{n\in\mathbb{N}} \subset U$ such that $d_n = \|u_n - v\| \to d$. Applying the parallelogram law with $a = u_n - v$ and

$b = u_m - v$ we get

$$\begin{aligned}\|u_n - u_m\|^2 &= 2\|u_n - v\|^2 + 2\|u_m - v\|^2 - \|u_n + u_m - 2v\|^2 \\ &= 2d_n^2 + 2d_m^2 - 4\|\frac{u_n + u_m}{2} - v\|^2.\end{aligned}$$

Since $(u_n + u_m)/2 \in U$, we have by definition that

$$\|\frac{u_n + u_m}{2} - v\| \geq d$$

and thus

$$\|u_n - u_m\|^2 \leq 2(d_n^2 - d^2) + 2(d_m^2 - d^2)$$

which shows that $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Since U is closed, the sequence converges to an element $u \in U$ which satisfies $\|u - v\| = d$. Consequently, for any $x \in U$, the function

$$f(t) = \|u + tx - v\|^2, \qquad t \in \mathbb{R}$$

has a minimum at $t = 0$. The condition $f'(0) = 0$ then yields

$$\langle v - u, x\rangle = 0, \qquad \forall x \in U \tag{3.7}$$

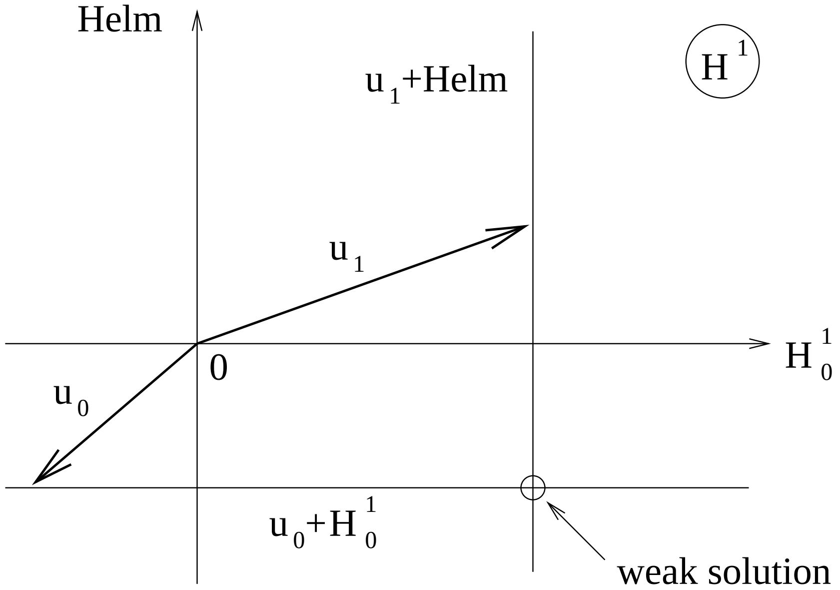
which shows that $w = v - u$ is an element of $U^\perp$.

Uniqueness of the decomposition follows easily from (3.7). If $\hat{u}$ is another vector such that $\langle v - \hat{u}, x\rangle = 0$ for all $x \in U$, the difference with (3.7) yields

$$\langle u - \hat{u}, x\rangle = 0 \qquad \forall x \in U.$$

In particular, for $x = u - \hat{u}$ we find $\|u - \hat{u}\|^2 = 0$ which implies $u = \hat{u}$. ■

Now, we can geometrically interpret the concept of weak solutions for the Helmholtz problem with general Dirichlet boundary conditions. If $u_0 \in H^1(\Omega)$ has the required boundary values g, the weak solution u is restricted to the affine linear subspace $u_0 + H_0^1(\Omega)$ which contains all functions in $H^1(\Omega)$ which have g as trace.



Conversely, if we just consider the Helmholtz equation $-\Delta u + u = f$ without boundary conditions, again the set of solutions makes up an affine linear subspace of $H^1(\Omega)$. If u_1 is a special solution then the solution manifold is $u_1 + \mathrm{Helm}(\Omega)$ where $\mathrm{Helm}(\Omega)$ contains all homogeneous solutions $-\Delta u + u = 0$. In the weak formulation, boundary conditions are neglected by using test functions which are compactly supported in Ω. Thus

$$\mathrm{Helm}(\Omega) = \left\{ u \in H^1(\Omega) \,:\, -\Delta u + u = 0 \quad \text{in } \mathscr{D}'(\Omega) \right\}. \tag{3.8}$$

It is easy to see that $\mathrm{Helm}(\Omega)$ is really the orthogonal complement of $H_0^1(\Omega)$ as indicated in the figure. If $u \in H_0^1(\Omega)^\perp$, then $\langle u, \phi \rangle_1 = 0$ for all $\phi \in \mathscr{D}(\Omega)$ since, by definition, test functions are dense in $H_0^1(\Omega)$. An integration by parts then yields $\langle -\Delta u + u, \phi \rangle = 0$ in the sense of distributions so that $u \in \mathrm{Helm}(\Omega)$. Conversely, if $u \in \mathrm{Helm}(\Omega)$ then u is, by definition, orthogonal to a dense subset of $H_0^1(\Omega)$ and thus to $H_0^1(\Omega)$ itself.

We conclude that the space $H^1(\Omega)$ is the orthogonal sum of two closed subspaces $H_0^1(\Omega)$ and $\mathrm{Helm}(\Omega)$. One of these subspaces takes care of the boundary conditions, the other one of the partial differential equation. A specific problem gives rise to affine linear spaces $u_0 + H_0^1(\Omega)$ and $u_1 + \mathrm{Helm}(\Omega)$. The first one contains all elements with the correct boundary value problems, the second one is the collection of all solutions of the (inhomogeneous) PDE. The solution of the boundary value problem is then naturally given by the intersection of the two spaces which is a single point (due to orthogonality). In other words, the weak solution is the projection of $u_1 + \mathrm{Helm}(\Omega)$ onto the correct boundary values $u_0 + H_0^1(\Omega)$ (or, vice versa, the projection of $u_0 + H_0^1(\Omega)$ on the solution space $u_1 + \mathrm{Helm}(\Omega)$). Geometrically (in 2D) it is clear that a slight change of u_0 and u_1 (the data of the problem) only changes the intersection point (the solution) a little bit. This is a geometrical manifestation of the continuous dependence on the data.

Finally, let us use the projection theorem to prove the Riesz representation theorem.

Theorem 3.7. *The dual space of a Hilbert space is isometric to the Hilbert space itself. In particular, for every $u \in V$ the linear functional on V defined by*

$$l_u(v) := \langle u, v \rangle_V$$

is bounded with norm $\|l_u\|_{V'} = \|u\|_V$. Moreover, for every $l \in V'$, there exists a unique $u \in V$ such that

$$l(v) = \langle u, v \rangle_V \qquad \forall v \in V,$$

and, furthermore, $\|u\|_V = \|l\|_{V'}$.

Proof: Boundedness of l_u is obtained with the Cauchy–Schwarz inequality

$$|l_u(v)| = |\langle u, v \rangle| \leq \|u\| \, \|v\|$$

which shows that $\|l_u\| \leq \|u\|$. Equality follows upon taking $v = u$. (We again neglect subscripts on norms and scalar products.) Conversely, if $l \in V'$ is given, we define $U = \ker l$ which is a closed subspace of V. For $l \neq 0$ (otherwise $u = 0$ is the unique representation), $U^\perp$ is one dimensional. Indeed, if w_1, w_2 are vectors in $U^\perp$ then

$$l(w_2)w_1 - l(w_1)w_2 \in U \cap U^\perp = \{0\}$$

which shows that w_1 and w_2 are linearly dependent. (Here we have used the projection theorem which shows that V is the orthogonal sum of U and $U^\perp$ if U is a closed subspace.) Now, let $\hat{u} \in U^\perp$ with $\|\hat{u}\| = 1$. Then for $v \in V$

$$v = \langle \hat{u}, v \rangle \, \hat{u} + (v - \langle \hat{u}, v \rangle \, \hat{u})$$

is an orthogonal decomposition with $\langle \hat{u}, v \rangle \, \hat{u} \in U^\perp$ so that $v - \langle \hat{u}, v \rangle \, \hat{u} \in U = \ker l$ which gives

$$l(v) = \langle \hat{u}, v \rangle \, l(\hat{u}) = \langle l(\hat{u})\hat{u}, v \rangle .$$

Setting $u = l(\hat{u})\hat{u}$ we thus have $l = l_u$. The vector u is clearly unique for if $\bar{u}$ is another such vector then

$$\langle \bar{u} - u, v \rangle = 0 \qquad \forall v \in V$$

which implies $\bar{u} = u$ (by setting $v = \bar{u} - u$). We have already seen that $\|l_u\| = \|u\|$ which completes the proof. ∎

The isometry which assigns to a functional the corresponding Riesz representation is called *Riesz operator.*

Corollary 3.8. *Let V be a Hilbert space. There exists an isometry $\mathcal{R} : V' \mapsto V$ (i.e. $\mathcal{R}$ and $\mathcal{R}^{-1}$ are bounded linear operators with $\|\mathcal{R}l\| = \|l\|$ for all $l \in V'$) which is uniquely determined by the condition*

$$\langle \mathcal{R}l, v \rangle_V = l(v) = \langle l, v \rangle_{V' \times V} \qquad \forall v \in V, \, l \in V'.$$

$\mathcal{R}$ is called Riesz operator.

3.1.4. The Dirichlet Poisson problem. In the previous example we have seen how problems with inhomogeneous Dirichlet conditions can be reduced to the homogeneous case. Therefore, we will only consider

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{3.9}$$

As before, the weak formulation is obtained by multiplying the equation with a test function and integrating by parts. We find

$$\langle \nabla u, \nabla v \rangle_0 = \langle f, v \rangle_0 + \int_{\partial\Omega} v \frac{\partial u}{\partial n}\, d\sigma.$$

Taking u and v from $H_0^1(\Omega)$ we obtain

Definition 3.9. *Let $\Omega \subset \mathbb{R}^n$ be open and $f \in \mathbb{L}^2(\Omega)$. A function $u \in H_0^1(\Omega)$ is called* weak solution *of* (3.9) *if*

$$\langle \nabla u, \nabla v \rangle_0 = \langle f, v \rangle_0 \qquad \forall v \in H_0^1(\Omega).$$

In order to apply Theorem 3.5 we first have to show that $V = H_0^1(\Omega)$ together with the scalar product $\langle u, v \rangle_V = \langle \nabla u, \nabla v \rangle_0$ is a Hilbert space (i.e. that $\langle u, u \rangle_V = 0$ implies $u = 0$ – and that V Cauchy sequences converge in the V norm). This is accomplished with the well known Poincaré inequality. An application of Theorem 3.5 then shows that the weak formulation of (3.9) admits a unique solution which depends continuously on the data if Ω is bounded in at least one direction.

Theorem 3.10. **(Poincaré inequality)**
Let $\Omega \subset \mathbb{R}^n$ be an open set which can be enclosed between two parallel hyperplanes of distance $d > 0$. Then, the $H^m(\Omega)$ seminorm

$$|u|_{m,2} = \left(\sum_{|\alpha|=m} \langle \nabla^\alpha u, \nabla^\alpha u \rangle_0 \right)^{\frac{1}{2}}$$

is equivalent to $\|\cdot\|_{m,2}$ on $H_0^m(\Omega)$. In particular, there exists a constant $C > 0$ which just depends on m and d such that

$$C\|u\|_{m,2} \le |u|_{m,2} \le \|u\|_{m,2} \qquad \forall u \in H_0^m(\Omega)$$

and $H_0^m(\Omega)$ is a Hilbert space with the scalar product

$$\langle u, v \rangle := \sum_{|\alpha|=m} \langle \nabla^\alpha u, \nabla^\alpha v \rangle_0 .$$

Proof: We assume, without loss of generality, that Ω is contained in the strip $\{x \in \mathbb{R}^n \,:\, 0 \le x_1 \le d\}$. For $u \in C_0^\infty(\Omega)$ we obtain with integration by parts

$$\|u\|_{0,2}^2 = \int_\Omega 1 \cdot u^2\, dx = -\int_\Omega x_1 2u \frac{\partial u}{\partial x_1}\, dx \le 2d \int_\Omega |u| \left| \frac{\partial u}{\partial x_1} \right| dx.$$

Using Cauchy–Schwarz inequality

$$\|u\|_{0,2}^2 \le 2d\|u\|_{0,2}\|\frac{\partial u}{\partial x_1}\|_{0,2}$$

and after dividing by $\|u\|_{0,2}$

$$\|u\|_{0,2} \le 2d\|\frac{\partial u}{\partial x_1}\|_{0,2}.$$

Applying this result to $\nabla^\beta u$ with $|\beta| < m$ and $u \in H_0^m(\Omega)$ we get (since $\nabla^\beta u \in H_0^1(\Omega)$)

$$\|\nabla^\beta u\|_{0,2} \le 2d\|\nabla^{\beta+e_1}u\|_{0,2}.$$

Then, by repeating the argument for $\beta + e_1$ (if $|\beta| + 1 < m$) we arrive at

$$\|\nabla^\beta u\|_{0,2} \le (2d)^{m-|\beta|}\|\nabla^{\beta+(m-|\beta|)e_1}u\|_{0,2} \le (2d)^{m-|\beta|}|u|_{m,2}.$$

Summing up over all $|\beta| < m$ then yields

$$\|u\|_{m,2}^2 \le \left(1 + \sum_{|\beta|<m}(2d)^{2(m-|\beta|)}\right)|u|_{m,2}^2$$

which completes the proof since the other estimate $|u|_{m,2} \le \|u\|_{m,2}$ is trivial. ■

3.1.5. The Neumann Poisson problem. Repeating the derivation of the weak formulation, we find for

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ \frac{\partial u}{\partial n} &= 0 && \text{on } \partial\Omega \end{aligned} \tag{3.10}$$

Definition 3.11. *Let $\Omega \subset \mathbb{R}^n$ be open and $f \in \mathbb{L}^2(\Omega)$. A function $u \in H^1(\Omega)$ is called* weak solution *of* (3.10) *if*

$$\langle\nabla u, \nabla v\rangle_0 = \langle f, v\rangle_0 \qquad \forall v \in H^1(\Omega).$$

We remark that for bounded Ω the weak solution is not unique. If $u \in H^1(\Omega)$ is a solution, the same holds for $u + c \in H^1(\Omega)$ where c is any constant. Consequently, $V = H^1(\Omega)$ together with $\langle u, v\rangle_V = \langle\nabla u, \nabla v\rangle_0$ cannot be a Hilbert space since otherwise the solution would be unique due to Theorem 3.5. Indeed, $\langle u, u\rangle_V$ vanishes exactly on the subspace U of constant functions if Ω is connected (see Theorem B.9; if Ω is not connected, the critical functions are those which are constant on each connected component with not necessarily the same constants). To fix this arbitraryness in the solution we should therefore work in $U^\perp$. If Ω is bounded, then the constant function $e(x) = 1$ is an element of $V = H^1(\Omega)$ and, for connected Ω, the projection on U is given by

$$P_U(v) = \langle e, v\rangle_V\, e = (\langle\nabla e, \nabla v\rangle_0 + \langle e, v\rangle_0)e = \int_\Omega v\,dx\,e$$

Consequently, $U^\perp$ can be characterized as

$$U^\perp = \left\{ u \in H^1(\Omega) \, : \, \int_\Omega u \, dx = 0 \right\}.$$

To show that $U^\perp$ with $\langle \cdot, \cdot \rangle_V$ is a Hilbert space we need a generalization of the Poincaré inequality.

Theorem 3.12. **(Bramble–Hilbert)**
Let $\Omega \subset \mathbb{R}^n$ be a bounded, connected $C^{0,1}$ domain and let $m \in \mathbb{N}$, $1 \leq p < \infty$. Assume further that $\{\mu_1, \ldots, \mu_k\}$ is a family of continuous seminorms on $W^{m,p}(\Omega)$ with the property that $\mu_i(P) = 0$ for all polynomials of order less than m and all $i = 1, \ldots, k$ implies $P \equiv 0$. Then, the norm

$$\|u\|_\mu = \left(\sum_{|\alpha|=m} \|\nabla^\alpha u\|^p_{\mathbb{L}^p(\Omega)} + \sum_{i=1}^k \mu_i(f)^p \right)^{\frac{1}{p}}$$

is equivalent to the $W^{m,p}(\Omega)$ norm.

Proof: Continuity of μ_i implies for some constant C_i

$$\mu_i(u) \leq C_i \|u\|_{m,p}, \qquad \forall u \in W^{m,p}(\Omega)$$

so that $\|u\|_\mu \leq C\|u\|_{m,p}$ is easy. A converse inequality is shown by contradiction. We assume that there is a sequence $(u_l)_{l\in\mathbb{N}} \subset W^{m,p}(\Omega)$ such that $\|u_l\|_{m,p} = 1$ but $\|u_l\|_\mu < 1/l$. The latter condition implies

$$\lim_{l\to\infty} \|\nabla^\alpha u_l\|_{\mathbb{L}^p(\Omega)} = 0, \qquad \forall |\alpha| = m, \tag{3.11}$$

and

$$\lim_{l\to\infty} \mu_i(u_l) = 0, \qquad \forall i = 1, \ldots, k. \tag{3.12}$$

Since $W^{m,p}(\Omega)$ is compactly embedded in $W^{m-1,p}(\Omega)$ (see Theorem C.2), the $\|\cdot\|_{m,p}$ bounded sequence $(u_l)_{l\in\mathbb{N}}$ has a convergent subsequence in $W^{m-1,p}(\Omega)$ (which we denote again $(u_l)_{l\in\mathbb{N}}$ to avoid subscripts). If u denotes the limit of this subsequence, we have $\nabla^\alpha u = 0$. Indeed, for any $\phi \in \mathscr{D}(\Omega)$ and $|\alpha| = m$ we have

$$\langle \nabla^\alpha u, \phi \rangle = (-1)^{|\alpha|} \langle u, \nabla^\alpha \phi \rangle$$

and since $u_l \to u$ in $\mathbb{L}^p(\Omega)$ and thus also in $\mathscr{D}'(\Omega)$

$$\langle \nabla^\alpha u, \phi \rangle = (-1)^{|\alpha|} \lim_{l\to\infty} \langle u_l, \nabla^\alpha \phi \rangle = \lim_{l\to\infty} \langle \nabla^\alpha u_l, \phi \rangle .$$

Due to (3.11) we thus find

$$\langle \nabla^\alpha u, \phi \rangle = 0 \qquad \forall \phi \in \mathscr{D}(\Omega), \; |\alpha| = m$$

so that $\|u - u_l\|_{m,p} \to 0$ for $l \to \infty$. Since u satisfies $\nabla^\alpha u = 0$ for all $|\alpha| = m$, Theorem B.9 implies that u is a polynomial of degree less than m. Now (3.12) shows (due to continuity of μ_i) that $\mu_i(u) = 0$ for all i so that by assumption $u = 0$. This however is in contradiction to the assumption $\|u_l\|_{m,p} = 1$ for all l since $\lim_{l\to\infty} \|u_l\|_{m,p} = \|u\|_{m,p} = 0$. ∎

Now, we are in the position to prove existence and uniqueness for the Neumann Poisson problem on a bounded domain.

Theorem 3.13. *Let $\Omega \subset \mathbb{R}^n$ be bounded and connected with a $C^{0,1}$ boundary. If $f \in \mathbb{L}^2(\Omega)$ satisfies $\langle 1, f\rangle_0 = 0$ then problem* (3.10) *admits a solution which is unique under the condition $\langle 1, u\rangle_0 = 0$. Furthermore, we have $\|u\|_{1,2} \leq C\|f\|_{0,2}$ for some constant C which is independent of u and f.*

Proof: We consider the Hilbert space

$$V = \{\, u \in H^1(\Omega) \,:\, \langle 1, u\rangle_0 = 0 \,\}$$

with the scalar product $\langle u, v\rangle_V = \langle \nabla u, \nabla v\rangle_0$. On V, the corresponding norm is equivalent to the $H^1(\Omega)$ norm and thus V is a Hilbert space with this scalar product (just apply Theorem 3.12 with the seminorm $\mu_1(u) = |\langle 1, u\rangle_0|$). Then Theorem 3.5 yields a unique solution u to the weak formulation of (3.10) whose norm can be estimated by the norm of the functional

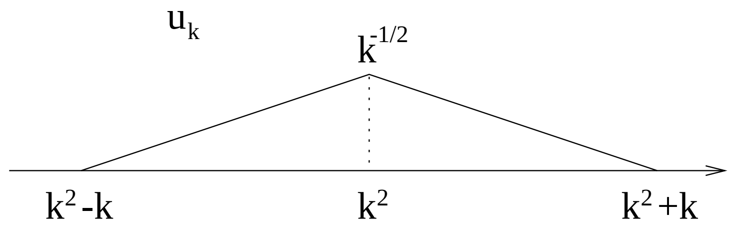
$$l_f(v) = \langle f, v\rangle_0 \qquad v \in V$$

in the dual space V'. (Note that the condition $\langle 1, f\rangle_0 = 0$ ensures that $f \in V$ and thus $l_f \in V'$.) By definition of the dual norm, we have

$$\|l_f\|_{V'} = \sup_{\|v\|_V=1} l_f(v) = \sup_{\|v\|_V=1} \langle f, v\rangle_0 \leq \sup_{\|v\|_V=1} \|f\|_{0,2}\,\|v\|_{0,2} \leq \|f\|_{0,2}.$$

■

If Ω is not connected, one can apply the same considerations to each of the connected components. In the case of unbounded Ω, uniqueness of the solution is no problem if the constant function $e \equiv 1$ is not contained in $H^1(\Omega)$. Indeed, $\nabla u = 0$ implies $u = const$ (Theorem B.9) which again implies $u = 0$ (which is, by assumption, the only constant function in $H^1(\Omega)$). However, Theorem 3.12 cannot be applied directly in that case since boundedness of Ω is important in the proof to ensure that $H^1(\Omega)$ is compactly embedded in $\mathbb{L}^2$. To see this, consider the functions $u_k \in H^1(0,\infty)$ given by the figure below.



It is easily checked that $\|\nabla u_k\|_{0,2} = \mathcal{O}(1/k)$ while $\|u_k\|_{0,2} = \mathcal{O}(1)$. Hence, the H^1 norm can never be controled by the norm of the gradient alone. For such cases one has to require a certain decay of the functions and consider weighted Sobolev spaces instead of $H^1(\Omega)$ as starting point.

3.2. The general Dirichlet problem

We consider a general linear operator on an open set $\Omega \subset \mathbb{R}^n$

$$L(x, \nabla) = \sum_{|\alpha| \le k} a_\alpha(x) \nabla^\alpha \qquad x \in \Omega$$

which is elliptic, i.e.

$$L^p(x, i\xi) = i^k \sum_{|\alpha| = k} a_\alpha(x) \xi^\alpha \neq 0 \qquad \forall \xi \in \mathbb{R}^n \backslash \{0\}.$$

It is easy to see, that (for $n \ge 2$) k must be even and with some smoothness assumptions on the coefficients a_α one can show that L can be written in *divergence form*

$$L(x, \nabla)u = \sum_{0 \le |\sigma|, |\gamma| \le m} (-1)^{|\sigma|} \nabla^\sigma (a_{\sigma_\gamma}(x) \nabla^\gamma u) \qquad x \in \Omega$$

with suitable coefficients $a_{\sigma\gamma}$ (see Exercise 3.1). The divergence form strongly simplifies a weak formulation of boundary value problems related to L.

As example, let us consider the classical Dirichlet problem, to find $u \in C^{2m}(\Omega) \cap C^m(\bar{\Omega})$ such that

$$\begin{aligned} L(x, \nabla)u &= f \qquad \text{in } \Omega \\ \nabla^\alpha u &= 0 \qquad \text{on } \partial\Omega, \ |\alpha| \le m - 1 \end{aligned} \tag{3.13}$$

where $f \in C(\Omega)$. To set up a weak formulation, we multiply the equation by a test function $v \in C_0^\infty(\Omega)$ and integrate by parts. This leads to the equation

$$\sum_{0 \le |\sigma|, |\gamma| \le m} \int_\Omega a_{\sigma_\gamma} \nabla^\gamma u \nabla^\sigma v \, dx = \int_\Omega f v \, dx.$$

Introducing the bilinear form

$$a(u, v) = \sum_{0 \le |\sigma|, |\gamma| \le m} \int_\Omega a_{\sigma_\gamma} \nabla^\gamma u \nabla^\sigma v \, dx \tag{3.14}$$

and the linear form

$$l(v) = \int_\Omega f v \, dx$$

we obtain

Definition 3.14. *Let $\Omega \subset \mathbb{R}^n$ be open. In* (3.14)*, let $a_{\sigma,\gamma} \in \mathbb{L}^\infty(\Omega)$ for all $0 \le |\sigma|, |\gamma| \le m$ and $l \in H^{-m}(\Omega)$ (the dual space of $H_0^m(\Omega)$). A function $u \in H_0^m(\Omega)$ is called weak solution of the Dirichlet problem* (3.13) *if*

$$a(u, v) = l(v) \qquad \forall v \in H_0^m(\Omega).$$

Similar to the examples in the previous section, we can generalize the homogeneous case to more general Dirichlet conditions. For example, we can require that $\nabla^\alpha u = \nabla^\alpha g$ on $\partial\Omega$ for all $|\alpha| \le m - 1$ where $g \in H^m(\Omega)$ is a given function.

3.3. The Lax Milgram Theorem

The weak formulation derived in the previous section can easily be stated completely in functional analytic terms:

Definition 3.15. *Let V be a Hilbert space, $a : V \times V \mapsto \mathbb{R}$ a bilinear form and $l \in V'$ an element of the dual space. The problem to find $u \in V$ such that*

$$a(u,v) = l(v) \qquad \forall v \in V \tag{3.15}$$

is called abstract variational problem.

The name is motivated by the observation that (3.15) is equivalent to an optimization problem, if a is a *symmetric bilinear form.*

Lemma 3.16. *Let a be a symmetric bilinear form on a Hilbert space V which satisfies $a(v,v) \geq 0$ for all $v \in V$. Further, let $l \in V'$ be a bounded linear functional. Then* (3.15) *is equivalent to the variational problem to find $u \in V$ which minimizes the functional*

$$\Theta(u) = \frac{1}{2}a(u,u) - l(u).$$

Proof: If u is a minimizer of the functional Θ, then

$$f(t) := \Theta(u + tv)$$

is minimal at $t = 0$ for any $v \in V$. In particular,

$$\frac{d}{dt}f(t)\bigg|_{t=0} = 0.$$

To calculate the derivative, we first use the bilinearity of a to write

$$f(t) = \Theta(u) + (a(u,v) - l(v))t + \frac{1}{2}a(v,v)t^2. \tag{3.16}$$

Clearly, the vanishing derivative in $t = 0$ implies

$$a(u,v) - l(v) = 0 \qquad \forall v \in V$$

which shows that u solves (3.15). Conversely, if u is a solution of the abstract variational problem, we find for any $v \in V$ by setting $t = 1$ in (3.16)

$$\Theta(u+v) = \Theta(u) + (a(u,v) - l(v)) + \frac{1}{2}a(v,v).$$

Since $(a(u,v) - l(v)) = 0$ and $a(v,v) \geq 0$ we see that $\Theta(u+v) \geq \Theta(u)$ for all $v \in V$ which shows that u is a minimizer of Θ. ∎

For problems arising from physical models, one very often finds that Θ has the meaning of an energy. For example, in electrostatics the equation for the potential ϕ of an electric field $E = \nabla\phi$ which is generated by a charge distribution ρ is just Poisson's equation $-\Delta\phi = \rho$ (for example with boundary conditions $\phi = 0$ if the boundary of the domain is 'grounded'). We already know the bilinear form a corresponding to the Dirichlet Poisson problem. It is given by the $\mathbb{L}^2$ scalar product of the gradients of the functions. Since the gradient of ϕ is the electric field E, the functional Θ has the form

$$\Theta(\phi) = \frac{1}{2}\int_\Omega \|E\|^2\,dx - \int_\Omega \phi\rho\,dx.$$

While the first integral describes the electric energy stored in the field E, the second term is the potential energy of the charge distribution ρ in the potential ϕ so that Θ is indeed an energy functional.

In order to get a general solution theory for abstract variational problems we need as minimal requirements *continuity* and *ellipticity* of the bilinear form a.

Definition 3.17. *Let $a : V \times V \mapsto \mathbb{R}$ be a bilinear form on the Hilbert space V. We say that a is continuous if there exists a constant $C_1 > 0$ such that*

$$|a(u,v)| \le C_1\|u\|_V\|v\|_V \qquad \forall u, v \in V.$$

a is called V–elliptic (or simply elliptic) if, for some $C_2 > 0$

$$a(u,u) \ge C_2\|u\|_V^2 \qquad \forall u \in V.$$

We remark that ellipticity of L is a weaker condition than ellipticity of the corresponding bilinear form a (see Exercise 3.5). Also, $-\Delta$ is an elliptic operator but the corresponding form

$$a(u,v) = \langle \nabla u, \nabla v\rangle_0 \tag{3.17}$$

is in general not $H^1(\Omega)$–elliptic (reflecting the non uniqueness of the Neumann Poisson problem). Obviously, a contains not only information about the differential operator but also about the boundary conditions of the problem (which are hidden in the domain of definition of a). For the bilinear form (3.17) on the Hilbert space $V = H_0^1(\Omega)$ (which appears in the connection with the Dirichlet Poisson problem), continuity follows immediately from the Cauchy-Schwarz inequality in $\mathbb{L}^2$

$$|a(u,v)| \le \|\nabla u\|_{\mathbb{L}^2}\|\nabla v\|_{\mathbb{L}^2} \le \|u\|_{1,2}\|v\|_{1,2}.$$

Typically, the more difficult part is to prove ellipticity. Indeed, for the Poisson problem we needed the Poincaré inequality to show

$$a(u,u) = \langle \nabla u, \nabla u\rangle_0 \ge C_2\|u\|_{1,2}^2 \qquad \forall u \in H_0^1(\Omega).$$

If a is a symmetric bilinear form (as in the Poisson problem) a simple application of the Riesz representation theorem yields a unique solution of the abstract variational problem.

Lemma 3.18. *Let V be a Hilbert space and let $a : V \times V \mapsto \mathbb{R}$ be a symmetric, continuous and elliptic bilinear form. Then* (3.15) *admits a unique solution u which depends continuously on the data (i.e. $\|u\|_V \le C\|l\|_{V'}$).*

Proof: The form a is clearly a scalar product since it is bilinear, symmetric and satisfies $a(u,u) \ge C_2\|u\|_V^2 > 0$ for all $u \ne 0$. Using continuity and ellipticity we find for the associated norm $\|u\|_a := \sqrt{a(u,u)}$

$$\sqrt{C_2}\|u\|_V \le \|u\|_a \le \sqrt{C_1}\|u\|_V$$

which shows that $\|\cdot\|_a$ gives rise to the original (complete) topology on V. Altogether, V is Hilbert space also with the scalar product $a(\cdot,\cdot)$. Due to the equivalence of norms, any linear functional $l \in V'$ is also continuous with respect to $\|\cdot\|_a$. Thus Theorem 3.5 can be applied and we obtain a unique solution u of the abstract variational problem which satisfies

$$\|u\|_a = \sup_{\|v\|_a=1} |l(v)|.$$

Using ellipticity twice, we get

$$\sqrt{C_2}\|u\|_V \le \|u\|_a \le \|l\|_{V'} \sup_{\|v\|_a=1} \|v\|_V \le \frac{1}{\sqrt{C_2}}$$

which yields the continuous dependence in the original norm of V. ■

Unfortunately, the Lemma applies only to symmetric bilinear forms so that many elliptic operators are excluded. For example, let us consider a simple, one dimensional situation with a second order operator with constant coefficients

$$Lu = -(a_{11}u')' - (a_{10}u)' + a_{01}u' + a_{00}u.$$

In this case, the bilinear form is

$$a(u,v) = \int_\Omega a_{11}u'v' + a_{10}uv' + a_{01}u'v + a_{00}uv\,dx.$$

Assuming symmetry of a, the relation $0 = a(u,v) - a(v,u)$ implies

$$(a_{10} - a_{01}) \int_\Omega uv' - vu'\,dx = 0 \qquad \forall u, v \in H_0^1(\Omega).$$

Using integration by parts, we get

$$2(a_{10} - a_{01}) \int_\Omega u'v\,dx = 0 \qquad \forall u, v \in H_0^1(\Omega).$$

Finally, by choosing a function $v \ge 0$ in $C_0^\infty(\Omega)$ and a function u with $u' > 0$ on the support of v we can conclude $a_{01} = a_{10}$. Going back to the corresponding operator L

$$Lu = -a_{11}u'' + (a_{01} - a_{10})u' + a_{00}u = -a_{11}u'' + a_{00}u$$

we see that L cannot contain first derivatives if a is a symmetric bilinear form. To incorporate also non symmetric forms in our general treatment we therefore need to extend the Riesz representation theorem. The corresponding result for general bilinear forms is called Lax Milgram Theorem. It plays a central role in the theory of linear elliptic partial differential equations because it provides existence and uniqueness results for weak solutions.

Theorem 3.19. (Lax Milgram)
Let V be a Hilbert space, $l \in V'$ a continuous linear functional and $a : V \times V \mapsto \mathbb{R}$ a continuous, elliptic bilinear form. Then, the abstract variational problem (3.15) admits a unique solution $u \in V$ which satisfies $\|u\|_V \leq C\|l\|_{V'}$.

Proof: In the first step, we transform the abstract variational problem into an operator equation of the form $\tilde{A}u = \tilde{l}$. To do this, we fix $u \in V$ and define the functional

$$a_u(v) := a(u, v) \qquad \forall v \in V.$$

Clearly, a_u is linear and we have

$$|a_u(v)| = |a(u, v)| \leq C_1\|u\|_V\|v\|_V$$

which implies $\|a_u\|_{V'} \leq C_1\|u\|_V$. In particular, $a_u \in V'$ and we can define

$$\tilde{A}u := \mathcal{R}a_u$$

where $\mathcal{R}$ is the Riesz operator from Corollary 3.8. By definition of $\mathcal{R}$, we have

$$a(u, v) = a_u(v) = \langle \mathcal{R}a_u, v\rangle_V = \left\langle \tilde{A}u, v\right\rangle_V$$

which leads to the equivalent reformulation of problem (3.15)

$$a(u, v) = l(v)\ \forall v \in V \quad \Leftrightarrow \quad a_u = l \quad \Leftrightarrow \quad \tilde{A}u = \mathcal{R}l =: \tilde{l}$$

which is the desired operator formulation.
In the second step of the proof we show that $\tilde{A}$ is a linear bounded operator which has a continuous inverse. This implies that $\tilde{A}u = \tilde{l}$ (and thus also the variational problem) has a unique solution $u = \tilde{A}^{-1}\tilde{l}$. Continuity of $\tilde{A}^{-1}$ obviously gives continuous dependence on the data.
Since $\tilde{A}$ is the composition of $\mathcal{R}$ and a_u linearity follows from the observation that

$$\begin{aligned} a_{\alpha_1 u_1 + \alpha_2 u_2}(v) &= a(\alpha_1 u_1 + \alpha_2 u_2, v) \\ &= \alpha_1 a(u_1, v) + \alpha_2 a(u_2, v) = (\alpha_1 a_{u_1} + \alpha_2 a_{u_2})(v). \end{aligned}$$

Boundedness follows from

$$\|\tilde{A}u\|_V = \|\mathcal{R}a_u\|_V = \|a_u\|_{V'} \leq C_1\|u\|.$$

Existence of an inverse operator is shown with the Banach fixed point theorem (the idea is taken from [**5**]). We transform $\tilde{A}u = \tilde{l}$ in a fixed point problem $Tu = u$ with

$$Tu = u - \rho(\tilde{A}u - \tilde{l}) \qquad u \in V$$

T is an affine linear map which is a contraction mapping if we choose ρ in a suitable way. Indeed,

$$\begin{aligned}\|Tu_1 - Tu_2\|_V^2 &= \|u_1 - u_2\|_V^2 - 2\rho \langle A(u_1 - u_2), u_1 - u_2\rangle_V + \rho^2 \|A(u_1 - u_2)\|_V^2 \\ &\le (1 + \rho^2 C_1^2)\|u_1 - u_2\|_V^2 - 2\rho a(u_1 - u_2, u_1 - u_2) \\ &\le (1 - 2C_2\rho + \rho^2 C_1^2)\|u_1 - u_2\|_V^2 \\ &= \left(1 - \frac{C_2^2}{C_1^2}\right)\|u_1 - u_2\|_V^2\end{aligned}$$

where the last line follows if we choose $\rho = C_2/C_1^2$. According to the Banach fixed point theorem, T has a unique fixed point u which is the limit of the sequence $u_{n+1} = Tu_n$ (in fact, this is *Richardson iteration with relaxation* – see Appendix A). Due to the construction of T, the fixed point u is a solution of $\tilde{A}u = \tilde{l}$ which shows existence and uniqueness. To get continuity of $\tilde{A}^{-1}$ (or equivalently continuous dependence on the data) we can use again a result from the Banach fixed point theorem

$$\|u\|_V = \|Tu\|_V \le \|Tu - T0\|_V + \|T0\|_V \le \sqrt{1 - \frac{C_2^2}{C_1^2}}\|u\|_V + \rho\|\tilde{l}\|_V.$$

Another possibility is

$$\begin{aligned}\|\tilde{A}^{-1}\tilde{l}\|_V^2 = \frac{1}{C_2}C_2\|u\|_V^2 \le \frac{1}{C_2}a(u,u) &= \frac{1}{C_2}\left\langle \tilde{A}u, u\right\rangle_V \\ &\le \frac{1}{C_2}\|\tilde{A}u\|_V\|u\|_V = \frac{1}{C_2}\|\tilde{l}\|_V\|\tilde{A}^{-1}\tilde{l}\|_V.\end{aligned}$$

■

In the proof of the Lax Milgram theorem we have transformed the abstract variational problem into an operator equation. This approach turns out to be useful also for other purposes. We slightly reformulate the idea.

Corollary 3.20. *To any continuous bilinear form a on $V \times V$ we can uniquely associate a bounded linear operator $A : V \mapsto V'$ such that*

$$a(u,v) = \langle Au, v\rangle_{V'\times V} \qquad u, v \in V.$$

If a is elliptic, the operator A is a continuous bijection with $\|A^{-1}\| \le 1/C_2$ where C_2 is the ellipticity constant of a.

Proof: Let $\tilde{A}$ be the operator constructed in the proof of the Lax Milgram theorem. We set $A := \mathcal{R}^{-1}\tilde{A}$ so that $Au = \mathcal{R}^{-1}\tilde{A}u = \mathcal{R}^{-1}\mathcal{R}a_u = a_u$. Then

$$\langle Au, v\rangle_{V'\times V} = a_u(v) = a(u,v) \qquad \forall v \in V.$$

The remaining properties of A follow directly from those of $\tilde{A}$ and $\mathcal{R}$. ■

The representation of a in terms of A gives rise to an operator formulation of the abstract variational problem.

Let V be a Hilbert space, $l \in V'$ and let $a : V \times V \mapsto \mathbb{R}$ be a continuous bilinear form. If A denotes the operator associated to a then problem (3.15) is equivalent to finding $u \in V$ such that

$$Au = l.$$

Note that if a arises from an elliptic problem like (3.13), the operator A contains information about the differential operator L *and* the boundary conditions (which are hidden in the domain of definition of A).

In the particular case where a arises from the homogeneous Dirichlet problem (Section 3.2), the associated operator A can be considered as a direct extension of the differential operator L. This is due to the fact that $C_0^\infty(\Omega)$ is dense in $V = H_0^m(\Omega)$ and that the action of A on this dense subset is exactly the one of L. More precisely, for $\phi, \psi \in C_0^\infty(\Omega)$ we have by definition

$$\langle A\phi, \psi \rangle_{H^{-m} \times H_0^m} = a(\phi, \psi)$$

and using integration by parts, we find

$$\langle A\phi, \psi \rangle_{H^{-m} \times H_0^m} = \langle L\phi, \psi \rangle_{\mathbb{L}^2} = \left\langle \mathcal{R}_{\mathbb{L}^2}^{-1} L\phi, \psi \right\rangle_{(\mathbb{L}^2)' \times \mathbb{L}^2} \tag{3.18}$$

where $\mathcal{R}_{\mathbb{L}^2}$ is the Riesz operator on $\mathbb{L}^2$. We now want to identify the $\mathbb{L}^2$–functional $\mathcal{R}_{\mathbb{L}^2}^{-1} L\phi$ with an H_0^m–functional (to get an equivalent expression for $A\phi$). This is possible since $H_0^m(\Omega)$ is continuously embedded in $\mathbb{L}^2(\Omega)$.

Definition 3.21. *Let V, U be Banach spaces. We say that V is continuously embedded in U if there exists a bounded linear and injective mapping $I : V \mapsto U$ (a so called inclusion operator). Based on I, we can identify V with the linear subspace $I(V) \subset U$. If there is no ambiguity about I, we also write simply $V \hookrightarrow U$.*

Lemma 3.22. *Let V, U be Hilbert spaces and $V \hookrightarrow U$ based on the inclusion I. If $I(V)$ is dense in U then $U' \hookrightarrow V'$ based on the inclusion I' defined by*

$$\langle I'l, v \rangle_{V' \times V} := \langle l, Iv \rangle_{U' \times U} \qquad \forall l \in U',\ v \in V.$$

Proof: First, we note that I' is well defined. Indeed, since I is linear and bounded

$$\Lambda(v) := \langle l, Iv \rangle_{U' \times U}$$

defines an element of V' for each $l \in U'$. Setting $I'l := \Lambda$ we first check injectivity of the mapping. If $I'l_1 = I'l_2$ then clearly $\langle l_1 - l_2, Iv \rangle_{U' \times U} = 0$ for all $v \in V$. Since $I(V)$ is dense in U this implies $l_1 = l_2$. The linearity of I' is easily checked by using the linearity of the dual product and again the density of $I(V)$ in U. Finally

$$|(I'l)(v)| \leq \|l\|_{U'} \|Iv\|_U \leq \|I\| \, \|l\|_{U'} \|v\|_V.$$

shows that $\|I'\| \leq \|I\|$. ∎

In the particular case $V = H_0^m(\Omega)$ and $U = \mathbb{L}^2(\Omega)$ a natural inclusion is the identity mapping on $\mathbb{L}^2(\Omega)$ restricted to $H_0^m(\Omega)$ (since H_0^m is, by definition, a subset of $\mathbb{L}^2$). Moreover, since $C_0^\infty(\Omega) \subset H_0^m(\Omega)$ is dense in $\mathbb{L}^2(\Omega)$ we can apply Lemma 3.22 which shows that $(\mathbb{L}^2)' \hookrightarrow H^{-m}$ based on the adjoint of the natural inclusion of H_0^m into EL^2. Consequently, we can continue equation (3.18) as

$$\langle A\phi, \psi \rangle_{H^{-m}\times H_0^m} = \left\langle \mathcal{R}_{\mathbb{L}^2}^{-1} L\phi, \psi \right\rangle_{H^{-m}\times H_0^m} \qquad \forall \phi, \psi \in C_0^\infty(\Omega).$$

(The inclusion operator corresponding to $(\mathbb{L}^2)' \hookrightarrow H^{-m}$ is typically not mentioned because it is the natural choice.) Since $C_0^\infty(\Omega)$ is dense in $H_0^m(\Omega)$ we get

$$A\phi = \mathcal{R}_{\mathbb{L}^2}^{-1} L\phi \qquad \forall \phi \in C_0^\infty(\Omega)$$

which shows that A is the unique extension of $\mathcal{R}_{\mathbb{L}^2}^{-1} L$ from $C_0^\infty(\Omega)$ to $H_0^m(\Omega)$. We can even suppress the Riesz operator if we agree on the chain of inclusions

$$H_0^m(\Omega) \hookrightarrow \mathbb{L}^2(\Omega) = \mathbb{L}^2(\Omega)' \hookrightarrow H^{-m}(\Omega). \tag{3.19}$$

(This line has to be read in the following way: H_0^m is embedded in $\mathbb{L}^2$ based on the natural inclusion; $\mathbb{L}^2$ is identified with $(\mathbb{L}^2)'$ based on the Riesz isometry $\mathcal{R}_{\mathbb{L}^2}$; $(\mathbb{L}^2)'$ is included in H^{-m} based on the adjoint of the natural inclusion $H_0^m \hookrightarrow \mathbb{L}^2$.) If $I : H_0^m(\Omega) \mapsto H^{-m}(\Omega)$ denotes the inclusion generated by the chain (3.19) (i.e. the composition of natural inclusion, Riesz operator and adjoint of natural inclusion) we get again

$$\langle IL\phi, \psi \rangle_{H^{-m}\times H_0^m} = \langle A\phi, \psi \rangle_{H^{-m}\times H_0^m} \qquad \forall \phi, \psi \in C_0^\infty(\Omega).$$

A is the unique extension of IL which motivates the notation $A = IL$ (or even $A = L$ if it is clear that the inclusion is based on the chain (3.19)). An important feature of the inclusion (3.19) is its compactness. Indeed, $H_0^m(\Omega)$ is a compact subset of $\mathbb{L}^2(\Omega)$ which is equivalent to saying that the natural inclusion operator is compact. Since compositions of compact and bounded operators are compact, we conclude that I is a compact mapping. In particular, if a is an elliptic form, we know that A is invertible (mapping H^{-m} into H_0^m), so that IA^{-1} is a compact operator on $H^{-m}(\Omega)$ (or again, if we agree on I as the natural inclusion we can simply say that A^{-1} is compact). Recalling that $A^{-1}l$ is the solution of the abstract variational problem we conclude that the solution operator corresponding to boundary value problems which are associated to elliptic bilinear forms is a compact operator.

This important result enables us to apply spectral theory of compact operators and thus yields information about eigenvalues and eigensolutions of elliptic boundary value problems.

We conclude with the remark that the inclusion $I : H_0^m(\Omega) \mapsto H^{-m}(\Omega)$ based on (3.19) is not the only possible choice. Indeed, based on the inclusion $\mathcal{R}^{-1} : H_0^m(\Omega) \mapsto H^{-m}(\Omega)$ (the Riesz operator on $H_0^m(\Omega)$) one can even *identify* H_0^m with its dual H^{-m}. Note, however, that this embedding cannot be compact since the inverse of a compact operator cannot be bounded.

Also, the inclusion I is the most natural in connection with eigenvalue problems

$$\begin{aligned} L(x,\nabla)u &= \lambda u && \text{in } \Omega \\ \nabla^\alpha u &= 0 && \text{on } \partial\Omega,\ |\alpha| \le m-1 \end{aligned} \tag{3.20}$$

Indeed, the weak formulation of (3.20) is to find $u \in H_0^m(\Omega)$ such that

$$a(u,v) = \lambda \langle u, v\rangle_{\mathbb{L}^2} \qquad \forall v \in H_0^m(\Omega).$$

Note that the inclusion operator I is just the operator associated to the bilinear form $\langle u, v\rangle_{\mathbb{L}^2}$, i.e.

$$\langle Iu, v\rangle_{H^{-m}\times H_0^m} = \langle u, v\rangle_{\mathbb{L}^2}\,.$$

Consequently, the operator notation of the eigenvalue problem is

$$Au = \lambda I u$$

where A is associated to a. With our considerations above, we can as well write

$$I(Lu - \lambda u) = 0$$

(or even $Lu - \lambda u = 0$ if we suppress the injection I).

3.4. General eigenvalue problems

Up to now, we have tacitly assumed that V is a Hilbert space over $\mathbb{R}$. In the case of spectral analysis, however, it is favorable to work with the field $\mathbb{C}$ of complex numbers. The standard procedure in this case is a complexification of V, where V is embedded in a $\mathbb{C}$–space $\hat{V}$. More precisely, we define

$$\hat{V} := \{(u,v) \,:\, u, v \in V\} = V \times V$$

and use as notation $(u,v) = u + iv$. (Note that the complex unit i is only a symbol which is used to memorize the following definitions of addition, multiplication and the rules for extensions of functionals, mappings and bilinear forms.) We set

$$(u+iv) + (x+iy) := u + x + i(v+y)$$

and define multiplication with $\lambda = \alpha + i\beta \in \mathbb{C}$ by

$$(\alpha + i\beta)(u+iv) := (\alpha u - \beta v) + i(\alpha v + \beta u).$$

The projection on the first component of $V\times V$ is denoted $\Re$ and on the second component $\Im$. It is then natural to identify V with the subspace $\Re(V) = V \times \{0\}$ of *real vectors*. Based on $\langle\cdot,\cdot\rangle_V$ we define a scalar product on $\hat{V}$

$$\langle u+iv, x+iy\rangle_{\hat{V}} := \langle u,x\rangle_V + \langle v,y\rangle_V + i(\langle v,x\rangle_V - \langle u,y\rangle_V).$$

The definition is done in such a way that

$$\langle w,z\rangle_{\hat{V}} = \overline{\langle z,w\rangle_{\hat{V}}} \qquad w,z \in \hat{V}.$$

Moreover, if we apply $\langle\cdot,\cdot\rangle_{\hat{V}}$ to real vectors $w,z \in V$ (i.e. $\Im z = \Im w = 0$) we recover the V product. The same approach is used to extend any other bilinear form a. Note that the extension $\hat{a}$ is a sesquilinear form on V, i.e. it is linear in the first component, additive in the second and $\hat{a}(w, \lambda z) = \bar{\lambda}\hat{a}(w, z)$. All other extensions are done along the same lines. For example, a mapping $A : V \mapsto U$ is extended to $\hat{A} : \hat{V} \mapsto \hat{U}$ by

$$\hat{A}(u + iv) := Au + iAv.$$

(If $\hat{U}$ is already a $\mathbb{C}$ vector space, then $\hat{U} = U$ otherwise $\hat{U}$ is the complexification of U.) In the following, we can always think of V as a subset of a complex vector space. In this sense, spectral theory can be applied to objects defined on V by using their complex extensions. Of course, the results are in general related to the complexification $\hat{V}$ and perhaps have to be translated back into results for the original objects on V. However, since all extensions used in the complexification are very natural it is clear from the context how expressions involving $\mathbb{C}$-objects are to be understood.

With these remarks in mind, we now turn to the eigenvalue problem. It turns out that the underlying bilinear form a need not be elliptic. It suffices if a is V–coercive.

Definition 3.23. *Let V be a Hilbert space which is dense and continuously embedded into another Hilbert space U. A bilinear form $a : V \times V \mapsto \mathbb{R}$ is called V–coercive with respect to U if a is continuous and if there exists $C_V, C_U > 0$ such that*

$$a(v, v) + C_U\|v\|_U^2 \geq C_V\|v\|_V^2 \qquad \forall v \in V.$$

Note that coercivity is closely connected with ellipticity. Indeed, if a is V–elliptic then A is also V–coercive with respect to any U in which V is dense and continuously embedded. Conversely, if a is V–coercive with respect to U we can define a new bilinear form

$$\tilde{a}(u, v) = a(u, v) + C\,\langle u, v\rangle_U \qquad u, v \in V$$

which is V–elliptic if $C \geq C_U$. Note that the operator $\tilde{A}$ related to $\tilde{a}$ is given by

$$\tilde{A} = A + CI$$

where A is related to a. The operator $I : V \mapsto V'$ is the natural inclusion operator from V into V' related to the chain

$$V \hookrightarrow U = U' \hookrightarrow V'.$$

The first inclusion $V \hookrightarrow U$ of the chain is assumed in the current context. Since V is dense in U, we know from Lemma 3.22 that also $U' \hookrightarrow V'$ which explains the last inclusion of the chain. The identification $U = U'$ is naturally based on the Riesz operator on U. Hence, I is defined by the relation

$$\langle Iu, v\rangle_{V'\times V} = \langle u, v\rangle_U \qquad \forall u, v \in V. \tag{3.21}$$

Typically, the action of inclusion operators is unambiguous so that they need not be mentioned explicitly. In this spirit, we can write (3.21) in the form

$$\langle u, v\rangle_{V'\times V} = \langle u, v\rangle_U \qquad \forall v \in V, u \in U.$$

If the embedding $V \hookrightarrow U$ is compact (i.e. if V–bounded sequences have U–convergent subsequences), we have the following result.

Theorem 3.24. *Let V, U be Hilbert spaces which are arranged in the chain $V \hookrightarrow U = U' \hookrightarrow V'$ where V is assumed to be dense and compactly embedded in U. The inclusion operator from V into V' related to the chain is denoted I. Let $a : V \times V \mapsto \mathbb{R}$ be V–coercive with respect to U and let $A : V \mapsto V'$ be the operator related to a.*

a) *For every $\lambda \in \mathbb{C}$ one of the two alternatives holds*
   i) *$(A - \lambda I)^{-1}$ is a bounded linear operator from V' into V. In particular, $Au - \lambda Iu = l$ is uniquely solvable for every $l \in V'$ (i.e. $a(u, v) - \lambda \langle u, v\rangle_U = l(v)$ for all $v \in V$).*
   ii) *The subspace $E(\lambda) = \ker(A - \lambda I)$ has dimension $1 \leq \dim E(\lambda) < \infty$. ($E(\lambda)$ is the eigenspace corresponding to the eigenvalue λ). In particular, there exists $0 \neq u \in E(\lambda)$ such that $Au = \lambda Iu$ (i.e. $a(u, v) = \lambda \langle u, v\rangle_U$ for all $v \in V$).*

b) *The spectrum $\sigma(A)$ of A consists of at most countably many eigenvalues which cannot have an accumulation point in $\mathbb{C}$.*

Proof: Since a is assumed to be V–coercive with respect to U we have already seen that $\tilde{a}(u, v) = a(u, v) + C_U \langle u, v\rangle_U$ is V–elliptic. Hence, the associated operator $\tilde{A} = A + C_U I$ has a bounded inverse. Using that I is a compact mapping we find that

$$K := (A + C_U I)^{-1} I$$

is a compact operator on V. Applying spectral theory for compact operators [**8**] we conclude that K has a pure point spectrum with zero as only possible accumulation point. Moreover, the eigenspaces are finite dimensional.

In order to use these results for the original mapping A, we consider the relation between eigenvalues of K and those of A. If I_V is the identity mapping on V we have

$$\begin{aligned} K - \mu I_V &= -\mu \left(I_V - \frac{1}{\mu}(A + C_U I)^{-1} I \right) \\ &= -\mu (A + C_U I)^{-1} \left(A + C_U I - \frac{1}{\mu} I \right) \\ &= -\mu (A + C_U I)^{-1} (A - \lambda I) \end{aligned}$$

where

$$\lambda = \frac{1}{\mu} - C_U \qquad \text{respectively} \qquad \mu = \frac{1}{\lambda + C_U}.$$

Given $\lambda \in \mathbb{C}$ and neglecting the case $\lambda = -C_U$ for which we know already that $A - \lambda I$ has a bounded inverse, there are two alternatives: if $\mu = 1/(\lambda + C_U) \notin \sigma(K)$ then K and thus also

$$A - \lambda I = -\frac{1}{\mu}(A + C_U I)(K - \mu I_V) \tag{3.22}$$

has a bounded inverse. On the other hand, if $\mu \in \sigma(K)$ then $\ker(K - \mu I_V) \neq \{0\}$ is a finite dimensional space. Due to (3.22), the kernel of $K - \mu I_V$ is in one to one correspondence with the kernel of $A - \lambda I$ (via the invertible mapping $A + C_U I$). Hence, the eigenspace $E(\lambda) = \ker(A - \lambda I)$ has the same properties as $\ker(K - \mu I_V)$. Finally, since $\mu = 0$ is the only possible accumulation point of $\sigma(K)$, the spectrum of A has no finite accumulation point. ∎

A simple consequence of Theorem 3.24 is given by

Corollary 3.25. *On top of the conditions of Theorem 3.24 we assume that $a(u,u) = 0$ for $u \in V$ implies $u = 0$. Then, the variational problem to find $u \in V$ such that for $l \in V'$ the relation $a(u,v) = l(v)$ holds for all $v \in V$ has a unique solution which depends continuously on the data.*

Proof: We will show that under the given assumptions $0 \notin \sigma(A)$ so that A^{-1} is a bounded operator (and hence $u = A^{-1}l$ is the unique solution with $\|u\|_V \leq \|A^{-1}\| \, \|l\|_{V'}$). If $0 \in \sigma(A)$ then we know from Theorem 3.24 that $\dim E(0) \geq 1$. Hence, there exists $w \neq 0$ such that $Aw = 0$. In particular,

$$0 = (Aw)(w) = \langle Aw, w \rangle_{V' \times V} = a(w,w)$$

which leads to the contradiction $w = 0$. ∎

3.5. Solving the Dirichlet problem

In order to get existence and uniqueness for solutions of the general Dirichlet problem (together with continuous dependence on the data) we have to check the assumptions of the Lax Milgram Theorem for the particular bilinear form

$$a(u,v) = \sum_{0 \leq |\sigma|,|\gamma| \leq m} \int_\Omega a_{\sigma_\gamma} \nabla^\gamma u \nabla^\sigma v \, dx \tag{3.23}$$

on the Hilbert space $V = H_0^m(\Omega)$. It turns out that ellipticity of the related operator

$$L(x, \nabla)u = \sum_{0 \leq |\sigma|,|\gamma| \leq m} (-1)^{|\sigma|} \nabla^\sigma (a_{\sigma_\gamma}(x) \nabla^\gamma u) \qquad x \in \Omega \tag{3.24}$$

is not sufficient to guarantee $H_0^m(\Omega)$–ellipticity of a (see Exercise 3.5). Recall that L is elliptic if the symbol $L^p(x, i\xi)$ of the principal part does not vanish on $R^n \backslash \{0\}$. For an operator in divergence form, the principal part is given by

$$L^p(x, \nabla) = \sum_{|\sigma|=|\gamma|=m} (-1)^{|\sigma|} a_{\sigma_\gamma}(x) \nabla^\sigma \nabla^\gamma$$

so that

$$L^p(x, i\xi) = \sum_{|\sigma|=|\gamma|=m} a_{\sigma_\gamma}(x) \xi^\sigma \xi^\gamma$$

and L is elliptic if

$$\sum_{|\sigma|=|\gamma|=m} a_{\sigma_\gamma}(x)\xi^\sigma\xi^\gamma > 0 \qquad \forall \xi \in \mathbb{R}^n\backslash\{0\},\, x \in \Omega$$

(we use the convention that the symbol should be positive). To show ellipticity of the associated bilinear form (3.23), we need the following stronger condition.

Definition 3.26. *The general operator L defined in* (3.24) *is called uniformly elliptic on $\Omega \subset \mathbb{R}^n$ if for some $\theta > 0$*

$$\sum_{|\sigma|=|\gamma|=m} a_{\sigma_\gamma}(x)\xi^\sigma\xi^\gamma \geq \theta\|\xi\|^{2m} \qquad \forall \xi \in \mathbb{R}^n.$$

For certain uniformly elliptic operators L we can then show that a is $H_0^m(\Omega)$ elliptic.

Lemma 3.27. *Let $\Omega \subset \mathbb{R}^n$ be open and assume the coefficients $a_{\sigma\gamma}$ of L in* (3.24) *are constants. Moreover, let $a_{00} > 0$ and $a_{\sigma\gamma} = 0$ for $0 < |\sigma|, |\gamma| < m$. If L is uniformly elliptic then the corresponding bilinear form a defined in* (3.23) *is $H_0^m(\Omega)$–elliptic. If Ω is bounded, the result remains true if $a_{00} = 0$.*

Proof: To check ellipticity, we take $u \in C_0^\infty(\Omega)$ and try to estimate $a(u,u)$ from below. By definition

$$a(u,u) = \sum_{|\sigma|,|\gamma|=m} a_{\sigma_\gamma} \int_\Omega \nabla^\gamma u \nabla^\sigma u\, dx + a_{00}\int_\Omega u^2\, dx.$$

Since u is compactly supported in Ω, we can replace the domain of integration by $\mathbb{R}^n$. Using standard properties of the Fourier transform

$$\hat{u}(\xi) = (\mathcal{F}u)(\xi) = \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} e^{-ix\cdot\xi}u(x)\, dx$$

we obtain

$$a(u,u) = \sum_{|\sigma|,|\gamma|=m} a_{\sigma_\gamma} \int_{\mathbb{R}^n} (i\xi)^\gamma \hat{u}\overline{(i\xi)^\sigma \hat{u}}\, d\xi + a_{00}\int_{\mathbb{R}^n} |\hat{u}|^2\, d\xi \tag{3.25}$$

where the bar denotes complex conjugation. We have used the property

$$\int_{\mathbb{R}^n} uv\, dx = \int_{\mathbb{R}^n} \hat{u}\overline{\hat{v}}\, d\xi$$

and the fact that $\mathcal{F}$ transforms derivatives into polynomials. (Note that for general operators $L(\nabla)$ with constant coefficients the transform $\mathcal{F}(L(\nabla)u) = L(i\xi)\hat{u}$ leads to the *symbol of L*. Applying $L(\nabla)$ to u thus corresponds to a *multiplication* of $\hat{u}$ with the symbol which is a polynomial in ξ. The principal part $L^p(i\xi)$ involves the highest order terms of this polynomial. For large ξ ('high frequencies'), the principal part obviously dominates the lower order terms and thus dictates the behavior of the differential operator.) We continue with the estimate (3.25) by using the uniform ellipticity of L

$$a(u,u) \geq \int_{\mathbb{R}^n} \left(\theta\|\xi\|^{2m} + a_{00}\right) |\hat{u}|^2 \, d\xi. \tag{3.26}$$

Assuming that $a_{00} > 0$, we easily see that

$$\theta\|\xi\|^{2m} - d_\alpha \xi^{2\alpha} + a_{00} \geq 0$$

for suitable $d_\alpha > 0$. Choosing the smallest constant d_α among $|\alpha| \leq m$, we obtain

$$\sum_{\|\alpha|\leq m} \xi^{2\alpha} \leq \frac{1}{C_2}\left(\theta\|\xi\|^{2m} + a_{00}\right)$$

with $C_2 > 0$. Using this estimate in (3.26) and transforming back yields

$$a(u,u) \geq C_2 \int_{\mathbb{R}^n} \sum_{\|\alpha|\leq m} |\xi^\alpha \hat{u}|^2 \, d\xi = C_2 \int_{\mathbb{R}^n} \sum_{\|\alpha|\leq m} |\nabla^\alpha u|^2 \, dx = C_2 \|u\|_{m,2}^2.$$

In the case where Ω is bounded and $a_{00} = 0$ we use $|\xi_i| \leq \|\xi\|$ so that

$$|\xi^\alpha| \leq \|\xi\|^m \qquad \forall |\alpha| = m$$

and thus

$$a(u,u) \geq \theta \int_{\mathbb{R}^n} \sum_{\|\alpha|=m} |\nabla^\alpha u|^2 \, dx = \theta |u|_{m,2}^2$$

with the notation from Theorem 3.10 (Poincaré inequality). Using the result of that theorem, we conclude further

$$a(u,u) \geq C_2 \|u\|_{m,2}^2.$$

Since $C_0^\infty(\Omega)$ is dense in $H_0^m(\Omega)$ the ellipticity estimate carries over to general H_0^m functions due to continuity. ∎

In the special case of second order operators we can easily generalize these considerations to x–dependent coefficients.

Lemma 3.28. *If L is a second order operator with the structure as in Lemma 3.27 but with coefficients in $\mathbb{L}^\infty(\Omega)$ (and $a_{00} \geq \mu > 0$ in the general case or $a_{00} \geq 0$ for bounded Ω), the corresponding bilinear form is $H_0^1(\Omega)$–elliptic.*

In this case, Fourier transform is not required because we can apply the ellipticity condition directly to $\xi = \nabla u$ since $\nabla^\sigma u = (\nabla u)^\sigma$ for $|\sigma| = 1$

$$\begin{aligned} a(u,u) &= \sum_{|\sigma|=|\gamma|=1} \int_\Omega a_{\sigma\gamma} (\nabla u)^\sigma (\nabla u)^\gamma \, dx + \int_\Omega a_{00} u^2 \, dx \\ &\geq \theta \int_\Omega \|\nabla u\|^2 dx + \int_\Omega a_{00} u^2 \, dx \end{aligned}$$

If $a_{00}(x) \geq \mu > 0$ then $a(u,u) \geq C_2 \|u\|_{1,2}^2$. Otherwise, the result follows with the Poincaré inequality. ■

If we include first order derivatives in the operator L, ellipticity of a is in general not given. For the special case of first order derivatives with constant coefficients, we find a similar result as in Lemma 3.28 (see Exercise 3.4). For the general second order operator

$$L(x,\nabla)u := -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial u}{\partial x_j} + \sum_{i=1}^{n} b_i(x) \frac{\partial u}{\partial x_i} + c(x)u, \tag{3.27}$$

with associated bilinear form

$$a(u,v) = \sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i}\, dx + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v\, dx + \int_\Omega cuv\, dx \tag{3.28}$$

we typically get coercivity results with respect to $\mathbb{L}^2(\Omega)$.

Theorem 3.29. *Let Ω be a bounded domain in $\mathbb{R}^n$ and L defined in (3.27) be uniformly elliptic on Ω with $\mathbb{L}^\infty(\Omega)$ coefficients. Then there exists C_3 and $\hat{\lambda} \geq 0$ such that*

$$a(u,u) + \hat{\lambda}\|u\|_{\mathbb{L}^2}^2 \geq C_3 \|u\|_{1,2}^2 \qquad \forall u \in H^1(\Omega)$$

where a is given in (3.28).

Proof: Apart from the estimate of the principal part used in Lemma 3.28, we have

$$\begin{aligned} &+\sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v\, dx + \int_\Omega cuv\, dx \\ &\qquad\qquad \geq -\max_{i=1}^{n} \|b_i\|_{\mathbb{L}^\infty} \int_\Omega \|\nabla u\|\ |u|\, dx - \|c\|_{\mathbb{L}^\infty} \int_\Omega u^2\, dx. \end{aligned}$$

Applying Cauchy-Schwarz inequality, we get

$$-\int_\Omega \|\nabla u\|\ |u|\, dx \geq -\|\nabla u\|_{\mathbb{L}^2} \|u\|_{\mathbb{L}^2}.$$

The trick

$$-ab = -(\epsilon a)\left(\frac{b}{\epsilon}\right) \geq -\frac{1}{2}\left(\epsilon^2 a^2 + \frac{b^2}{\epsilon^2}\right)$$

then yields

$$-\|\nabla u\|_{\mathbb{L}^2} \|u\|_{\mathbb{L}^2} \geq -\frac{1}{2}\epsilon^2 \|\nabla u\|_{\mathbb{L}^2}^2 - \frac{1}{2\epsilon^2}\|u\|_{\mathbb{L}^2}^2.$$

Choosing ϵ^2 small enough we get

$$a(u,u) \geq \frac{1}{2}\theta \|\nabla u\|_{\mathbb{L}^2}^2 - \tilde{C}\|u\|_{\mathbb{L}^2}^2 \geq C_3 \|u\|_{1,2}^2 - \hat{\lambda}\|u\|_{\mathbb{L}^2}^2.$$

■

According to our considerations in Section 3.4 coercivity enables us to apply spectral theory. In particular, if the spectrum of L (as mapping from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$) does not contain $\lambda = 0$, then the original Dirichlet problem has a unique solution which corresponds continuously on the data. (Even if $0 \in \sigma(L)$, spectral theory yields results about the solution space $\{u \in H_0^m(\Omega) : Lu = f\}$.) In view of these remarks, the following theorem is very important for the case of general uniformly elliptic operators.

Theorem 3.30. **(Gårding's inequality)**
Let $\Omega \subset \mathbb{R}^n$ be a bounded open set and let L in (3.24) be uniformly elliptic with coefficients of the principal part in $C(\bar{\Omega})$ and all other coefficients in $\mathbb{L}^\infty(\Omega)$. Then a given in (3.23) is $H_0^m(\Omega)$–coercive with respect to $\mathbb{L}^2(\Omega)$, i.e.

$$a(u,u) + C_{\mathbb{L}^2}\|u\|_{\mathbb{L}^2}^2 \geq C_{H^m}\|u\|_{H^m}^2 \qquad u \in H_0^m(\Omega)$$

Proof: For the complete proof see the articles quoted in [**5**]. Here, we just consider the case where the coefficients of the principal part are constant. (The full proof reduces the general case to this situation by a partition of unity argument and exploitation of the uniform continuity of the coefficients of the principal part.) In Lemma 3.27 we have already seen that the principal part satisfies

$$a^p(u,u) = \sum_{|\sigma|,|\gamma|=m} a_{\sigma_\gamma} \int_\Omega \nabla^\gamma u \nabla^\sigma u \, dx \geq C_2 \|u\|_{m,2}^2.$$

The remaining terms are then estimated by Cauchy-Schwarz inequality and the $\mathbb{L}^\infty$ bounds on the coefficients

$$\begin{aligned} |a(u,u) - a^p(u,u)| &\leq \sum_{\substack{0 \leq |\sigma|,|\gamma| \leq m \\ |\sigma|+|\gamma|<2m}} \int_\Omega \left| a_{\sigma_\gamma} \nabla^\gamma u \nabla^\sigma u \right| \, dx \\ & C \sum_{\substack{0 \leq |\sigma|,|\gamma| \leq m \\ |\sigma|+|\gamma|<2m}} \|\nabla^\gamma u\|_{\mathbb{L}^2} \|\nabla^\sigma u\|_{\mathbb{L}^2} \\ &\leq C\|u\|_{m,2}\|u\|_{m-1,2}. \end{aligned}$$

Using the ϵ–trick, we can separate the product into a sum of squares with an arbitrarily small contribution to the $\|\cdot\|_{m,2}$ norm

$$|a(u,u) - a^p(u,u)| \leq \frac{\epsilon}{2}\|u\|_{m,2}^2 + C(\epsilon)\|u\|_{m-1,2}^2.$$

Since $X = H^m(\Omega)$ is compactly embedded in $Y = H^{m-1}(\Omega)$ and Y is continuously embedded in $Z = \mathbb{L}^2(\Omega)$, we can find for every $\epsilon > 0$ a constant $D(\epsilon)$ such that

$$\|u\|_Y \leq \epsilon\|u\|_X + D(\epsilon)\|u\|_Z \qquad \forall u \in X. \tag{3.29}$$

(This is *Ehrling's Lemma* which is proved by contradiction: we assume that there is $\epsilon_0 > 0$ for which (3.29) fails. Then, there is a sequence $(x_n)_{n\in\mathbb{N}}$ in X

such that $\|x_n\|_X = 1$ and

$$\|x_n\|_Y > \epsilon_0 + n\|x_n\|_Z. \tag{3.30}$$

Since $X \hookrightarrow Y$, $(x_n)_{n\in\mathbb{N}}$ is bounded in Y so that $(x_n)_{n\in\mathbb{N}}$ converges to zero in Z. A convergent subsequence in Y (which exists since $X \hookrightarrow Y$ is compact) must then also converge to zero which contradicts (3.30).) Using this result, we get

$$\|u\|_{m-1,2}^2 \le \frac{\epsilon}{2}\|u\|_{m,2}^2 + D(\epsilon)\|u\|_{\mathbb{L}^2}^2$$

and thus

$$\begin{aligned} a(u,u) &\ge a^p(u,u) - |a(u,u) - a^p(u,u)| \\ &\qquad\qquad \ge C_2\|u\|_{m,2}^2 - \left(\epsilon\|u\|_{m,2}^2 + E(\epsilon)\|u\|_{\mathbb{L}^2}^2\right). \end{aligned}$$

Hence, for ϵ small enough we get coercivity of a. ■

3.6. Other boundary conditions

3.6.1. Natural boundary conditions. In Section 3.2 we have seen that the choice of the Hilbert space $V = H_0^m(\Omega)$ is related to homogeneous Dirichlet conditions. Similarly, the unrestricted case $V = H_0^m(\Omega)$ implies certain boundary conditions. Since they are not explicitly given through the definition of V they are called *natural boundary conditions.* For the Helmholtz and Poisson problem we have already seen that the natural conditios are Neumann conditions. In general, the natural boundary conditions are more complicated because they are determined by the coefficients of L so that especially for higher order operators the boundry conditions involve many parameters. For simplicity, we will only consider second order operators in divergence form

$$Lu = -\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}u\right) - \frac{\partial}{\partial x_i}(a_{i0}u) + a_{0j}\frac{\partial}{\partial x_j} + a_{00}u$$

with corresponding bilinear form

$$a(u,v) = \int_\Omega a_{ij}\frac{\partial v}{\partial x_i}\frac{\partial u}{\partial x_j} + a_{i0}\frac{\partial v}{\partial x_i}u + a_{0j}\frac{\partial u}{\partial x_j}v + a_{00}uv\,dx$$

(Einstein's summation convention is used). If $u \in H^1(\Omega)$ is a solution of the abstract variational problem

$$a(u,v) = l(v) \qquad \forall v \in H^1(\Omega) \tag{3.31}$$

and if u, the coefficients of L, the right hand side and the boundary of the domain have sufficient regularity, we can recover the classical formulation of the problem (3.31). Using $v \in C_0^\infty(\Omega)$ and assuming $l(v) = \langle f, v\rangle_0$ we get after integration by parts

$$\int_\Omega (Lu - f)v\,dx = 0 \qquad \forall v \in C_0^\infty(\Omega)$$

which implies $Lu = f$. If we now take $v \in C^\infty(\bar\Omega)$ in (3.31) integration by parts gives rise to boundary terms

$$0 = a(u,v) - \langle f, v\rangle_0 = \int_\Omega (Lu - f)v\,dx + \int_{\partial\Omega}\left(a_{ij}\frac{\partial u}{\partial x_j} + a_{i0}u\right) n_i v\,d\sigma.$$

Since $Lu - f = 0$ and traces of $C(\bar\Omega)$ functions are dense in $\mathbb{L}^2(\partial\Omega)$ we conclude

$$\left(a_{ij}\frac{\partial u}{\partial x_j} + a_{i0}u\right) n_i = 0. \tag{3.32}$$

Introducing the matrix $A_L = (a_{ij})$ of the second order coefficients of L as well as the vector $a_L = (a_{i0})$ we can rewrite (3.32) in the compact form

$$(A_L \nabla u)\cdot n + (a_L \cdot n)u = 0$$

If $L = -\Delta$ then $a_{ij} = \delta_{ij}$ and $a_L = 0$ so that the natural boundary conditions are the well known Neumann conditions $0 = n\cdot\nabla u = \partial u/\partial n$. In order to get inhomogeneous boundary conditions

$$(A_L \nabla u)\cdot n + (a_L \cdot n)u = g$$

with $g \in \mathbb{L}^2(\partial\Omega)$ it suffices to modify the right hand side of (3.31) by adding an extra linear form

$$v \mapsto \int_{\partial\Omega} gBv\,d\sigma$$

where B is the trace operator which is continuous on $H^1(\Omega)$ (if $\partial\Omega$ is regular enough).

3.6.2. Mixed boundary conditions. If Γ_N is an open subset of $\partial\Omega$ with complement Γ_D we can consider the problem to find a function u such that $Lu = f$ in Ω with Dirichlet conditions on Γ_D and natural conditions on Γ_N. To get a weak formulation we obviously have to generalize the space $H_0^m(\Omega)$ to the case where the zero conditions are only imposed on $\Gamma_D \subset \partial\Omega$. Similar to the definition of H_0^m we introduce the space $H^m_{\Gamma_D}(\Omega)$ as the H^m–closure of all $C^\infty(\bar\Omega)$ functions which vanish in a neighborhood of Γ_D. Then, the variational problem to find $u \in H^m_{\Gamma_D}(\Omega)$ such that $a(u,v) = l(v)$ for all $v \in H^m_{\Gamma_D}(\Omega)$ is the natural weak formulation of the mixed boundary value problem.

3.6.3. General boundary conditions. A linear elliptic operator L of order $2m$ requires in general m boundary conditions to obtain a well posed boundary value problem. The corresponding boundary operators B_j are of the form

$$B_j u = \sum_{|\alpha|\le m_j} b_{j\alpha} B(\nabla^\alpha u) \qquad j = 1,\dots,m$$

where B is the trace operator and $0 \le m_j < 2m$. Of course, $\{B_j : j = 1,\dots,m\}$ must be linearly independent but independency might not be sufficient to obtain well posed problems. A characterization of reasonable boundary conditions is given by the so called *complementing condition for elliptic systems* (see [**14**] for an introduction). For second order operators a single boundary condition is required. Since the order of the boundary operator is at most one, the most general case is given by

$$B_{gen}u := b_{gen} \cdot \nabla u + c_{gen}u.$$

We assume that the vector b_{gen} always has a normal component, i.e.

$$b_{gen}(x) \cdot n(x) \geq \mu > 0 \qquad \forall x \in \partial\Omega \tag{3.33}$$

Note that the natural boundary conditions are of the same form if we set $b_{nat} = {A_L}^T n$ and $c_{nat} = a_L \cdot n$ since

$$n \cdot (A_L \nabla u) + (a_L \cdot n)u = b_{nat} \cdot \nabla u + c_{nat}u.$$

If L is uniformly elliptic, the vector b_{nat} always has a normal component. Indeed,

$$b_{nat} \cdot n = ({A_L}^T n) \cdot n = \sum_{i,j=1}^{n} a_{ij} n_i n_j \geq \theta \|n\|^2 = \theta > 0. \tag{3.34}$$

Starting from the bilinear form a which is related to L and the natural conditions we will apply modifications which leave L unchanged but have some effect on the bilinear form and thus also on the boundary conditions.

The problem can be seperated into three steps. In the first step we transform c_{nat} into the required function c_{gen} without modifying the first order term $b_{nat} \cdot \nabla u$. In a second step, we adjust the first order term without changing the zeroth order part so that both steps are independent of each other. Finally, the results are combined and the condition $b_{gen} \cdot n = b_{nat} \cdot n$ which we require in the second step is relaxed. We remark that the modifications presented here are not unique. However, they are done in such a way that the resulting bilinear form is $H^1(\Omega)$ coercive with respect to $\mathbb{L}^2(\Omega)$ if the same holds for the original form a.

We start with the observaion that due to the product rule

$$L_1 u = -\frac{\partial}{\partial x_i}(\beta_i u) + \beta_i \frac{\partial}{\partial x_i} u + u \frac{\partial}{\partial x_i} \beta_i = 0$$

whenever u and β are smooth enough (we still use Einstein's summation convention over repeated indices). Hence, the modified operator $L + L_1$ acts exactly like L but leads to the different bilinear form $\tilde{a}(u, v) = a(u, v) + a_1(u, v)$ with

$$a_1(u, v) = \int_\Omega \beta_i u \frac{\partial}{\partial x_i} v + \beta_i v \frac{\partial}{\partial x_i} u + uv \frac{\partial}{\partial x_i} \beta_i \, dx.$$

The associated natural boundary conditions are given by

$$\tilde{B}u = b_{nat} \cdot \nabla u + c_{nat}u + (\beta \cdot n)u = 0.$$

Now we use the freedom in β to adjust the zeroth order part. If

$$\beta(x) = c_{gen}(x) n(x) - a_L(x) \qquad \forall x \in \partial\Omega$$

then $\beta \cdot n = c_{gen} - a_L \cdot n = c_{gen} - c_{nat}$ showing that $\tilde{B}u = b_{nat} \cdot \nabla u + c_{gen} u$. Of course, the problem is to find a smooth function β in Ω which has the required values on the boundary. If $\partial\Omega$ and the baoundary data are smooth enough such a function can always be constructed by extension along the normal direction and smooth cut off. Since the bilinear form corresponding to L_1

$$a_1(u, v) = \int_\Omega \beta_i u \frac{\partial}{\partial x_i} v + \beta_i v \frac{\partial}{\partial x_i} u + uv \frac{\partial}{\partial x_i} \beta_i \, dx$$

involves only first order terms we can use the standard trick to estimate

$$\begin{aligned} a_1(u, u) &\geq -2\|\beta\|_{\mathbb{L}^\infty} \|u\|_{\mathbb{L}^2} \|u\|_{H^1} - \|\operatorname{div} \beta\|_{\mathbb{L}^\infty} \|u\|^2_{\mathbb{L}^2} \\ &\geq -\epsilon \|u\|^2_{H^1} - C(\epsilon) \|u\|^2_{\mathbb{L}^2}. \end{aligned}$$

Hence, if a is H^1–coercive with respect to $\mathbb{L}^2$ we can choose ϵ so small that $a + a_1$ has the same property.

In the next step we use a similar approach to fix the first order part of the boundary condition. Assuming that $b_{nat} \cdot n = b_{gen} \cdot n$, we have

$$b_{gen} \cdot \nabla + c_{nat} = b_{nat} \cdot \nabla + c_{nat} + d \cdot \nabla$$

where $d = b_{gen} - b_{nat}$ is a tangential vector. The trick is again to find a bilinear form a_2 which has the additional term $d \cdot \nabla$ as natural boundary condition so that together with a the modified conditions are obtained. Since the first order term in the natural boundary conditions has the structure $b_{nat} = {A_L}^T n$ we have to find an operator L_2 such that $L_2 u = 0$ and ${A_{L_2}}^T n = d$. The second condition is easily satisfied by the tensor product $(n \otimes d)_{ij} = n_i d_j$. However, to keep coercivity and to satisfy the condition $L_2 u = 0$ we use the antisymmetric matrix

$$A^{(2)}_{ij} = n_i d_j - n_j d_i.$$

Since $n \perp d$, we have ${A^{(2)}}^T n = -A^{(2)} n = -(n \cdot d) n + (n \cdot n) d = d$. As in the first step the main problem is to extend the matrix valued function $A^{(2)}$ to a function in Ω (keeping antisymmetry). Having achieved that, we define the operator L_2 as

$$L_2 u = -\frac{\partial}{\partial x_i} \left(A^{(2)}_{ij} \frac{\partial}{\partial x_j} u \right) + \left(\frac{\partial}{\partial x_i} A^{(2)}_{ij} \right) \frac{\partial}{\partial x_j} u.$$

Due to the antisymmetry of $A^{(2)}$ we have

$$L_2 u = -A^{(2)}_{ij} \frac{\partial^2}{\partial x_i \partial x_j} u = -\frac{1}{2} \left(A^{(2)}_{ij} + A^{(2)}_{ji} \right) \frac{\partial^2}{\partial x_i \partial x_j} u = 0.$$

Hence, by construction, the modified operator $L + L_2$ acts exactly like L (on smooth functions) and gives rise to the bilinear form $\bar{a}(u, v) = a(u, v) + a_2(u, v)$ with

$$a_2(u, v) = \int_\Omega A^{(2)}_{ij} \frac{\partial v}{\partial x_i} \frac{\partial u}{\partial x_j} + \left(\frac{\partial}{\partial x_i} A^{(2)}_{ij} \right) v \frac{\partial}{\partial x_j} u \, dx.$$

The natural boundary conditions of the combined form $\bar{a}$ is

$$b_{nat} \cdot \nabla + c_{nat} + ({A^{(2)}}^T n) \cdot \nabla = (b_{nat} + d) \cdot \nabla + c_{nat} = b_{gen} \cdot \nabla + c_{nat}.$$

Note that again coercivity is not destroyed since

$$\begin{aligned}
a_2(u,u) &= \int_\Omega A^{(2)}_{ij} \frac{\partial v}{\partial x_i} \frac{\partial u}{\partial x_j} + \left(\frac{\partial}{\partial x_i} A^{(2)}_{ij} \right) v \frac{\partial}{\partial x_j} u \, dx \\
&= \int_\Omega \frac{1}{2} \left(A^{(2)}_{ij} + A^{(2)}_{ji} \right) \frac{\partial v}{\partial x_i} \frac{\partial u}{\partial x_j} + \left(\frac{\partial}{\partial x_i} A^{(2)}_{ij} \right) v \frac{\partial}{\partial x_j} u \, dx \\
&= \int_\Omega \left(\frac{\partial}{\partial x_i} A^{(2)}_{ij} \right) v \frac{\partial}{\partial x_j} u \, dx \\
&\geq -C \|u\|_{\mathbb{L}^2} \|u\|_{H^1} \geq -\epsilon \|u\|^2_{H^1} - C(\epsilon) \|u\|^2_{\mathbb{L}^2}.
\end{aligned}$$

In the last step we combine the previous results. First, the boundary condition

$$b_{gen} \cdot \nabla u + c_{gen} u = 0$$

is modified by multiplying with $\eta = (b_{nat} \cdot n)/(b_{gen} \cdot n)$. According to (3.34) and assumption (3.33) we have $\eta > \mu/\theta > 0$ so that

$$(\eta b_{gen}) \cdot \nabla u + (\eta c_{gen}) u = 0 \tag{3.35}$$

is an equivalent condition for which we construct a_1 and a_2 as before (observe that $(\eta b_{gen}) \cdot n = b_{nat} \cdot n$). The resulting bilinear form $a + a_1 + a_2$ on the Hilbert space $V = H^1(\Omega)$ yields a weak formulation of the general boundary value problem. As we have seen before, coercivity of a carries over to the new bilinear form. Also, inhomogeneous conditions can be obtained by adding the functional

$$v \mapsto \int_{\partial\Omega} \eta g B v \, d\sigma \qquad v \in H^1(\Omega)$$

on the right hand side of the abstract variational problem.

3.7. The Ritz–Galerkin Method

In the theory of classical solutions it is natural to use approximation procedures like Finite Difference Methods which are based on a point wise evaluation of functions and differential operators. When dealing with weak solutions, this approach cannot be taken over because point values of functions in $H^m(\Omega)$ are in general not well defined. In fact, the formulation of the original problem as abstract variational problem suggests a different strategy to convert the infinite dimensional setup into a finite dimensional one which then allows a numerical treatment. If the abstract variational problem is to find u in a Hilbert space V such that for some $l \in V'$

$$a(u, v) = l(v) \qquad \forall v \in V$$

then a discretization is obtained by replacing V with a finite dimensional subspace V_h. (Here, $h > 0$ is related to the dimension of V_h and will later play the role of a discretization length.) To get a numerical approximation, the idea of the *Ritz–Galerkin Method* is to find $u_h \in V_h$ such that

$$a(u_h, v) = l(v) \qquad \forall v \in V_h \tag{3.36}$$

which is a finite dimensional problem. By introducing a basis $\{\psi_1, \psi_2, ..., \psi_N\}$ of V_h we can reformulate (3.36) as

$$a(u_h, \psi_i) = l(\psi_i), \quad i = 1, \ldots, N.$$

Using

$$u_h = \sum_{j=1}^{N} z_j \psi_j$$

we get the system of equations

$$\sum_{j=1}^{N} a(\psi_j, \psi_i) z_j = l(\psi_i), \quad i = 1, \ldots, N$$

which we may write in the form

$$Az = b$$

with the matrix $A_{ij} = a(\psi_j, \psi_i)$ and right hand side $b_i = l(\psi_i)$. In engineering science the matrix A is called the *stiffness matrix*. Note that A is positive definite if a is V–elliptic since

$$z^T Az = \sum_{i,j} z_i A_{ij} z_j = a\left(\sum_j z_j \psi_j, \sum_i z_i \psi_i\right) = a(u_h, u_h) \geq C_2 \|u_h\|_V^2$$

In particular, $Az = 0$ implies $u_h = 0$ and thus also $z = 0$ which shows that A is invertible. Consequently, the Ritz–Galerkin solution always exists if a is elliptic. Moreover, if a is symmetric, i.e. $a(u, v) = a(v, u)$ for all $u, v \in V$ then A is a symmetric matrix. The basic estimate on the quality of a numerical approximation u_h based on the sub space V_h is the following lemma.

Lemma 3.31. (Céa Lemma)
Let a be a bilinear, V-elliptic form and let u and u_h be the solutions of the variational problem in V and $V_h \subset V$. Then

$$\|u - u_h\|_V \leq \frac{C_1}{C_2} \inf_{v_h \in V_h} \|u - v_h\|_V$$

where C_1 is the constant of continuity and C_2 the constant of ellipticity of a.

Proof: By definition we have $a(u, v) = l(v)$ for all $v \in V$ and $a(u_h, v) = l(v)$ for all $v \in V_h$. Since $V_h \subset V$ we can subtract the two equations which yields

$$a(u - u_h, v) = 0 \qquad \forall v \in V_h.$$

Now, let $v_h \in V_h$. With $v = v_h - u_h \in V_h$ we have $a(u - u_h, v_h - u_h) = 0$ and

$$\begin{aligned} C_2\|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &\leq C_1\|u - u_h\|_V\|u - v_h\|_V \end{aligned}$$

Hence, it follows that $C_2\|u - u_h\|_V \leq C_1\|u - v_h\|_V$ for all $v_h \in V_h$. ■

The Céa Lemma gives information about the convergence of the Ritz–Galerkin method. Obviously, the distance between approximation u_h and exact solution u vanishes if the distance of u to the space V_h becomes arbitrarily small. (The distance of a point u to a set V_h is defined by $d(u, V_h) := \inf\{\|u - v_h\|_V : v_h \in V_h\}$.) Consequently, convergence can only be obtained if the spaces V_h become larger for $h \to 0$ and eventually exhaust the Hilbert space V.

Corollary 3.32. *Let $V_h \subset V$ $(h > 0)$ be a family of subspaces with the property*

$$d(u, V_h) \xrightarrow[h \to 0]{} 0 \qquad \forall u \in V.$$

If a is a V–elliptic bilinear form and if u is the solution of an abstract variational problem related to a, the Ritz–Galerkin approximation u_h converges

$$\|u - u_h\|_V \xrightarrow[h \to 0]{} 0.$$

Proof: Immediate consequence of Lemma 3.31. ■

The Ritz–Galerkin Method turns into an efficient solution algorithm if V_h has good approximation properties and if the stiffness matrix A remains either small or sparse. The latter property is obtained if the basis is chosen almost a–orthogonal so that each row of A has only a few non zero entries. Apart from that, it is useful if the basis is related to the problem under consideration. More precisely, if V is a function space, the basis functions should be selected in such a way that high resolution is possible where the solution is expected to develop important structures. In the next section, a few exemplary choices of V_h are discussed.

3.8. Examples of Finite Element spaces

3.8.1. First remarks. In a first example, we consider a Helmholtz problem on the unit interval with homogeneous Dirichlet conditions

$$\begin{aligned} -u'' + u &= f \qquad \text{in } (0,1) \\ u &= 0 \qquad \text{on } \{0,1\} \end{aligned} \tag{3.37}$$

The corresponding weak formulation is to find $u \in H_0^1(0,1)$ such that

$$\langle u, v\rangle_1 = \langle f, v\rangle_0 \qquad \forall v \in H_0^1(0,1).$$

As finite dimensional subspaces we choose

$$V_h = \{\theta_n \,:\, n = 1, \dots, N\} \qquad h = 1/N$$

where θ_n are orthonormal basis functions in $H_0^1(0,1)$

$$\theta_n(x) = \left(\frac{2}{1+n^2\pi^2}\right)^{\frac{1}{2}} \sin(n\pi\, x).$$

(Here, h does not correspond to a discretization length. For a derivation of the orthonormal basis see Exercise C.4.) This particular choice is very well suited for the Helmholtz problem because θ_n are eigenfunctions of the Helmholtz operator. Indeed,

$$-\theta_n'' + \theta_n = (1 + n^2\pi^2)\theta_n \qquad \forall n \in \mathbb{N}.$$

Another side effect is that the nested sequence of spaces $(V_{1/N})_{N\in\mathbb{N}}$ exhausts $H_0^1(0,1)$ because each element can be represented in terms of the basis vectors. Finally, the orthonormality of θ_n greatly simplifies the structure of the stiffness matrix. We have

$$A_{ij} = a(\theta_j, \theta_i) = \langle \theta_j, \theta_i\rangle_1 = \delta_{ij}$$

so that A is the identity matrix. Hence, the Ritz–Galerkin solution is just

$$u_h = \sum_{n=1}^{N} z_n\theta_n \qquad \text{with} \qquad z_n = b_n = \langle f, \theta_n\rangle_0\,.$$

Since $\theta_n \in C^\infty(0,1)$, the approximation u_h is a smooth function and we can check in which sense it approximates a solution of the problem in the classical formulation. Since θ_n are eigenfunctions of the Helmholtz operator, we have

$$-u_h'' + u_h = \sum_{n=1}^{N} (1 + n^2\pi^2)\, \langle f, \theta_n\rangle_0\, \theta_n = \sum_{n=1}^{N} \langle f, \psi_n\rangle_0\, \psi_n$$

where $\psi_n = \sqrt{2}\sin(n\pi\, x)$ forms an orthonormal basis of $\mathbb{L}^2(0,1)$ (see Exercise C.4). Note that

$$\sum_{n=1}^{N} \langle f, \psi_n\rangle_0\, \psi_n(x) = 2\sum_{n=1}^{N} \int_0^1 f(y) \sin(n\pi\, y)\, dy \sin(n\pi\, x)$$

is the truncated Fourier sine series of f. In particular, we see that the weak solution of the Helmholtz problem is the limiting case of classical solutions to Helmholtz problems with modified right hand sides. (There is a close relation to the *separation of variables* approach in Section 1.3.1.) Rewriting the approximate solution in terms of ψ_n we find

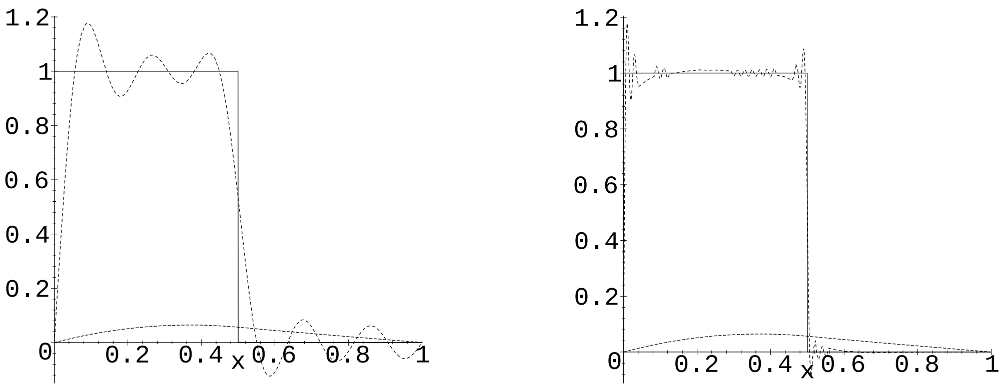
$$u_h = \sum_{n=1}^{N} \frac{1}{1+n^2\pi^2} \langle f, \psi_n \rangle_0 \, \psi_n$$

which shows that u_h is obtained from the Fourier sine series of f by damping highly oscillatory terms (large n) which are responsible for a possible rough behavior of f. Consequently, u_h is much smoother than the right hand side which demonstrates the smoothing properties of the inverse Dirichlet Helmholtz operator.

For the case

$$f(x) = \begin{cases} 1 & 0 \le x \le \frac{1}{2} \\ 0 & \frac{1}{2} < x \le 1 \end{cases}$$

with $N = 10$ and $N = 100$, the approximation u_h, the truncated Fourier sine series of f and f itself are shown in the figures below.



$\dim V_h = 10$ (left) and $\dim V_h = 100$ (right)

The simplicity of our first example is related to the availability of a basis which is orthonormal with respect to the bilinear form of the abstract variational problem. As a consequence, $A_{ij} = a(\theta_j, \theta_i)$ is the identity matrix so that the linear problem which determines u_h is trivial. However, for more general geometries or other differential operators the construction of an orthonormal basis is as difficult as solving the original problem so that we have to resort to other basis functions. As a model problem for the general case we consider the Poisson equation on the unit square with homogeneous Dirichlet boundary conditions,

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega = [0,1]^2 \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{3.38}$$

The restriction to the 2D case is only for simplicity. It should be clear, how the ideas can be generalized to other dimensions.

We know from Section 3.1.4 that the bilinear form corresponding to (3.38) is $a(u, v) = \langle \nabla u, \nabla v \rangle_0$ on $H_0^1(\Omega)$. Although a sine basis is again available (see Section 1.3.1), we choose V_h as the set of all functions $v \in C^0(\Omega)$, which are linear in each triangle of the triangulation shown in Fig. 3.1 and which vanish on the boundary of Ω.

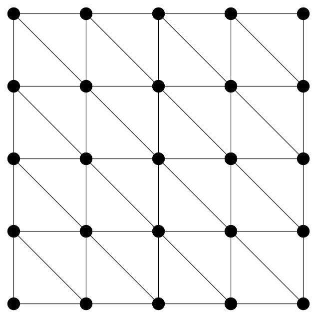


Figure 3.1: Triangulation of the Unit Square

More precisely, in each triangle, a function $v \in V_h$ has the form

$$v(x, y) = a + bx + cy$$

and the three values at the corners of a triangle determine the parameters a, b and c. Consequently, v is given by the values at the N interior grid points (x_j, y_j), $j = 1, \ldots, N$. The dimension of V_h is equal to N and we choose a basis ψ_i, $i = 1, \ldots, N$ with $\psi_i(x_j, y_j) = \delta_{ij}$. Since the basis functions should be linear on each triangle, the support of an exemplary basis function at point number C is non zero on all triangles apart from II and V (see Figure 3.2).

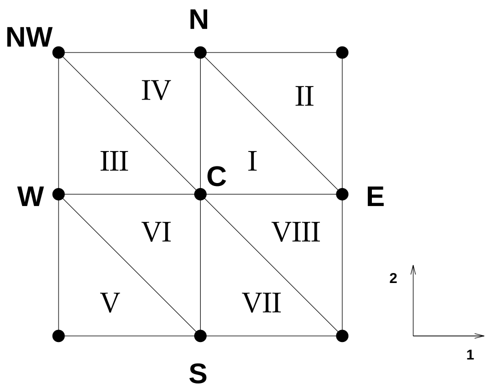


Figure 3.2: Ordering of triangles around point C

Since the values of ψ_C drop linearly from one at C to zero at the neighboring points, the partial derivatives are either zero or constant $\pm 1/h$ on the neighboring triangles ($h > 0$ denotes the spacing of grid points). In the following table, we list the derivatives of the basis function ψ_C on the triangles depicted in Figure 3.2

	I	II	III	IV	V	VI	VII	VIII
$\partial_1\psi_C$	$-h^{-1}$	0	h^{-1}	0	0	h^{-1}	0	$-h^{-1}$
$\partial_2\psi_C$	$-h^{-1}$	0	0	$-h^{-1}$	0	h^{-1}	h^{-1}	0
$\|\nabla\psi_C\|^2$	$2h^{-2}$	0	h^{-2}	h^{-2}	0	$2h^{-2}$	h^{-2}	h^{-2}

The directions 1 and 2 are chosen as indicated in Figure 3.2. Using the fact that the area of a single triangle is $h^2/2$, we get

$$a(\psi_C, \psi_C) = \int\limits_{I,\ldots,VIII} |\nabla\psi_C|^2\, dxdy = 8h^{-2}\frac{1}{2}h^2 = 4.$$

The basis functions are not a–orthogonal because the constant derivatives of neighboring basis functions overlap giving contributions to the bilinear form. For example, ψ_N and ψ_C have overlapping support on the triangles I and IV. We find

$$a(\psi_C, \psi_N) = \int\limits_{I+IV} (\nabla\psi_C)\cdot(\nabla\psi_N)dxdy$$

$$= \int\limits_{I+IV} \partial_2\psi_C\partial_2\psi_N dxdy = \int\limits_{I+IV} (-h^{-1})h^{-1}\mathrm{d}x\mathrm{d}y = -1$$

Similarly, we obtain

$$a(\psi_C, \psi_E) = a(\psi_C, \psi_S) = a(\psi_C, \psi_W) = a(\psi_C, \psi_N) = -1$$

For the remaining corners, one computes

$$a(\psi_C, \psi_{NW}) = \int\limits_{III+IV} (\partial_1\psi_C\partial_1\psi_{NW} + \partial_2\psi_C\partial_2\psi_{NW})\,\mathrm{d}x\mathrm{d}y = 0$$

and the same for the other points NE, ES and WS. Hence, in the i^{th} row of the matrix $A_{ij} = a(\psi_j, \psi_i)$ there are five non zero entries referring to the point i and its neighboring points in north, south west and east direction. Using the stencil notation to describe the action of A in the neighborhood of the point i, we recover the standard 5 point stencil

$$\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$$

as in the Finite Difference approach. Consequently, in this particular case, the matrix A is the same as in the standard Finite Difference scheme (which is not true in general). Finally, the right hand side of the Ritz–Galerkin Method is obtained by

$$b_i = l(\psi_i) = \int_\Omega f(x)\psi_i(x)\,dx.$$

Observe that, in contrast to the finite difference approach, we use integral averages of f instead of point values.

The resulting system $Az = b$ now has to be solved numerically in order to find the Ritz Galerkin approximation. Note that due to the properties of a, the matrix A is symmetric and positive definite. The system consists of N equations where N is the number of interior nodes. Since the basis functions all have very little overlap, the resulting matrix A is sparse.

The finite dimensional space V_h of piecewise linear functions on Ω is a typical example of a *Finite Element Space*. In general, finite element spaces are characterized by:

(i) The partition of the domain Ω:
The domain Ω is represented as the union of simple subdomains (triangles, rectangles, etc.). The single subdomains are called *elements*.
(ii) The basis functions:
Typically, the basis functions are defined on each element as polynomials which match at the interfaces up to a certain order of differentiability. The polynomial building blocks are called *shape functions*.

Definition 3.33. *A partition $\mathcal{T} = \{T_1, T_2, ..., T_M\}$ of Ω into elements elements is called admissible, if $\bar{\Omega} = \bigcup_{i=1}^{M} T_i$ and if $T_i \cap T_j \neq \emptyset$ the intersection must be a common vertex or a common edge of the elements. If $h > 0$ is the maximal side length of all elements of $\mathcal{T}$, we also write $\mathcal{T}_h$ instead of $\mathcal{T}$.*

Note that a situation as shown in Fig. 3.3 is ruled out in an admissible partition.

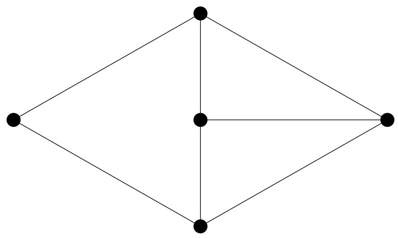


Figure 3.3: Non–admissible partition into three triangles

The reason for the exclusion is that on such a configuration, for example, a piecewise linear function is, in general, not continuous.

3.8.2. A complete example. To give a detailed analysis of a finite element space, we focus on the simple situation of *linear, triangular elements* for Ritz Galerkin approximations in $V = H_0^1(\Omega)$ (i.e. the elements are triangles and the shape functions are linear polynomials). To ensure that we can find a partition of Ω into triangles, we assume that $\Omega \subset \mathbb{R}^2$ is a bounded set with a polygonal boundary. However, the triangles need not be completely regular, as in the second example of the last section. In general, we allow any distortion of the triangles as long as the partition (respectively *triangulation*) is admissible (see Figure 3.4 for examples).

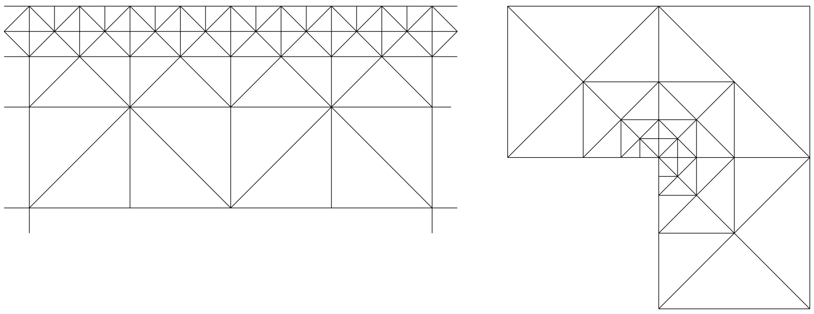


Figure 3.4: Admissible triangulations

The basic finite dimensional space V_h we investigate in the following is defined as

$$V_h := \left\{u \in C(\bar{\Omega}) \,:\, u|_{\partial\Omega} = 0, \quad u|_{T_i} \text{ is linear for all } T_i \in \mathcal{T}_h\right\}. \tag{3.39}$$

Lemma 3.34. *V_h defined in (3.39) is a linear subspace of $H_0^1(\Omega)$. Every function $v \in V_h$ is uniquely characterized by its values at the interior nodes (i.e. the corners of $T_i \in \mathcal{T}_h$ in the interior of Ω). The dimension of V_h is equal to the number of interior nodes.*

Proof: Since, by definition, each $v \in V_h$ is continuous and vanishes on $\partial\Omega$, it suffices to check $v \in H^1(\Omega)$. Obviously, $v \in \mathbb{L}^2(\Omega)$ since v is continuous and Ω is bounded. To show that also $\frac{\partial v}{\partial x_i}$ are in $\mathbb{L}^2$, we first have to calculate the generalized derivatives. Taking any test function $\phi \in C_0^\infty(\Omega)$, we have

$$\begin{aligned}\left\langle v, \frac{\partial \phi}{\partial x_i}\right\rangle = \int_\Omega v\frac{\partial \phi}{\partial x_i}\,dx &= \sum_{T_j\in\mathcal{T}_h}\int_{T_j} v\frac{\partial \phi}{\partial x_i}\,dx \\ &= \sum_{T_j\in\mathcal{T}_h}\left(-\int_{T_j}\frac{\partial v}{\partial x_i}\phi\,dx + \int_{\partial T_j} v\phi n_i\,d\sigma\right).\end{aligned}$$

Since ϕ vanishes on $\partial\Omega$, edges of triangles which coincide with parts of the boundary of Ω do not have to be considered. All other integrals along edges of the triangles appear twice in the sum with opposite orientation of the outer normal n but the same values of ϕ and v (due to continuity). Taking the cancellation into account, we find

$$\left\langle v, \frac{\partial \phi}{\partial x_i}\right\rangle = -\sum_{T_j\in\mathcal{T}_h}\int_{T_j}\frac{\partial v}{\partial x_i}\phi\,dx$$

so that the weak derivative of v is, by definition, the piecewise constant function

$$\frac{\partial v}{\partial x_i}(x) = \sum_{T_j\in\mathcal{T}_h}\mathcal{X}_{T_j}(x)\frac{\partial}{\partial x_i}\left(v|_{T_j}\right)$$

which is certainly square integrable on Ω.

Since a linear function on a triangle $T_i \in \mathcal{T}_h$ is uniquely determined by its values at the corners, it follows from $\bar{\Omega} = \bigcup_{i=1}^M T_i$ that $v \in V_h$ is determined by its values at the interior nodes $P_1, \dots, P_N$.

To show that $\dim V_h = N$, we construct basis functions $\psi_i \in V_h$ by the condition

$$\psi_i(P_j) = \delta_{ij} \qquad i, j = 1, \dots, N. \tag{3.40}$$

Certainly, on each triangle we can find a linear function satisfying (3.40) at its corners. To show that the resulting function on Ω is continuous, we need that $\mathcal{T}_h$ is an admissible triangulation. This assumption guarantees that each edge of $T_i \in \mathcal{T}_h$ is either part of the boundary or an edge of a neighboring triangle. On a common edge of two neighboring triangles, however, both linear functions are uniquely determined by the linear combination of the two values at the endpoints of the edge which gives continuity. With $\{\psi_i : i = 1, \dots, N\}$ being a well defined subset of V_h we can now check that it is a basis. Given any $v \in V_h$, the superposition

$$x \mapsto \sum_{i=1}^{N} v(P_i)\psi_i(x)$$

is obviously contained in V_h and coincides with v on all nodes of the partition. Due to the uniqueness result, we thus have

$$v = \sum_{i=1}^{N} v(P_i)\psi_i.$$

The required linear independence of the functions ψ_i follows from the uniqueness of the trivial representation of the zero function. ■

The basis functions $\psi_i \in V_h$ constructed in the proof of Lemma 3.34 are characterized by the condition (3.40). A basis consisting of functions which satisfy such a condition is called *nodal basis*. We also remark that the mapping

$$u \mapsto I_h u := \sum_{i=1}^{N} u(P_i)\psi_i \qquad u \in C(\bar{\Omega})$$

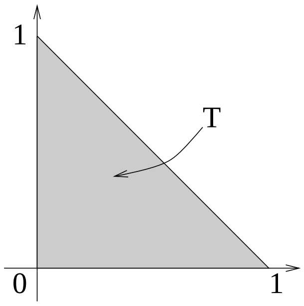
describes a linear interpolation of the nodal values $u(P_i)$ of u.

If we have a family of admissible partitions $\mathcal{T}_h$, it is clear that the dimension of V_h increases for $h \to 0$. In order to apply the Céa Lemma, however, it is necessary to show that V_h comes arbitrarily close to any function $u \in H_0^1(\Omega)$. The idea how to prove that is essentially the following: First, we use the fact that any $u \in H_0^1(\Omega)$ can be approximated by a more regular function $\tilde{u} \in S \cap H_0^1(\Omega)$ (where S can be, for example, $C_0^\infty(\Omega), C^k(\bar{\Omega})$ or $H^2(\Omega)$). For the smooth function $\tilde{u}$ we then show that the distance to the linear interpolation $I_h\tilde{u}$ becomes arbitrarily small in the H^1–norm for $h \to 0$. Altogether, this implies that any $u \in H_0^1(\Omega)$ can be approximated by elements of V_h to any accuracy if h is small enough.

To carry out this program, we choose $S = H^2(\Omega)$. The reason for this choice is that we can get more precise convergence estimates of the Ritz–Galerkin method under the additional assumption that u is H^2–regular (i.e. that $u \in$

$H^2(\Omega) \cap H_0^1(\Omega)$ which is indeed the case if the domain and the data have some extra smoothness).

We start our considerations with the so called *reference triangle* T shown in the figure below.



The main tool to obtain estimates is the Bramble-Hilbert Lemma 3.12.

Lemma 3.35. *Let T be the reference triangle and $u \in H^2(T)$. Then*

$$\|u\|_{H^2(T)}^2 \leq C \left(|u(0)|^2 + |u(e_1)|^2 + |u(e_2)|^2 + |u|_{2,2,T}^2 \right).$$

Similarly, for $u \in H^2(hT)$ with $h > 0$, we find

$$\|\nabla^\beta u\|_{\mathbb{L}^2(hT)}^2 \leq C h^{2-2|\beta|} \left(|u(0)|^2 + |u(he_1)|^2 + |u(he_2)|^2 + h^2 |u|_{2,2,hT}^2 \right).$$

Proof: According to Theorem C.2, $H^2(T)$ is embedded in the set of bounded continuous functions on T. In particular, for any $y \in T$, the evaluation functional $\delta_y(u) = u(y)$ is well defined and we have

$$|u(y)| = |\delta_y(u)| \leq \|u\|_{C_B(T)} \leq C \|u\|_{H^2(T)}.$$

Consequently,

$$\mu_0(u) := |u(0)|, \qquad \mu_1(u) := |u(e_1)|, \qquad \mu_2(u) := |u(e_2)|$$

are continuous semi-norms on $H^2(T)$. Moreover, any linear polynomial Q on T which satisfies $\mu_i(Q) = 0$ for all $i = 0, 1, 2$ is necessarily identically zero. Thus we can apply the Bramble-Hilbert Lemma which yields the estimate

$$\|u\|_{H^2(T)}^2 \leq C \left(|u|_{2,2,T}^2 + \sum_{i=0}^{2} \mu_i^2(u) \right)$$

and thus gives the first part of the Lemma.

The second estimate on the scaled triangle hT is reduced to the previous one by the transformation $v(y) = u(x)$ with $y = hx$ and $dy = h^2\, dx$.

$$\int_{hT} |\nabla_x^\beta u|^2 \, dx = \int_T |\nabla_x^\beta u|^2 h^2 \, dy = \int_T h^{-2|\beta|} |\nabla_y^\beta v|^2 h^2 \, dy. \tag{3.41}$$

Using the result for the reference triangle we get

$$\int_{hT} |\nabla_x^\beta u|^2 \, dx \leq C h^{2-2|\beta|} \left(|v(0)|^2 + |v(e_1)|^2 + |v(e_2)|^2 + |v|_{2,2,T}^2 \right).$$

Finally, by using (3.41) with $|\beta| = 2$, we can translate the right hand side back into terms involving u

$$\int_{hT} |\nabla_x^\beta u|^2 \, dx \le C h^{2-2|\beta|} \left(|u(0)|^2 + |u(he_1)|^2 + |u(he_2)|^2 + h^2 |u|^2_{2,2,hT} \right).$$

■

In order to extend the result to an estimate which applies uniformly to all triangles T_i of the partition $\mathcal{T}_h$, we consider T_i as an affine transformation $\Phi(h_i T)$ of a scaled reference triangle. An estimate as in Lemma 3.35 then follows from uniform bounds on all the involved transformations.

Setting $\Phi(x) = Bx + a$ for $x \in T$, the origin is mapped to a and the edges e_1 and e_2 turn into $\eta = Be_1$ and $\xi = Be_2$.

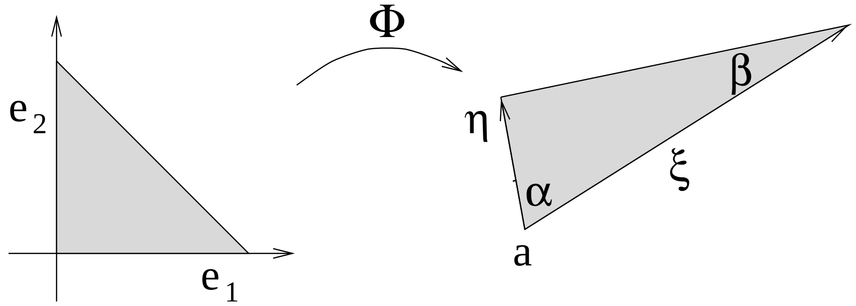


Figure 3.5: Affine linear deformation of the reference triangle.

Since the behavior under scaling is already understood, we just have to investigate distortions. Without loss of generality, we assume $|\eta| \le |\xi| = 1$. Then, all entries of the matrix

$$B = \begin{pmatrix} \eta_1 & \xi_1 \\ \eta_2 & \xi_2 \end{pmatrix}$$

are less than one and the norm of B just depends on the chosen matrix norm (in the matrix norm associated to the Euclidean norm, we have $\|B\| \le 2$). To estimate the norm of the inverse we use the representation

$$B^{-1} = \frac{1}{\det B} \begin{pmatrix} \xi_2 & -\xi_1 \\ -\eta_2 & \eta_1 \end{pmatrix}.$$

so that $\|B^{-1}\| \le 2/|\det B|$. The determinant of B is given by

$$\det B = \eta_1 \xi_2 - \eta_2 \xi_1$$

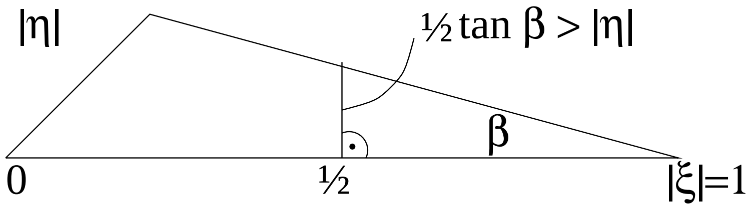
and can be related to geometric properties of the image triangle due to the relation

$$(\eta_1 \xi_2 - \eta_2 \xi_1)^2 = |\eta|^2 |\xi|^2 - |\eta \cdot \xi|^2 = |\eta|^2 |\xi|^2 (1 - \cos^2 \alpha) = |\eta|^2 |\xi|^2 \sin^2 \alpha$$

which implies

$$|\det B| = |\eta| \sin \alpha.$$

(det B describes the volume change under the mapping Φ and thus is twice the area of the image triangle.) Note that $\det B$ is small either if α or the ratio of the two edges $|\eta|/|\xi| = |\eta|$ is close to zero. To get a simple description of these extreme cases, we relate a possible bad ratio of the edges to a condition on the interior angle β (see Figure 3.5). In the case $|\eta| < 1/2$ we can estimate $|eta$ from below by $\frac{1}{2}\tan\beta$ as depicted in the figure below.



Hence, $|\eta| \geq \frac{1}{2}\min\{1, \tan\beta\}$ so that

$$|\det B| \geq \frac{1}{2}\min\{1, \tan\beta\}\sin\alpha$$

and

$$\|B^{-1}\| \leq \frac{4}{\min\{1, \tan\beta\}\sin\alpha}.$$

Consequently, if α_h denotes the smallest interior angle of all triangles in $\mathcal{T}_h$, we have a uniform estimate

$$\det B, \|B\|, (\det B)^{-1}, \|B^{-1}\| \leq C(\alpha_h).$$

Lemma 3.36. *Let $T_i \in \mathcal{T}_h$ be a triangle with corners $P_0^{(i)}, P_1^{(i)}, P_2^{(i)}$. Then, for any $u \in H^2(T_i)$ and $|\beta| \leq 2$ we have*

$$\|\nabla^\beta u\|^2_{\mathbb{L}^2(T_i)} \leq \tilde{C}(\alpha_h) h^{2-2|\beta|} \left(\sum_{j=0}^{2} |u(P_j^{(i)})|^2 + h^2 |u|^2_{2,2,T_i} \right)$$

where $\tilde{C}(\alpha_h)$ just depends on the smallest interior angle α_h of all triangles in $\mathcal{T}_h$.

Proof: Let $\Phi(x) = Bx + a$ be the affine linear transformation introduced above. Then T_i is the image of the scaled reference triangle $h_i T$ under Φ where $h_i \leq h$. According to Lemma 3.35 we have for $v \in H^2(h_i T)$

$$\|\nabla^\beta v\|^2_{\mathbb{L}^2(h_i T)} \leq C h^{2-2|\beta|} \left(|v(0)|^2 + |v(he_1)|^2 + |v(he_2)|^2 + h^2 |v|^2_{2,2,h_i T} \right).$$

The effect of the additional distortion due to Φ leads to additional constants which just depend on α_h (due to our control on B). We demonstrate the estimate for the case $|\beta| = 2$.

With $v(y) = u(\Phi(y))$ for $y \in h_i T$, we obtain a function $v \in H^2(h_i T)$ to which we can apply the above estimate. Hence, the main ingredients of the proof are transformation of variables in integrals and the chain rule.

$$\int_{T_i} \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 dx = \int_{h_i T} \left| \frac{\partial^2 u}{\partial x_i \partial x_j}(\Phi(y)) \right|^2 |\det B|\, dy.$$

Since

$$\frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial^2}{\partial x_i \partial x_j} v \circ \Phi^{-1} = \sum_{k,l=1}^{n} B_{ki}^{-1} B_{lj}^{-1} \frac{\partial^2 u}{\partial y_k \partial y_l}$$

Cauchy Schwarz inequality yields

$$\left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 \leq \sum_{k,l=1}^{n} \left(B_{ki}^{-1} B_{lj}^{-1} \right)^2 \sum_{k,l=1}^{n} \left| \frac{\partial^2 u}{\partial y_k \partial y_l} \right|^2 .$$

Finally, since $|B_{ki}^{-1}| \leq \|B^{-1}\| \leq C(\alpha_h)$ and $|\det B| \leq C(\alpha_h)$ we get with our result for the scaled triangle $h_i T$

$$\int_{T_i} \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 dx \leq \hat{C}(\alpha_h) h^{2-2|\beta|} \left(|v(0)|^2 + |v(he_1)|^2 + |v(he_2)|^2 + h^2 |v|^2_{2,2,h_i T} \right) .$$

To obtain the required estimate, $|v|^2_{2,2,h_i T}$ has to be transformed back to the distorted triangle

$$\sum_{k,l=1}^{n} \int_{h_i T} \left| \frac{\partial^2 u}{\partial y_k \partial y_l} \right|^2 dy$$
$$= \sum_{k,l=1}^{n} \frac{1}{\det B} \int_{T_i} \left| \sum_{i,j=1}^{n} B_{ik} B_{jl} \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 dx \leq \frac{\|B\|^2}{\det B} |u|^2_{2,2,T_i}$$

Since $|\det B|^{-1}, \|B\| \leq C(\alpha_h)$, the proof is complete. ∎

Now we are in the position to estimate the quality of linear interpolation of H^2–functions.

Theorem 3.37. *Let $\mathcal{T}_h$ be an admissible triangulation with interior nodes $P_1, \ldots, P_N$. and let $\{\psi_1, \ldots, \psi_N\}$ be the nodal basis of V_h. For $u \in H^2(\Omega) \cap H_0^1(\Omega)$ we set*

$$I_h u := \sum_{i=1}^{N} u(P_i) \psi_i$$

and obtain for $m = 0, 1$

$$\|u - I_h u\|_{H^m(\Omega)} \leq \tilde{C}(\alpha_h) h^{2-m} \|u\|_{H^2(\Omega)}$$

where α_h is the smallest interior angle of all triangles in $\mathcal{T}_h$.

Proof: For abbreviation, we set $w = u - I_h u$. Note that w vanishes on all nodes of the partition. We have, using Lemma 3.36, for $|\beta| = m < 2$

$$\|w\|^2_{H^m(\Omega)} = \sum_{T_i \in \mathcal{T}_h} \|w\|^2_{H^m(T_i)} \leq \sum_{T_i \in \mathcal{T}_h} \tilde{C}(\alpha_h) h^{4-2m} |w|^2_{2,2,T_i} .$$

Since all second order derivatives of $I_h u$ vanish on T_i, we find $|w|^2_{2,2,T_i} = |u|^2_{2,2,T_i}$ so that the result follows. ∎

If we have a family of triangulations $\{\mathcal{T}_h : h > 0\}$, the result of Theorem 3.37 applies simultaneously to all $\mathcal{T}_h$ if we can assure that the minimal interior angles are uniformly bounded away from zero. A family of triangulations satisfying this condition is called *quasi-uniform*. Thus, if V_h is constructed from a quasi-uniform family of triangulations we can apply the Céa Lemma and obtain convergence of the Ritz-Galerkin method.

Theorem 3.38. *Let $\{\mathcal{T}_h : h > 0\}$ be a quasi-uniform family of admissible triangulations of Ω with associated spaces $V_h \subset H_0^1(\Omega)$ consisting of continuous piecewise linear functions. Then, for any $u \in H_0^1(\Omega)$ we have*

$$\inf_{v\in V_h} \|u - v\|_{H^1(\Omega)} = d(u, V_h) \xrightarrow[h\to 0]{} 0.$$

If, moreover, $u \in H^2(\Omega) \cap H_0^1(\Omega)$ we get

$$d(u, V_h) \leq Ch\|u\|_{H^2(\Omega)}$$

respectively

$$\inf_{v\in V_h} \|u - v\|_{\mathbb{L}^2(\Omega)} \leq Ch^2\|u\|_{H^2(\Omega)}.$$

Proof: Theorem 3.37 shows the last two estimates. For the first one, density of $H^2(\Omega) \cap H_0^1(\Omega)$ in $H_0^1(\Omega)$ together with Theorem 3.37 is used. ■

A situation as described in the last theorem is typical in the theory of finite elements: the interpolation is of higher order, if the function, which is to be approximated, possesses more smoothness. In view of the Ritz-Galerkin method, it is therefore of interest, to classify abstract variational problems according to possible extra smoothness of solutions.

Definition 3.39. *Let a be a V–elliptic bilinear form where $H_0^m(\Omega) \subset V \subset H^m(\Omega)$ and $m \geq 1$. The abstract variational problem to find $u \in V$ such that*

$$a(u, v) = \langle f, v\rangle_0, \quad \forall v \in V$$

is called H^s–regular ($s \geq 2m$), if the unique solution $u \in V$ is also in $H^s(\Omega)$ such that

$$\|u\|_s \leq c\|f\|_{s-2m}$$

for all $f \in H^{s-2m}(\Omega)$

For second order problems with homogeneous Dirichlet conditions, one can find regularity theorems of the following kind.

Theorem 3.40. *Let a be a H_0^1–elliptic bilinear form with smooth coefficients*

(i) *If Ω is convex, then the Dirichlet problem in H^2-regular.*
(ii) *Let $s \geq 2$. If Ω has a C^s–boundary, then the Dirichlet problem is H^s-regular.*

Proof: See [**9**]. ■

Up to now, we have assumed that Ω has a polygonal boundary such that admissible triangulations can be constructed. This assumption, is not compatible with $\partial\Omega$ being very smooth. Using the part on convex sets, however, we derive

Theorem 3.41. *Let $\{\mathcal{T}_h : h > 0\}$ be a quasi–uniform family of admissible triangulations of the convex set Ω. Moreover, let a be a $H_0^1(\Omega)$–elliptic bilinear form with smooth coefficients and u as the unique solution of the corresponding abstract variational problem $a(u,v) = \langle f, v\rangle_0$ for all $v \in H_0^1(\Omega)$. Then, the Ritz–Galerkin approximation $u_h \in V_h$ satisfies*

$$\|u - u_h\|_1 \leq ch\|f\|_0$$

Proof: Since the domain Ω is convex, we have H^2–regularity and the estimate $\|u\|_2 \leq c_1\|f\|_0$ holds. Due to Theorem 3.38 we have $d(u, V_h) \leq Ch\|u\|_{H^2}$ so that the Céa Lemma yields the desired result. ■

While Theorem 3.41 yields $\mathcal{O}(h)$ convergence in the H^1-norm, one can expect that a higher order of convergence is obtained if we measure the distance between u and u_h in a weaker norm. Looking at the result of Theorem 3.38 we see that the approximation of H^2-functions by linear elements is indeed one order better in the weaker $\mathbb{L}^2$-norm. However, this result cannot be used directly to get a higher order of convergence because the Céa Lemma only holds for the H^1-distance. One can nevertheless show that second order convergence is obtained in the $\mathbb{L}^2$-norm if the *adjoint problem* is H^2-regular.

Definition 3.42. *Let a be a bilinear form on a Hilbert space V. Then,*

$$a^*(u,v) := a(v,u) \qquad \forall u, v \in V$$

is called the adjoint bilinear form to a. The corresponding abstract variational problem to find $u \in V$ such that

$$a^*(u,v) = l(v) \qquad \forall v \in V$$

is called adjoint problem.

Theorem 3.43. *Under the conditions of Theorem 3.41 we have*

$$\|u - u_h\|_0 \leq ch^2\|f\|_0$$

Proof: Since $a(u,u) = a^*(u,u)$, H_0^1-ellipticity of a implies H_0^1-ellipticity of a^*. Of course, if a has smooth coefficients, the same holds for a^* so that Theorem 3.40 shows that the adjoint problem is H^2-regular. Hence, for any $g \in \mathbb{L}^2(\Omega)$, the solution w of the adjoint problem with linear form $l(v) = \langle g, v\rangle_0$ satisfies $\|w\|_2 \leq C\|g\|_0$. According to Theorem 3.38 we then find

$$\inf_{v_h \in V_h} \|w - v_h\|_1 \leq Ch\|g\|_0. \tag{3.42}$$

(Note that we use C as abbreviation for any constant, i.e. the value of C in different estimates is not necessarily the same.) If u is the solution of

$$a(u,v) = \langle f, v\rangle_0 \qquad \forall v \in H_0^1(\Omega) \tag{3.43}$$

and u_h is the corresponding Ritz-Galerkin approximation, i.e. $a(u_h, v) = \langle f, v\rangle_0$ for all $v \in V_h$, we have, as in the Céa Lemma

$$a(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h. \tag{3.44}$$

The main idea of the proof rests on the relation

$$\|u - u_h\|_0 = \sup_{\|g\|_0=1} |\langle g, u - u_h\rangle_0|.$$

Using the definition of the adjoint problem together with (3.44), we get

$$\begin{aligned} |\langle g, u - u_h\rangle_0| = |a^*(w, u - u_h)| &= |a(u - u_h, w)| \\ &= |a(u - u_h, w - v_h)| \le C\|u - u_h\|_1 \|w - v_h\|_1 \end{aligned}$$

Since v_h is an arbitrary element of V_h we can use (3.42) to estimate

$$|\langle g, u - u_h\rangle_0| \le Ch\|u - u_h\|_1 \|g\|_0$$

so that

$$\|u - u_h\|_0 \le Ch\|u - u_h\|_1.$$

Finally, we apply Theorem 3.41 which yields

$$\|u - u_h\|_1 \le Ch\|f\|_0$$

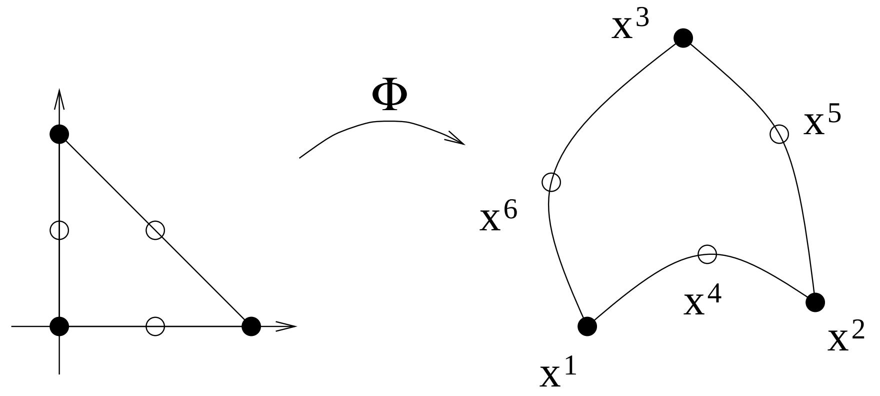
and the proof is complete. ■

3.8.3. Final remarks.

3.8.3.1. *Natural boundary conditions.* Replacing the Dirichlet condition, for example, by natural boundary conditions leads to some obvious modifications in the example of the last section. The nodal basis now includes also piecewise linear functions which are equal to one on boundary nodes. As a consequence, the dimension of the finite element space based on a partition $\mathcal{T}_h$ is larger in comparison with the corresponding space for Dirichlet conditions.

3.8.3.2. *Non-polygonal domains.* If Ω is not a polygonal domain, one can still apply the finite element method. A general approach is to use *isoparametric* finite elements. The idea is to replace the affine linear transformation Φ in Figure 3.5 by a more general, quadratic mapping

$$\Phi(x_1, x_2) = \begin{pmatrix} a_1 + a_2 x_1 + a_3 x_2 + a_4 x_1^2 + a_5 x_1 x_2 + a_6 x_2^2 \\ a_7 + a_8 x_1 + a_9 x_2 + a_{10} x_1^2 + a_{11} x_1 x_2 + a_{12} x_2^2 \end{pmatrix}, \qquad \forall (x_1, x_2) \in T$$



The coefficients $a_1, \ldots, a_{12}$ are uniquely determined by the images of corners and edge centers $x^1, \ldots x^6$ of the reference triangle. The shape function u on the isoparametric element is defined as the composition of the standard polynomial shape function v on the reference triangle and the mapping Φ^{-1} (i.e. $u = v \circ \Phi^{-1}$). A typical application of isoparametric elements is shown in Figure 3.6. The point x^4 is located on $\partial\Omega$ so that only the corresponding edge is not a linear segment.

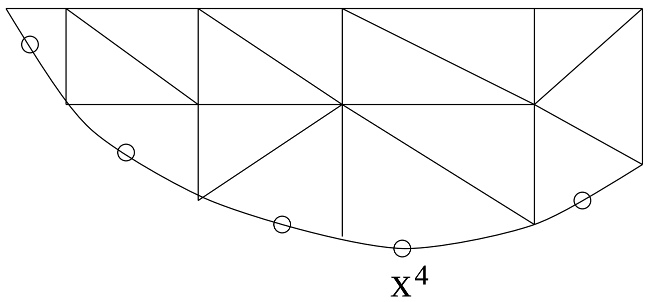


Figure 3.6: Application of isoparametric elements

3.8.3.3. *Non-conforming Finite Elements.* The condition $V_h \subset V$ in the Ritz-Galerkin method characterizes the *conforming Finite Element method.* Discretizations which are based on $V_h \not\subset V$ are called *non-conforming.* This situation appears if the shape functions on the elements are not connected in a sufficiently regular way (e.g. piecewise constant functions are not in H^1 because jumps lead to weak derivatives which are not square integrable). Nevertheless, the approach is appealing especially for higher order problems where the construction of finite dimensional subspaces of sufficiently smooth functions is difficult. If one decides to use non-conforming elements, an immediate problem arises in setting up the stiffness matrix because the original approach to evaluate the bilinear form at the basis functions can no longer be applied. Hence, an additional approximation $a \to a_h$ is required beside the discretization $V \to V_h$. In the case where V_h is based on a partition $\mathcal{T}_h$ and smooth (say polynomial) shape functions which are not connected regularly enough, one can set for example

$$a_h(u,v) := \sum_{T_i \in \mathcal{T}_h} \int_{T_i} \sum_{0 \le |\sigma|,|\gamma| \le m} a_{\sigma\gamma} \nabla^\gamma u \nabla^\sigma v \, dx.$$

However, the quality of the approximation $a \to a_h$ now also has to be investigated in order to ensure consistency of the discretization with the original problem.

3.8.3.4. *Other elements.* A generalization of linear triangular elements is obtained if we allow higher order polynomials as shape functions, i.e.

$$\mathcal{P}_d := \left\{ x \mapsto \sum_{|\alpha| \le d} c_\alpha x^\alpha \ : \ c_\alpha \in \mathbb{R} \right\}$$

with $d > 1$. Since $\mathcal{P}_d$ is invariant under affine transformations, it suffices to consider shape functions for the reference triangle. The case of general triangles can always be reduced to the reference situation.

Lemma 3.44. *Let $d \ge 0$. In the reference triangle T we define nodes on $d+1$ parallel lines as shown in Figure 3.7*

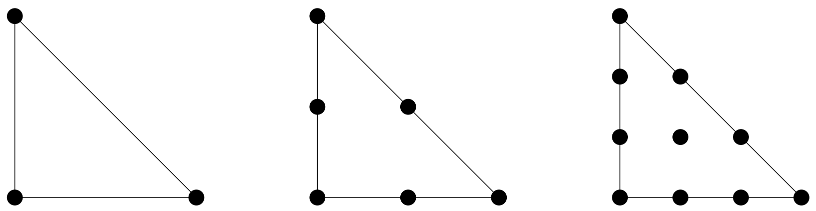


Figure 3.7: Nodes for linear, quadratic and cubic triangular elements

In this way we obtain $s = 1 + 2 + \cdots + (d+1)$ points which we call $z_1, \ldots, z_s$. To any $u \in C(T)$ there exists exactly one polynomial $Q \in \mathcal{P}_d$ which solves the interpolation problem

$$Q(z_i) = u(z_i) \qquad \forall i = 1, \ldots, s.$$

Proof: The proof is done by an induction argument. For $d = 0$, the claim is trivially true so that we can assume it holds for $d-1$. Our first observation is that for the $d+1$ points on the x-axis, we can find a unique polynomial Q_0 which takes the required values $u(z_i)$ in $z_1, \ldots, z_{d+1}$ (Lagrange interpolation). Moreover, due to induction assumption, there exists a polynomial Q_1 of degree $d-1$ which satisfies

$$Q(z_i) = \frac{1}{y_i}(u(z_i) - Q_0(z_i)) \qquad \forall i = d+2, \ldots, s$$

since the remaining nodes $z_{d+2}, \ldots z_s$ are located also on a triangle (which can be obtained by a simple scaling form the reference triangle). Obviously, $Q(x,y) = Q_0(x,y) + yQ_1(x,y)$ solves the interpolation problem. ∎

The concept of nodal bases can now easily be taken over.

Definition 3.45. (nodal basis)
Let V_h be a Finite Element space such that the functions in V_h are uniquely determined by their values at a finite number of points. Those functions which are equal to one at exactly one of these points form a basis of V_h. The basis is called nodal basis.

Similar to (3.39) we can now define C^0-elements based on an admissible triangulation $\mathcal{T}_h$ and shape functions from $\mathcal{P}_d$ with $d \geq 1$, i.e.

$$V_h := \left\{ u \in C(\bar{\Omega}) \, : \, u|_{\partial\Omega} = 0, \quad u|_{T_i} \in \mathcal{P}_d \; \forall T_i \in \mathcal{T}_h \right\}.$$

In order to construct a nodal basis of V_h, the $s = (d+1)(d+2)/2$ points in the reference triangle are mapped to each triangle $T_i \in \mathcal{T}_h$ under a suitable affine transformation. In particular, each edge of T_i carries exactly $d+1$ equally spaced points. In a first step, the prescription of zero and one as point values for a nodal basis function gives rise to uniquely defined polynomials on all triangles (due to Lemma 3.44). Considering two neighboring triangles, it remains to show that the corresponding polynomials match continuously at a common edge. Since the restrictions of both polynomials to the edge are again polynomials of degree d (in edge-coordinates) which coincide on $d+1$ separate points they must be identical. Consequently, the nodal basis function is automatically continuous and thus contained in V_h. Similar to Theorem 3.37 we have the result

Theorem 3.46. *Let $d \geq 2$ and $\mathcal{T}_h$ a quasi-uniform triangulation of Ω. Then we have for the interpolation using piecewise polynomials of degree $d-1$*

$$\|u - I_h u\|_{m,h} \leq c \cdot h^{d-m} |u|_{t,\Omega}, \quad u \in H^d(\Omega), \quad 0 \leq m \leq d$$

where the norm $\|\cdot\|_{m,h}$ is defined by

$$\|v\|_{m,h} = \sqrt{\sum_{T_j \in \mathcal{T}_h} \|v\|^2_{m,T_j}}.$$

Proof: The proof is similar to the one of Theorem 3.37. We refer to [**2**] for more details. ∎

If a Finite Element space in H^2 is required (e.g. for the Ritz-Galerkin method applied to a fourth order problem), overall continuity is not enough. Instead, one can show as in Lemma 3.34 that the functions have to be at least in $C^1(\bar{\Omega})$. The construction of so called C^1-elements is much more complicated than that of C^0-elements. As an example, we consider the construction of triangular C^1–elements using polynomials of degree five as shape functions, which was given by Argyris: for polynomials of degree five, we have to determine $(d+1)(d+2)/2|_{d=5} = 21$ degrees of freedom. First, we prescribe for each triangle the values of the polynomial at the three corners up to the second order derivatives, which fixes 18 degrees of freedom. For the *Argyris element*, we determine the resulting 3 degrees of freedom by the values of the normal derivatives at the midpoints of the three edges (see Fig. 3.8).

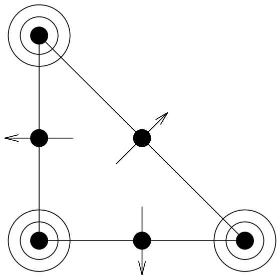


Figure 3.8: Argyris Element (C^1–element)

To see that the *Argyris element* yields a globally C^1–element, we consider the restriction of the polynomials on the edges of the triangles, which are one–dimensional polynomials of degree 5, i.e. we have 6 degrees of freedom. Moreover, the polynomials are determined by the values at the two nodes up to the second derivative. Hence, we have continuity and continuity of the tangential derivatives along the edges. The normal derivative is a polynomial of degree 4, i.e. we have five degrees of freedom. These degrees of freedom are fixed by the values up to the first order derivatives at the nodes and the value at the midpoints of the edges. The normal derivative is therefore uniquely determined and we have continuity of the normal derivative.

Note, that there is a variety of other C^1-elements (see [**2, 4**]).

Finally, let us consider a rectangular C^0-element.

Figure 3.9: Rectangle to define C^0–elements

While in triangular elements the shape functions are polynomials from the full class $\mathcal{P}_d$ for some $d \geq 1$ (so called *complete polynomials*), the shape functions for rectangular elements typically have a tensorial structure, i.e. they are contained in

$$\mathcal{Q}_d = \left\{ (x,y) \mapsto \sum_{0 \leq i,j \leq d} c_{ij} x^i y^j \ : \ c_{ij} \in \mathbb{R} \right\}.$$

In particular, the case $d = 1$ leads to bilinear shape functions

$$u(x,y) = a + bx + cy + dxy.$$

If the corners of all rectangles of a partition $\mathcal{T}_h$ are taken as nodes, the global continuity of the nodal basis functions follows from the fact that the restriction of a bilinear polynomial to an edge of a rectangle is linear and thus uniquely determined by the values at the endpoints of the edge. Again, one can find estimates for the interpolation property of bilinear rectangular elements similar to Theorem 3.37 (see [**2**]).

3.9. Exercises

Exercise 3.1. On $\Omega \subset \mathbb{R}^n$ with $n > 1$ let $L(x, \nabla) = \sum_{|\alpha| \le k} a_\alpha(x) \nabla^\alpha$ be a general linear differential operator.

a) Show that L can only be elliptic if k is even.
b) Show that, if $k = 2m$ and a_α are sufficiently smooth, L can be written in divergence form $L(x, \nabla)u = \sum_{0 \le |\sigma|, |\gamma| \le m} \nabla^\sigma (a_{\sigma\gamma}(x) \nabla^\gamma u)$.

Exercise 3.2. Let $L(x, \nabla) = \sum_{|\alpha| \le m} a_\alpha(x) \nabla^\alpha$ be defined on Ω with C^m coefficients. The formal adjoint L^* of L is defined by $\langle L^*\phi, \psi\rangle := \langle \phi, L\psi\rangle$ for all $\phi, \psi \in \mathscr{D}(\Omega)$.

a) Show that L^* is a linear differential operator.
b) Show that L^*L is in divergence form.
c) Use b) to show that the biharmonic operator $\Delta\Delta$ is in divergence form.

Exercise 3.3. Let $\Omega \subset \mathbb{R}^n$ be a bounded C^1 domain, $f \in C(\Omega)$ and

$$L(x, \nabla)u = \sum_{0 \le |\sigma|, |\gamma| \le m} (-1)^{|\sigma|} \nabla^\sigma (a_{\sigma\gamma}(x) \nabla^\gamma u)$$

with smooth coefficients. Show that a weak solution $u \in H_0^m(\Omega)$ of the Dirichlet problem

$$\begin{aligned} L(x, \nabla)u &= f \qquad && \text{in } \Omega \\ \nabla^\alpha u &= 0 && \text{on } \partial\Omega,\ |\alpha| \le m-1 \end{aligned} \tag{3.45}$$

which satisfies $u \in C^{2m}(\Omega) \cap C^m(\bar{\Omega})$ is also a classical solution (i.e. satisfies (3.45) point-wise).

Exercise 3.4. Let L be a second order, uniformly elliptic differential operator in divergence form with constant coefficients. Show that the corresponding bilinear form is $H_0^1(\Omega)$–elliptic if a_{00} (the coefficient of the zeroth order term in L) is positive, respectively $a_{00} \ge 0$ if Ω is bounded.

Exercise 3.5. Consider the bilinear form $a : H_0^1(0,1) \times H_0^1(0,1) \mapsto \mathbb{R}$ defined by

$$a(u, v) = \int_0^1 x^2 u'(x) v'(x)\, dx.$$

- What is the corresponding (ordinary) differential operator L?
- Show that L is elliptic but not uniformly elliptic.
- Show that a is not $H_0^1(0,1)$–elliptic.

- Show that the functional

$$\Theta(u) = a(u,u)/2 - \int_0^1 u\,dx$$

has no minimum on $H_0^1(0,1)$.

Exercise 3.6. Let $D = \{x \in \mathbb{R}^2 : |x| < 1\}$ be the unit disk in $\mathbb{R}^2$ and $\Omega := D\backslash\{0\}$ the disk without the origin. Consider the Dirichlet problem $\Delta u = 1$ in Ω and $u = 0$ on $\partial\Omega$.

a) Show that the Dirichlet problem has no classical solution $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$.
b) Show that the Dirichlet problem has a unique weak solution $u \in H_0^1(\Omega)$.
c) What is the weak solution? (Hint: Using the fact that $\delta(u) := u(0)$ defined on $C_0^\infty(D)$ cannot be extended to a continuous functional on $H_0^1(\Omega)$ show that $H_0^1(\Omega) = H_0^1(D)$.)
d) Show that the assumption $|\delta(u)| \leq C\|u\|_{H^1(D)}$ for all all $C_0^\infty(D)$ functions u (i.e. that δ can be extended to H_0^1) leads to the wrong result that $H_0^1(\mathbb{R}^2)$ is continuously embedded in the space of bounded continuous functions on $\mathbb{R}^2$ (this explains the hint in (c) a little further).

Exercise 3.7. Consider the boundary value problem on $\Omega = (0,1)$

$$\begin{aligned} u'' + p(x)u' + q(x)u &= f(x) \qquad &&\text{in } \Omega \\ u &= 0 &&\text{on } \partial\Omega \end{aligned}$$

where $p \in C^1[0,1]$ and $q \in C[0,1]$. Prove that a unique solution exists if $p' \geq 2q$.

Exercise 3.8. Let $\Omega \subset \mathbb{R}^n$ be a bounded connected domain with C^1–boundary. Let $\emptyset \neq \Gamma_D \subset \partial\Omega$ be open and define

$$H^1_{\Gamma_D} := \{u \in H^1(\Omega) : Bu = 0 \quad \text{on } \Gamma_D\}$$

where $B : H^1(\Omega) \mapsto \mathbb{L}^2(\partial\Omega)$ is the trace operator.

- Derive the classical boundary value problem which corresponds to finding $u \in H^1_{\Gamma_D}(\Omega)$ such that

$$\langle \nabla u, \nabla v\rangle_0 = \langle f, v\rangle_0 \qquad \forall v \in H^1_{\Gamma_D}(\Omega)$$

- Show that $a(u,v) = \langle \nabla u, \nabla v\rangle_0$ is $H^1_{\Gamma_D}(\Omega)$ elliptic. (Hint: use Bramble Hilbert Lemma with a suitable semi norm.)

APPENDIX A

Numerical linear algebra

In this part we want to give a quick review on solution methods for systems of linear equations. Of course, our attention is focussed on systems which arise from the discretization of linear elliptic PDEs. The existing algorithms for the solution of

$$(A.1)\qquad Ax = b, \qquad A \in \mathbb{R}^{m\times m}, \quad x, b \in \mathbb{R}^m, \quad m \in \mathbb{N}$$

mainly fall into two categories: *direct methods* which compute the exact answer to (A.1) after a finite number of steps (if we neglect roundoff errors) and *iterative methods* which do not produce the exact answer after a finite number of steps (in general) but decrease the error by some fraction in each iteration.

A.1. Direct methods

A typical example in this category is the *Gaussian elimination*. The main idea of the algorithm is to split the matrix A into the product of a lower triangular matrix L and an upper triangular matrix U (therefore the algorithm is also called LU factorization). After this splitting has been accomplished, the original problem (A.1) is easily solved in two steps: first, $L^{-1}b$ is calculated which is simple since L is a lower triangular matrix. Then U^{-1} is applied to the result of step one which is again simple due to the triangular structure. The numerical cost for the final steps is $\mathcal{O}(m^2)$ and thus is dominated by the cost for the splitting which is $\mathcal{O}(m^3)$. In general, the algorithm can be applied to A only after a suitable reorganization of rows and/or columns (so called *pivoting*). This extra effort is not necessary if A is symmetric and positive definite (spd). In this special case, Gaussian elimination is essentially equivalent to *Cholesky decomposition* where A factors into LL^T with a lower triangular matrix L. In standard form, both Gaussian elimination and Cholesky decomposition do not take care of a possible sparse structure of A. (However, there are modifications, so called *direct sparse solvers*, which try to do that. See [**7**] for references.) If (A.1) originates in the discretization of a PDE, we easily get $m > 10^6$ with only very few non zero elements per line (depending on the size of the stencil). Hence, there are several arguments why direct solvers are not the methods of choice:

- standard solvers need memory of order m^2,
- the computation time is in general of order m^3,
- we get an exact solution to a wrong (or rather discretized) problem.

A.2. Iterative methods

With iterative methods, we compute the first members of a sequence $(x_k)_{k\in\mathbb{N}}$ of approximations to the solution of (A.1). Formally, we write

$$x_k = \Phi^{(k)}(x_{k-1}, \dots, x_0)$$

and require

a) $(x_k)_{k\in\mathbb{N}}$ should converge quickly,
b) $\Phi^{(k)}$ should be simple to evaluate (i.e. not more expensive than matrix vector multiplication which are of the order m for sparse matrices).

The basic idea of classical iterative methods is to transform the original problem (A.1) into a fixed point problem. One way to do this is

$$Ax = b \qquad \Longleftrightarrow \qquad x = (I - A)x + b = \Phi(x). \tag{A.2}$$

Then, the fixed point iteration $x_{k+1} = \Phi(x_k)$ converges under certain conditions on the matrix A (if $I - A$ is a contraction). Certainly, requirement b) is satisfied by this approach. As far as property a) is concerned, we have the following fundamental

Theorem A.1. *The fixed point iteration $x_{k+1} = Gx_k + c$ converges for every initial vector x_0 if and only if $\rho(G) < 1$, where $\rho(G)$ is the spectral radius of G, i.e. $\rho(g) = \max\{|\lambda_i| : \lambda_i$ eigenvalue of $G\}$.*

Proof: (for symmetric G)
There are rotation matrices $R \in \mathbb{R}^{m\times m}$ such that

$$RGR^T = \Lambda, \qquad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_m), \qquad |\lambda_i| < 1.$$

The sequence $y_k = Rx_k$ then satisfies

$$y_{k+1} = \Lambda y_k + d, \qquad d = Rc$$

or equivalently,

$$y_{k+1} = \Lambda^{k+1} y_0 + \sum_{i=0}^{k} \Lambda^i d.$$

Since $|\lambda_i| < 1$ we thus get

$$y_k \xrightarrow[k\to\infty]{} \operatorname{diag}\left(\frac{1}{1-\lambda_1}, \dots, \frac{1}{1-\lambda_m}\right) d = (I - \Lambda)^{-1} Rc.$$

multiplying by R^T from the right finally yields

$$x_k \xrightarrow[k\to\infty]{} (I - R^T\Lambda R)^{-1} c = (I - G)^{-1} c.$$

■

To check that $\rho(G) < 1$, it suffices to have $\|G\| < 1$ in some matrix norm $\|\cdot\|$ which is related to a vector norm. In such a norm it is then easy to see that

$$\|x_k - x\| \le \|G\|^k \|x_0 - x\|.$$

Thus, the smaller $\|G\|$, the faster the iterative scheme converges which gives some result concerning requirement a).

Applying these considerations to (A.2) (which leads to the so called *Richardson iteration*) we see that the scheme works only with special matrices A. If, for example, A is spd, the spectrum of A is located on the positive real line in $\mathbb{C}$. Consequently, $G = I - A$ has its spectrum to the left of 1 so that $\rho(G) < 1$ is only satisfied if $\rho(A) < 2$. This is of course a severe restriction (just think of the matrix $\bar{\mathcal{L}}_h$ obtained from the 1D Dirichlet Poisson problem in Section 2.7.1 where the maximal eigenvalue grows like $\mathcal{O}(m^2)$ (see Exercise 2.11)). Hence, (A.2) is not very helpful in our case and we have to think of a different fixed point formulation. The basic idea is to modify the original problem (A.1) by multiplying with the inverse of a matrix Q (so called *preconditioning*)

$$Ax = b \qquad \Longleftrightarrow \qquad Q^{-1}Ax = Q^{-1}b.$$

With the same idea as in (A.2) this leads to the fixed point formulation

$$Ax = b \qquad \Longleftrightarrow \qquad x = \underbrace{(I - Q^{-1}A)}_{G}\, x + \underbrace{Q^{-1}b}_{c}. \tag{A.3}$$

Compared to $Q = I$, which is the original formulation (A.2), the case $Q = A$ is at the other extreme. Indeed, if $Q = A$ then $G = 0$ and the first step of the iterative scheme yields the solution. However, to set up the scheme, A^{-1} has to be known already so that requirement b) is certainly violated: the evaluation of Φ is as hard as solving the problem itself. As so often, the truth lies somewhere in the middle which means in this case that we have to choose $Q \approx A$, so that $\rho(G) = \rho(I - Q^{-1}A) \approx 0$ but with a simple rule how to get Q^{-1}, or more precisely, how to solve (the simplified) problems of the form $Qx = b$. In other words, preconditioning amounts to having already a good guess about the solution of the problem. In the following, we list some ideas how to derive preconditioners (taken from [**16**]):

- *Diagonal scaling or Jacobi.* A very important preconditioner is given by $Q = \operatorname{diag}(A)$ provided this matrix is invertible. Similarly, one can take any invertible diagonal matrix but the entries should be related to the problem to get an improvement of the convergence behavior.
- *Incomplete Cholesky or LU factorization.* The idea is to perform similar transformations on A as in Cholesky or, for the non symmetric case, LU factorization, but to keep only those entries in the triangular factors which have nonzero counterparts in the matrix A.
- *Coarse–grid approximation.* While the first two examples work independent of the underlying problem, the coarse–grid approximation is designed for matrices which result from the discretization of a partial differential or integral equation. If the discretization which leads to A is very fine, the analogous discretization on a coarser grid may lead to a small system that is comparably easy to solve. Thus, a powerful

preconditioner may be obtained of the following schematic form:

$$Q = \langle \text{transfer to fine grid} \rangle \circ A_{\text{coarse}} \circ \langle \text{transfer to coarse grid} \rangle \,.$$

Applying this approach to a hierarchy of grids is one of the ideas in *multigrid schemes.*

- *Local approximation.* In a coarse–grid approximation, large scale structure of the problem is taken into account while ignoring some of the finer structure. If A contains coupling between near as well as far elements it can also be reasonable to consider the operator Q analogous to A but with the longer–range interactions omitted. In the simples case of this kind, Q may consist simply of a few of the diagonals of A near the main diagonal.
- *Block preconditioners and domain decomposition.* If a matrix has a block structure (say block tridiagonal), the Jacobi preconditioner can be generalized to block–diagonal or block–Jacobi form. This is another kind of local approximation, in that local effects within certain components are considered while connections to other components are ignored. Ideas of that kind have lately been generalized in the field of domain decomposition, in which solvers for certain subdomains of a problem are composed in flexible ways to form preconditioners for the global problem.
- *Low–order discretization.* If a differential or integral method is discretized by a higher order method one typically finds a gain in accuracy but stencils become bigger and thus A is less sparse. A lower approximation of the same problem, with its sparser matrix, may be an effective preconditioner.
- *Constant coefficient or symmetric approximation.* For a problem with variable coefficients, a constant–coefficient approximation implemented by a fast solver may be a good preconditioner. Analogously, if a differential equation is not self–adjoint but is close in some sense to a self–adjoint equation that can be solved more easily, then the latter may sometimes serve as preconditioner.
- *Splitting of a multi–term operator.* Many applications involve combinations of physical effects which can manifest itself in the discretization as a splitting of A into natural contributions $A = A_1 + \cdots + A_m$. If a partial sum of the matrices A_i is easily invertible, it may serve as a good preconditioner.
- *Dimensional splitting or ADI.* Another kind of splitting takes advantage of the fact that an operator such as the Laplacian in two or three dimensions is composed of analogous operators in each of the directions separately. This idea may form the basis for a preconditioner, and in one form goes by the name *ADI* or *alternating direction implicit* methods.
- *One step of a classical iterative method.* This is another idea which is used in multigrid schemes. Some of the classical iterative methods are presented below.

We now briefly describe some classical iterative methods. A detailed investigation can be found in [**3**]. For the case of a discretized Poisson problem on a square, the methods are presented in [**6**]. In the *Richardson iteration*, no preconditioning is performed. If A is spd, then the eigenvalues of $G = I - A$ are located left of $\lambda = 1$. The requirement on the spectral radius of G given in Theorem A.1, leads to the restriction $|1 - \rho(A)| < 1$. In other words, the Richardson iteration converges for spd matrices A provided $\rho(A) < 2$. For the next two methods, we split the matrix A into a lower triangle L and an upper triangle U (both with zeros in the diagonal) and the diagonal part $D = \operatorname{diag}(A)$ of A

$$A = L + D + U.$$

If we take D as preconditioner, we find

$$G = I - D^{-1}A = I - D^{-1}(L + D + U) = -D^{-1}(L + U)$$

giving rise to the *Jacobi iteration* respectively *single step method*. Under the assumption of strictly diagonal dominance,

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|, \qquad i = 1, \ldots, m$$

the Jacobi iteration converges. (Note that for strictly diagonal dominant matrices A we have $\|G\|_\infty = \|D^{-1}(L + U)\|_\infty < 1$.) Finally, if we take as preconditioner the lower triangle $Q = L + D$ (which is easy to invert due to the triangular structure) we obtain the so called *Gauß–Seidel* or *multi step method*. The matrix G is now of the form

$$G = I - (D + L)^{-1}(L + D + U) = -(D + L)^{-1}U$$

It can be shown that the Gauß–Seidel method converges for any symmetric and positive definite matrix A.

A general approach to construct a one–parameter family of iteration methods based on a given method is *relaxation*. Given an iteration method described by G and c, we take as new value x_{k+1} a linear combination of the old step x_k and original correction $Gx_k + c$

$$x_{k+1} = \omega(Gx_k + c) + (1 - \omega)x_k = \underbrace{(\omega G + (1 - \omega)I)}_{G_\omega} x_k + \omega c.$$

The advantage of this approach is that we can now determine the optimal scheme in the family by picking ω in such a way that $\rho(G_\omega)$ becomes minimal. If the corresponding ω is less than one, we speak of *under–relaxation*, if it is larger than one the notion is *over–relaxation*. Typically, the original scheme can be improved by relaxation, i.e. the optimal parameter ω will not be one. For an spd matrix A, the optimal relaxation of the Richardson iteration is obtained with

$$\omega = \frac{2}{\lambda_{\max}(A) + \lambda_{\min}(A)}$$

which yields the spectral radius

$$\rho(G_\omega) = \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} < 1.$$

Consequently, the relaxed scheme converges also in cases $\lambda_{\max}(A) \geq 2$ in which the original Richardson iteration fails.

For the special case $A = \mathcal{L}_h$, where A results from the discretization of an elliptic operator $\mathcal{L}$, we can also use positivity properties to show the convergence of iterative methods. As in Exercise 2.12, we introduce the notion *M–matrix* for matrices with a sign structure like $\mathcal{L}_h$.

Definition A.2. *$A \in \mathbb{R}^{m\times m}$ is called M–matrix, if $A_{ij} \leq 0$ for $i \neq j$, if A is regular and if A^{-1} is positive.*

It can be shown (see [**3**]) that both Jacobi and Gauß–Seidel iterations and their relaxed versions with $\omega \in (0,1)$ converge for any start vector. However, under– relaxation is not useful, because convergence is slowed down.

Generally the most efficient linear, first order iteration scheme is SOR (which expands to *successive over–relaxation*). It is nothing else than Gauß–Seidel plus relaxation (for $\omega < 1$, the method is still called successive over–relaxation although $\omega < 1$ is actually under–relaxation). For general spd matrices A, we have the result that SOR converges for any $\omega \in (0,2)$.

APPENDIX B

Generalized functions

B.1. Generalizing derivatives

In the theory of generalized functions it is possible to extend the domain of definition of differential operators from classically differentiable functions to a larger class in a natural way. Of course, in the theory of differential equations, this is of especial importance. To give an example, let us consider the function $f(x) = |x|$. What is a natural candidate for $f'(x)$? Clearly, for negative x, f is classically differentiable and $f'(x) = -1$. Similarly, $f'(x) = 1$ for positive x. At this point, we already see that any natural derivative of f cannot be continuous at $x = 0$ since its value should be one to the right and minus one to the left. Therefore, the object f' must be an element of some larger function space then $C^0(\mathbb{R})$. Due to the behavior at $x = 0$, it is convenient to look for $f' \in \mathbb{L}^1_{loc}(\mathbb{R})$ (that means f' is an equivalence class of functions which are integrable over any compact subset of $\mathbb{R}$ and which differ from each other at most on a set of Lebesgue measure zero). Indeed, the difference quotients

$$\left.\frac{f(x+h) - f(x)}{h}\right|_{x=0} = \frac{|h|}{h} = \operatorname{sign} h$$

oscillate between plus and minus one depending on the sign of h. Consequently, there is no natural value for the derivative of f at the single point $x = 0$, but in $\mathbb{L}^1_{loc}(\mathbb{R})$, the value of a single point has no relevance. Hence, we can select for example

$$g(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

and define $f' \in \mathbb{L}^1_{loc}(\mathbb{R})$ as the equivalence class of functions which contains g. A nice property of this generalized derivative is that the integration by parts formula is still correct. If we take a smooth function $\phi \in C^\infty(\mathbb{R})$ which is compactly supported (so that $f'\phi$ is integrable), we find

$$\int_{\mathbb{R}} f'\phi \, dx = -\int_{-\infty}^{0} \phi \, dx + \int_{0}^{\infty} \phi \, dx$$

and after applying integration by parts to both integrals on the right

$$\int_{\mathbb{R}} f'\phi \, dx = \int_{-\infty}^{0} x\phi' \, dx - \int_{0}^{\infty} x\phi' \, dx = -\int_{\mathbb{R}} |x|\phi' \, dx = -\int_{\mathbb{R}} f\phi' \, dx.$$

In the same way, a natural, generalized derivative of any continuous, piecewise differentiable function can be defined as an element of $\mathbb{L}^1_{loc}(\mathbb{R})$ and the integration by parts formula remains valid.

In the next step, one can ask for the derivative of a function with a jump (say g defined above). Again, we find classically $g' = 0$ for $x < 0$ and $g' = 0$ for $x > 0$. In $x = 0$, the difference quotient can now attain any positive number. The same approach as before would then imply $g' = 0$ in $\mathbb{L}^1_{loc}$ (i.e. g' is the equivalence class containing the zero function). Then, however, any information about the singular behavior of the difference quotient at $x = 0$ is lost. For the same reason, the integration by parts formula does not hold with this definition since

$$\int_{\mathbb{R}} g\phi' \, dx = -\int_{-\infty}^{0} \phi' \, dx + \int_{0}^{\infty} \phi' \, dx = -2\phi(0)$$

is in general different from

$$\int_{\mathbb{R}} 0 \cdot \phi \, dx = 0.$$

Hence, if we require that for the generalized derivative the integration by parts formula should hold, we have to find a function $h \in \mathbb{L}^1_{loc}(\mathbb{R})$ (our candidate for g') such that

$$\int_{\mathbb{R}} h\phi \, dx = -\int_{\mathbb{R}} g\phi' \, dx = 2\phi(0)$$

for any $\phi \in C_0^\infty(\mathbb{R})$. It is easy to see that no such function exists in $\mathbb{L}^1_{loc}(\mathbb{R})$ (because it must be zero almost everywhere in $\mathbb{R}\backslash\{0\}$). Consequently, one either has to give up the integration by parts formula, or the class of functions has to be more general than $\mathbb{L}^1_{loc}$. The latter implies, on the other hand, that also the notion of the integral has to be generalized such that it makes sense for non–$\mathbb{L}^1_{loc}$ functions.

In the theory of generalized functions, the integral

$$\int_{\mathbb{R}} g\phi \, dx$$

with $g \in \mathbb{L}^1_{loc}(\mathbb{R})$ is considered as a linear functional

$$\phi \mapsto \int_{\mathbb{R}} g\phi \, dx = I_g(\phi), \qquad \phi \in C_0^\infty(\mathbb{R})$$

acting on smooth functions ϕ. Since $g \in \mathbb{L}^1_{loc}(\mathbb{R})$ is uniquely determined if we know $I_g(\phi)$ for a sufficiently large class of functions ϕ, this interpretation is very natural (since no information about g is lost when going over to I_g). In our previous example, the derivative of f (or in the new interpretation I_f) was naturally given by g (resp. I_g) and the integration by parts formula

$$I_g(\phi) = -I_f(\phi'), \qquad \forall \phi \in C_0^\infty(\mathbb{R}) \tag{B.1}$$

was valid. Therefore, it was natural, to call $g = f'$ (resp. $I_g = I'_f$) and (B.1) can be written as

$$I'_f(\phi) = -I_f(\phi'), \qquad \forall \phi \in C_0^\infty(\mathbb{R}).$$

In the second example, we have

$$I_g(\phi') = -2\phi(0), \qquad \forall \phi \in C_0^\infty(\mathbb{R})$$

which is still a linear functional acting on ϕ (given essentially by $\phi \mapsto \Lambda(\phi) = 2\phi(0)$), but it can no longer be written in the form I_h with $h \in \mathbb{L}^1_{loc}(\mathbb{R})$. However, the integration by parts formula remains valid, if we set $I'_g = \Lambda$ since then

$$I'_g(\phi) = \Lambda(\phi) = 2\phi(0) = -I_g(\phi'), \qquad \forall \phi \in C_0^\infty(\mathbb{R}).$$

Consequently, by identifying functions $F \in \mathbb{L}^1_{loc}(\mathbb{R})$ with the functionals I_F acting on so called *test functions* $\phi \in C_0^\infty(\mathbb{R})$, we can define derivatives I'_F such that the integration by parts formula holds by setting

$$I'_F(\phi) := -I_F(\phi'), \qquad \forall \phi \in C_0^\infty(\mathbb{R})$$

(the left hand side being clearly a well defined linear functional on the test functions). The derivative I'_F really generalizes the classical derivative. Indeed, if $F \in C^1(\mathbb{R})$ (which is a subspace of $\mathbb{L}^1_{loc}(\mathbb{R})$), we get

$$I'_F(\phi) = -I_F(\phi') = -\int_{\mathbb{R}} F\phi'\,dx = -\left. F\phi\right|_{-\infty}^{\infty} + \int_{\mathbb{R}} F'\phi\,dx = I_{F'}(\phi)$$

i.e. $I'_F = I_{F'}$. Since I_F and $I_{F'}$ can be identified with the functions F and F', we find that the generalized derivative of F is nothing but the classical derivative F'.

It turns out that the identification of $F \in \mathbb{L}^1_{loc}(\mathbb{R})$ with the linear functional I_F is extremely powerful because it can be used to generalize a lot of concepts known for classical functions to a much wider class.

B.2. Aspects from the theory of generalized functions

Here we will briefly introduce some of the concepts which are important in the treatment of fundamental solutions and Green's functions. A more detailed investigation is given in [**11, 14, 15**]. We start with a more precise definition of the test function space and the generalized functions. In the following, Ω denotes an open set in $\mathbb{R}^n$.

Definition B.1. *A function ϕ defined on Ω is called a test function if $\phi \in C^\infty(\Omega)$ and there is a compact set $K \subset \Omega$ such that the support of ϕ lies in K. The set of all test functions on Ω is denoted by $\mathscr{D}(\Omega) = C_0^\infty(\Omega)$.*

Obviously, $\mathscr{D}(\Omega)$ is a linear space. To do analysis, we need a notion of convergence. It is possible to define a topology on $\mathscr{D}(\Omega)$ but for most purposes in PDEs this is not necessary; only a definition for the convergence of sequences is required.

Definition B.2. *Let ϕ_k, $k \in \mathbb{N}$ and ϕ be elements of $\mathscr{D}(\Omega)$. We say that ϕ_k converges to ϕ in $\mathscr{D}(\Omega)$, if there is a compact subset K of Ω such that the supports of all the ϕ_k (and of ϕ) lie in K and, moreover, ϕ_k and derivatives of ϕ_k of arbitrary order converge uniformly to those of ϕ.*

As motivated in the previous section, it is possible to view classical functions as linear functionals on the space of test functions which gives rise to

Definition B.3. *A distribution or generalized function is a linear mapping $\phi \mapsto \Lambda(\phi)$ from $\mathscr{D}(\Omega)$ to $\mathbb{R}$, which is continuous in the following sense: If $\phi_k \to \phi$ in $\mathscr{D}(\Omega)$, then $\Lambda(\phi_k) \to \Lambda(\phi)$. The set of all distributions is called $\mathscr{D}'(\Omega)$. A sequence Λ_k in $\mathscr{D}'(\Omega)$ converges to $\Lambda \in \mathscr{D}'(\Omega)$ if $\Lambda_k(\phi) \to \Lambda(\phi)$ for every $\phi \in \mathscr{D}(\Omega)$.*

It is customary, to write $\langle \Lambda, \phi \rangle$ instead of $\Lambda(\phi)$. Also, $\langle I_f, \phi \rangle$ is denoted $\langle f, \phi \rangle$ due to the possible identification of f and I_f for $f \in \mathbb{L}^1_{loc}(\Omega)$. If the test function ϕ depends on an additional parameter $\phi(x,y)$, it is sometimes not clear without further information on which variable the dual product $\langle \cdot, \cdot \rangle$ acts. Therefore we also write $\langle \Lambda(x), \phi(x,y) \rangle_x$ to indicate that Λ acts on x and not on y. Note, however, that Λ is not an ordinary function and that $\Lambda(x)$ therefore does not refer to the value of Λ at a point x (which is meaningless). It just clarifies how Λ is applied to test functions.

Let us now give a few examples how properties of classical functions can be extended to generalized functions by duality.

Definition B.4. *Let $\Lambda \in \mathscr{D}'(\Omega)$. Then, the derivative of Λ with respect to x_j is defined as*

$$\left\langle \frac{\partial \Lambda}{\partial x_j}, \phi \right\rangle := -\left\langle \Lambda, \frac{\partial \phi}{\partial x_j} \right\rangle, \qquad \forall \phi \in \mathscr{D}(\Omega).$$

Higher derivatives are obtained recursively. For a given multi–index α, we have

$$\langle \nabla^\alpha \Lambda, \phi \rangle = (-1)^{|\alpha|} \langle \Lambda, \nabla^\alpha \phi \rangle .$$

Motivated by the behavior of the integral under a linear transformation of variables $U(x) = Ax + b$ we get

Definition B.5. *Let $U(x) = Ax + b$ be a non singular linear transformation in $\mathbb{R}^n$, and let $U^{-1}(y) = A^{-1}(y - b)$ be the inverse transformation. Then, we set for any $\Lambda \in \mathscr{D}'(\mathbb{R}^n)$*

$$\langle U\Lambda, \phi \rangle := |\det A| \, \langle \Lambda, \phi \circ U \rangle , \qquad \forall \phi \in \mathscr{D}(\mathbb{R}^n).$$

Another way of denoting the linear transformation is

$$\left\langle \Lambda(U^{-1}(y)), \phi(y) \right\rangle_y = |\det A| \, \langle \Lambda(x), \phi(Ax + b) \rangle_x .$$

As example, let us mention transformations of Dirac's delta distribution $\delta \in \mathscr{D}'(\mathbb{R}^n)$ defined by

$$\langle \delta, \phi \rangle := \phi(0).$$

Taking any vector $x_0 \in \mathbb{R}^n$, we get the translated delta distribution by

$$\langle \delta(y - x_0), \phi(y) \rangle = \langle \delta(x), \phi(x + x_0) \rangle = \phi(x_0)$$

and for some constant $a \neq 0$

$$\langle \delta(ay), \phi(y) \rangle = \frac{1}{|a|^n} \langle \delta(x), \phi(x/a) \rangle = \frac{\phi(0)}{|a|^n}.$$

The product of a distribution with a smooth function is also given by duality.

Definition B.6. *Let $\psi \in C^\infty(\Omega)$ and $\Lambda \in \mathscr{D}'(\Omega)$. Then $\psi\Lambda$ is defined by*

$$\langle \psi\Lambda, \phi \rangle := \langle \Lambda, \psi\phi \rangle, \qquad \forall \phi \in \mathscr{D}(\Omega).$$

Note that if $\phi \in \mathscr{D}(\Omega)$ then also $\psi\phi$ is a compactly supported C^∞ function on Ω so that the right hand side is well defined.

While products between smooth functions and distributions can be defined, a generalization of the point wise product of functions is not possible in a natural way. More precisely, if Λ_1 and Λ_2 are two distributions, we cannot assign a distribution $\Pi(\Lambda_1, \Lambda_2)$ in such a way that it amounts to the usual product for regular distributions (those which correspond to an $\mathbb{L}^1_{loc}$ function) and is at the same time continuous in both arguments. To see this, just consider the regular distribution $\Lambda_k(x) = \sin(kx)$ in $\mathscr{D}'(\mathbb{R})$. Then $\Pi(\Lambda_k, \Lambda_k) = \sin^2(kx)$, but although $\Lambda_k \to 0$ in $\mathscr{D}'$, we get $\Pi(\Lambda_k, \Lambda_k) \to \frac{1}{2}$. Indeed, for any test function ϕ, we have

$$\langle \Lambda_k, \phi \rangle = \int_{\mathbb{R}} \sin(kx)\phi(x)\,dx = -\frac{1}{k}\int_{\mathbb{R}} \cos(kx)\phi'(x)\,dx$$

so that

$$|\langle \Lambda_k, \phi \rangle| \xrightarrow[k\to\infty]{} 0 \qquad \forall \phi \in \mathscr{D}(\mathbb{R})$$

which means $\Lambda_k \to 0$ in $\mathscr{D}'(\mathbb{R})$. Similarly,

$$\begin{aligned}\langle \Pi(\Lambda_k, \Lambda_k), \phi \rangle &= \int_{\mathbb{R}} \sin^2(kx)\phi(x)\,dx = \frac{1}{2}\int_{\mathbb{R}} (1 - \cos(2kx))\phi(x)\,dx \\ &= \frac{1}{2}\int_{\mathbb{R}} \phi(x)\,dx + \frac{1}{4k}\int_{\mathbb{R}} \sin(2kx)\phi'(x)\,dx \xrightarrow[k\to\infty]{} \frac{1}{2}\int_{\mathbb{R}} \phi(x)\,dx = \left\langle \frac{1}{2}, \phi \right\rangle.\end{aligned}$$

Although the point wise product cannot be generalized, an extension of tensor products is easily possible. That is, if $\Lambda_p \in \mathscr{D}'(\mathbb{R}^p)$ and $\Lambda_q \in \mathscr{D}'(\mathbb{R}^q)$, then $\Lambda_p(x)\Lambda_q(y)$ can be defined as a distribution on $\mathscr{D}'(\mathbb{R}^{p+q})$.

Definition B.7. *Let $\Lambda_p \in \mathscr{D}'(\mathbb{R}^p)$ and $\Lambda_q \in \mathscr{D}'(\mathbb{R}^q)$. Then the direct product $\Lambda_p \otimes \Lambda_q(x, y) = \Lambda_p(x)\Lambda_q(y)$ is the distribution on $\mathbb{R}^{p+q}$ given by*

$$\langle \Lambda_p \otimes \Lambda_q, \phi \rangle = \left\langle \Lambda_p(x), \langle \Lambda_q(y), \phi(x, y) \rangle_y \right\rangle_x, \qquad \forall \phi \in \mathscr{D}(\mathbb{R}^{p+q}).$$

This definition needs some justifications: The function $\psi(x)$ which is obtained when applying the distribution Λ_q to $\phi(x, y)$ is compactly supported and C^∞ smooth. Thus, the right hand side in the above relation is well defined. It remains to be shown that $\Lambda_p \otimes \Lambda_q$ is continuous on $\mathscr{D}(\mathbb{R}^{p+q})$ (see for example [**14**]).

As a simple example of a direct product, we note that $\delta(x)\delta(y) = \delta(x,y)$. If $\phi(x,y)$ itself is of product form $\phi(x,y) = \phi_p(x)\phi_q(y)$, we obtain

$$\langle \Lambda_p \otimes \Lambda_q, \phi \rangle = \langle \Lambda_p, \phi_p \rangle \langle \Lambda_q, \phi_q \rangle .$$

In [**14**], it is shown that linear combinations of products $\phi_p(x)\phi_q(y)$ with $\phi_p \in \mathscr{D}(\mathbb{R}^p)$ and $\phi_q \in \mathscr{D}(\mathbb{R}^q)$ are dense in $\mathscr{D}(\mathbb{R}^{p+q})$. This fact can be used to show properties of the direct product by verifying them only on the restricted set of test functions of product form. One immediate conclusion is that in the definition we can evaluate Λ_p and Λ_q in the opposite order, i.e. we also have

$$\langle \Lambda_p \otimes \Lambda_q, \phi \rangle = \left\langle \Lambda_q(y), \langle \Lambda_p(x), \phi(x,y) \rangle_x \right\rangle_y$$

which we express by the suggestive notation

$$\Lambda_p(x)\Lambda_q(y) = \Lambda_q(y)\Lambda_p(x).$$

Another property is the associative law

$$\Lambda_p(x)(\Lambda_q(y)\Lambda_r(z)) = (\Lambda_p(x)\Lambda_q(y))\Lambda_r(z).$$

Finally, we turn to the convolution of two distributions. To motivate the definition, let us consider the case of two classical functions f, g which decay rapidly at infinity. Then

$$\begin{aligned} \langle f * g, \phi \rangle &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} f(x-y) g(y)\, dy \right) \phi(x)\, dx \\ &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} f(x) g(y) \phi(x+y)\, dx\, dy. \end{aligned}$$

While the product $f(x)g(y)$ can be generalized as direct product of two distributions, the problem is that $\phi(x+y)$ is not a test function out of $\mathscr{D}(\mathbb{R}^{2n})$ since it is not compactly supported (unless $\phi \equiv 0$). Indeed, it turns out that the convolution of arbitrary distributions cannot be defined in a rational manner. There are, however, special cases where a meaning can be given to $\langle \Lambda_1(x)\Lambda_2(y), \phi(x+y) \rangle_{(x,y)}$. In the simplest such case, the support of $\phi(x+y)$ has a compact intersection with the support of $f(x)g(y)$. (In this case, we may replace $\phi(x+y)$ by any test function which agrees with $\phi(x+y)$ in a neighborhood of supp $(\Lambda_1(x)\Lambda_2(y))$.) To make this idea more precise, we have to clarify the notion *support of a distribution.*

Although generalized functions cannot be evaluated at points, they can be restricted to open sets. If G is an open subset of Ω, then $\mathscr{D}(G)$ is naturally a subset of $\mathscr{D}(\Omega)$. Consequently, we shall say $\Lambda \in \mathscr{D}'(\Omega)$ vanishes on G if $\langle \Lambda, \phi \rangle = 0$ for all $\phi \in \mathscr{D}(G)$. Two distributions are said to be equal on G if their difference vanishes on G.

It can be shown (by a partition of unity argument) that if Λ vanishes locally around every point of G then Λ vanishes on G. Now, the complement of the union of all open subsets of Ω on which Λ vanishes is a natural candidate for the support of Λ. For example, the support of the delta distribution is the set $\{0\}$. Although the δ cannot be evaluated at points, it makes sense to say that it vanishes except at the origin.

Let Λ be compactly supported on a set $K \subset \Omega$. We then take a test function $\phi_K \in \mathscr{D}(\Omega)$ which is equal to one on a neighborhood of K and find for any $\phi \in \mathscr{D}(\Omega)$

$$\langle \Lambda, \phi \rangle = \langle \Lambda, \phi\phi_K + (1-\phi_K)\phi \rangle = \langle \phi_K \Lambda, \phi \rangle$$

since $(1-\phi_K)\phi \in \mathscr{D}(\Omega \backslash K)$. Thus, $\Lambda = \phi_K \Lambda$ and we can even define the dual product with non–compactly supported test functions $\eta \in C^\infty(\Omega)$ by setting

$$\langle \Lambda, \eta \rangle := \langle \Lambda, \phi_K \eta \rangle \, .$$

To show that this definition is independent of the choice of ϕ_K is simple. If $\tilde{\phi}_K$ is another function with the same properties as ϕ_K, then

$$\langle \Lambda, \phi_K \eta \rangle = \left\langle \tilde{\phi}_K \Lambda, \phi_K \eta \right\rangle = \left\langle \Lambda, \tilde{\phi}_K \phi_K \eta \right\rangle = \left\langle \phi_K \Lambda, \tilde{\phi}_K \eta \right\rangle = \left\langle \Lambda, \tilde{\phi}_K \eta \right\rangle .$$

Coming back to the definition of the convolution, we see that if , for example , Λ_1 is compactly supported, we can replace Λ_1 by $\phi_K \Lambda_1$ with a suitable function $\phi_K \in \mathscr{D}(\mathbb{R}^n)$. Then,

$$\langle \Lambda_1 * \Lambda_2, \phi \rangle := \langle \Lambda_1(x)\Lambda_2(y), \phi(x+y) \rangle = \langle \Lambda_1(x)\Lambda_2(y), \phi_K(x)\phi(x+y) \rangle$$

and now $(x,y) \mapsto \phi_K(x)\phi(x+y)$ is really contained in $\mathscr{D}(\mathbb{R}^{2n})$ so that the convolution is well defined.

As example, we consider the convolution of a distribution Λ with the δ distribution (which is compactly supported)

$$\langle \delta * \Lambda, \phi \rangle = \langle \delta(x)\Lambda(y), \phi(x+y) \rangle = \langle \Lambda(y), \langle \delta(x), \phi(x+y) \rangle_x \rangle_y = \langle \Lambda(y), \phi(y) \rangle_y$$

so that $\delta * \Lambda = \Lambda$. If we take the convolution of a distribution Λ with a test function ψ, we find

$$\begin{aligned} \langle \Lambda * \psi, \phi \rangle &= \left\langle \Lambda(x), \langle \psi(y), \phi(x+y) \rangle_y \right\rangle_x \\ &= \left\langle \Lambda(x), \langle \psi(y-x), \phi(y) \rangle_y \right\rangle_x = \langle \langle \Lambda(x), \psi(x-y) \rangle_x , \phi(y) \rangle_y \end{aligned}$$

so that $\Lambda * \psi$ is represented by the smooth function $y \mapsto \langle \Lambda(x), \psi(y-x) \rangle_x$. Finally, the derivative of a convolution is

$$\begin{aligned} \left\langle \frac{\partial}{\partial x_j}(\Lambda_1 * \Lambda_2), \phi \right\rangle &= - \left\langle \Lambda_1(x)\Lambda_2(y), \frac{\partial \phi}{\partial x_j}(x+y) \right\rangle \\ &= - \left\langle \Lambda_2(y), \left\langle \Lambda_1(x), \frac{\partial \phi}{\partial x_j}(x+y) \right\rangle_x \right\rangle_y \\ &= - \left\langle \Lambda_2(y), \left\langle \frac{\partial \Lambda_1}{\partial x_j}(x), \phi \right\rangle_x \right\rangle_y = \left\langle \frac{\partial \Lambda_1}{\partial x_j} * \Lambda_2, \phi \right\rangle \end{aligned}$$

Similarly, we can show

$$\frac{\partial}{\partial x_j}(\Lambda_1 * \Lambda_2) = \frac{\partial \Lambda_1}{\partial x_j} * \Lambda_2 = \Lambda_1 * \frac{\partial \Lambda_2}{\partial x_j}$$

so that by induction for any multi–index α

$$\nabla^\alpha(\Lambda_1 * \Lambda_2) = (\nabla^\alpha \Lambda_1) * \Lambda_2 = \Lambda_1 * (\nabla^\alpha \Lambda_2).$$

One can also show a result on the support of the convolution $\Lambda_1 * \Lambda_2$. It is contained in the sum of the supports of Λ_1 and Λ_2

$$\operatorname{supp}(\Lambda_1 * \Lambda_2) \subset \operatorname{supp} \Lambda_1 + \operatorname{supp} \Lambda_2.$$

For the (easy) proof we refer to [**15**].

Finally, we want to show that convolution can be used to prove that any $\Lambda \in \mathscr{D}'(\Omega)$ can be approximated by test functions in the sense of distributions. The idea is to choose cut–off functions $\chi_k \in \mathscr{D}(\Omega)$ which are equal to one on large subsets Ω_k of Ω and drop to zero only close to the boundary of Ω. For

$$\Omega_k := \{\, x \in \Omega \,:\, \operatorname{dist}(x, \partial\Omega) > \frac{1}{k}, \quad |x| < k\}$$

it is easy to see that $\chi_k \Lambda \to \Lambda$ in $\mathscr{D}'(\Omega)$ (since any fixed test function will eventually be supported in Ω_k). Since $\chi_k \Lambda \in \mathscr{D}'(\mathbb{R}^n)$ has compact support it suffices to consider only such distributions.

Choosing a mollifier $\phi \in \mathscr{D}(B_1(0))$ which is nonnegative and symmetric ($\phi(x) = \phi(-x)$) and satisfies $\langle 1, \phi \rangle = 1$, we set up

$$\phi_\epsilon(x) = \frac{1}{\epsilon^n} \phi\left(\frac{x}{\epsilon}\right).$$

As we have seen, $\Lambda * \phi_\epsilon \in \mathscr{D}(\mathbb{R}^n)$. Moreover, for any $\psi \in \mathscr{D}(\mathbb{R}^n)$

$$\begin{aligned}
\langle \Lambda * \phi_\epsilon, \psi \rangle &= \left\langle \langle \Lambda(y), \phi_\epsilon(x-y) \rangle_y , \psi(x) \right\rangle_x \\
&= \left\langle 1, \langle \Lambda(y), \phi_\epsilon(x-y)\psi(x) \rangle_y \right\rangle_x \\
&= \langle 1 \cdot \Lambda(y), \phi_\epsilon(x-y)\psi(x) \rangle_{(x,y)} \\
&= \langle \Lambda(y), \langle 1, \phi_\epsilon(x-y)\psi(x) \rangle_x \rangle_y \\
&= \langle \Lambda(y), \phi_\epsilon * \psi(y) \rangle_y
\end{aligned}$$

(Note that in each step we have used definitions and results from above.) For smooth functions we can classically show that $\phi_\epsilon * \psi \to \psi$ in $\mathscr{D}(\mathbb{R}^n)$ for $\epsilon \to 0$. Thus

$$\langle \Lambda * \phi_\epsilon, \psi \rangle \xrightarrow[\epsilon \to 0]{} \langle \Lambda, \psi \rangle$$

so that $\Lambda * \phi_\epsilon \to \Lambda$ in $\mathscr{D}'(\mathbb{R}^n)$. Summarizing the result, we get:

Lemma B.8. *Let Ω and Ω' be open subsets of $\mathbb{R}^n$ with $\bar{\Omega}' \subset \Omega$ and let $\phi \in \mathscr{D}(B_1(0))$ be a test function which satisfies $\phi \geq 0$ and $\langle 1, \phi \rangle = 1$. With $\phi_\epsilon(x) = \phi(x/\epsilon)/\epsilon^n$ we have*

$$\Lambda * \phi_\epsilon\big|_{\Omega'} \xrightarrow[\epsilon\to 0]{\mathscr{D}'(\Omega')} \Lambda\big|_{\Omega'} .$$

Note that $\Lambda * \phi_\epsilon\big|_{\Omega'} \in C^\infty(\bar{\Omega}')$ for ϵ small enough.

B.3. A simple PDE

Theorem B.9. *Let $\Omega \subset \mathbb{R}^n$ be open and connected and let $m \in \mathbb{N}$. If $u \in \mathscr{D}'(\Omega)$ satisfies $\nabla^\alpha u = 0$ for all $|\alpha| = m$ then u is a polynomial of degree less than m.*

We start with a Lemma for the case $m = 1$.

Lemma B.10. *Let $\Omega \subset \mathbb{R}^n$ be open and connected. If $u \in \mathscr{D}'(\Omega)$ satisfies $\nabla u = 0$ then u is a constant function on Ω.*

Proof: Let $B \subset \Omega$ be a ball with $\bar{B} \subset \Omega$. Using a mollifier ϕ_ϵ as in Lemma B.8, we find

$$u * \phi_\epsilon \in C^\infty(\bar{B}).$$

Using the fact that derivatives can be put to each partner in the convolution, we find for $x \in B$

$$\nabla(u * \phi_\epsilon)(x) = ((\nabla u) * \phi_\epsilon)(x) = 0$$

so that $u * \phi_\epsilon$ is constant on B, say equal to c_ϵ. If $\psi_0 \in \mathscr{D}(B)$ is a test function which is normalized in the sense $\langle 1, \psi_0 \rangle = 1$, we have

$$c_\epsilon = \langle u * \phi_\epsilon, \psi_0 \rangle \xrightarrow[\epsilon\to 0]{} \langle u, \psi_0 \rangle .$$

(Here, convergence from Lemma B.8 has been used.) For any other $\psi \in \mathscr{D}(B)$ we get

$$\langle u * \phi_\epsilon - \langle u, \psi_0 \rangle , \psi \rangle = (c_\epsilon - \langle u, \psi_0 \rangle) \langle 1, \psi \rangle \xrightarrow[\epsilon\to 0]{} 0$$

which implies

$$\langle u - \langle u, \psi_0 \rangle , \psi \rangle = 0 \tag{B.2}$$

or, in other words,

$$u\big|_B = \langle u, \psi_0 \rangle .$$

We remark that the constant $c_B = \langle u, \psi_0 \rangle$ is, in fact, independent of ψ_0 and thus depends only on u and B. In fact, if ψ_1 is another test function which is normalized by $\langle 1, \psi_1 \rangle = 1$ we get from (B.2) $\langle u, \psi_1 \rangle = \langle u, \psi_0 \rangle$.

It is easy to see that the constants c_{B_1} and c_{B_2} of two intersecting balls B_1 and B_2 must be equal by taking $\psi \in \mathscr{D}(B_1 \cap B_2)$ with $\langle 1, \psi \rangle = 1$ and observing

$$c_{B_1} = \langle c_{B_1}, \psi \rangle = \langle u, \psi \rangle = \langle c_{B_2}, \psi \rangle = c_{B_2}.$$

Hence, we can define a mapping $f : \Omega \mapsto \mathbb{R}$ which assigns to $x \in \Omega$ the value c_B where $x \in B$ and B is any ball whose closure is contained in Ω. The function f is even continuous. If we take a sequence $(x_k)_{k\in\mathbb{N}}$ which converges to $x \in \Omega$ then, eventually, the members of the sequence will be in a small ball B around x so that $f(x_n) = c_B = f(x)$ for n sufficiently large. Finally, we consider the set

$$A = \{\, x \in \Omega \;:\; f(x) = f(x_0)\,\}$$

where $x_0 \in \Omega$ is an arbitrary point. Since f is continuous on Ω, A is closed in Ω. On the other hand, for any $x \in A$ we can find a ball B around x on which f is equal to $f(x)$ which is equal to $f(x_0)$ by definition of A. Thus, A is also open and since $x_0 \in A \neq \emptyset$ the connectedness of Ω implies $A = \Omega$ which yields $f = \mathit{const}$. Using a locally finite partition of unity on Ω based on balls $B \subset \Omega$, we can thus infer that $u = f$. ■

The remainder of the proof of Theorem B.9 is just a simple induction argument.
Proof: (of Theorem B.9)
The case $m = 1$ is treated in the Lemma so we assume the Theorem is correct for $m - 1$. Let $u \in \mathscr{D}'(\Omega)$ with $\nabla^\alpha u = 0$ for all $|\alpha| = m$. Set $v_\beta = \nabla^\beta u$ where $|\beta| = m - 1$ so that $\nabla v_\beta = 0$ for each β which implies $v_\beta = c_\beta$ by the Lemma above. Since

$$\nabla^\beta \left(\frac{1}{\beta!} c_\beta x^\beta \right) = c_\beta$$

we set up

$$\hat{u} = u - \sum_{|\beta|=m-1} \frac{1}{\beta!} c_\beta x^\beta \tag{B.3}$$

so that $\nabla^\beta \hat{u} = 0$ for all $|\beta| = m - 1$. Hence, $\hat{u}$ is a polynomial of degree less than $m - 1$ which concludes the proof due to (B.3). ■

APPENDIX C

Sobolev spaces

The following introduction to Sobolev spaces is not meant to be exhaustive. We just want to present the basic concepts and focus somewhat on the problem how to generalize boundary values.

Definition C.1. *Let $\Omega \subset \mathbb{R}^n$ be open and let $m \in \mathbb{N}_0$, $p \in [1, \infty]$. The Sobolev space $W^{m,p}(\Omega)$ consists of those functions $f \in \mathbb{L}^p(\Omega)$ whose distributional derivatives up to order m are again $\mathbb{L}^p(\Omega)$ functions.*

Of course, $W^{m,p}(\Omega)$ can naturally be interpreted as a subset of $\mathscr{D}'(\Omega)$. However, we can choose a stronger topology which is based on the norm

$$\|f\|_{m,p,\Omega} = \left(\sum_{|\alpha| \leq m} \|\nabla^\alpha f\|_{0,p}^p \right)^{\frac{1}{p}}.$$

(Note that $W^{0,p}(\Omega) = \mathbb{L}^p(\Omega)$.) It turns out that the norm topology on $W^{m,p}(\Omega)$ is complete [**1**] so that the Sobolev spaces are Banach spaces. The case $p = 2$ is of special importance because $W^{m,2}(\Omega)$ are even Hilbert spaces with the scalar products

$$\langle u, v \rangle_m = \sum_{|\alpha| \leq m} \int_\Omega \nabla^\alpha u \nabla^\alpha v \, dx.$$

A standard abbreviation of $W^{m,2}(\Omega)$ is $H^m(\Omega)$.

C.1. Embedding theorem

An important property of Sobolev spaces are the embedding relations among them. For a complete list of these relations we refer to [**1**]. Here, we just mention a few:

Theorem C.2. *Let Ω be a domain in $\mathbb{R}^n$ with a Lipschitz boundary and let p satisfy $1 \leq p < \infty$. Under the condition $p < n$ we find*

$$W^{1,p}(\Omega) \hookrightarrow \mathbb{L}^q(\Omega), \qquad p \leq q \leq np/(n-p).$$

If $mp > n$ for some $m \in \mathbb{N}$, then

$$W^{m,p}(\Omega) \hookrightarrow C_B(\Omega)$$

where $C_B(\Omega)$ is the space of bounded continuous functions on Ω (endowed with the supremum norm). If Ω is bounded and m is a nonnegative integer then

$$W^{m+1,p}(\Omega) \hookrightarrow W^{m,p}(\Omega)$$

and this embedding is even compact *(i.e. every $\|\cdot\|_{m+1,p}$– bounded sequence has a subsequence which converges in $W^{m,p}(\Omega)$.*

Proof: See [**1**]. ∎

In order to illustrate the compact embedding, let us consider the special case $\Omega = (0,1)$, with $p = 2$ and $m = 1$. More precisely, we will show that

$$H^1(0,1) \hookrightarrow \mathbb{L}^2(0,1)$$

is compact. Our argument is based on the knowledge of orthonormal bases in $\mathbb{L}^2(0,1)$ (see Exercise C.4). With

$$\psi_n(x) = \sqrt{2}\sin(n\pi x) \qquad n \in \mathbb{N}$$

we can represent any $v \in \mathbb{L}^2(0,1)$ by the ($\mathbb{L}^2$–convergent) series

$$v = \sum_{n\in\mathbb{N}} \langle v, \psi_n \rangle_0 \, \psi_n$$

where $\langle\cdot,\cdot\rangle_0$ is the scalar product on $\mathbb{L}^2$. According to Parseval's equality, we have

$$\|v\|^2_{\mathbb{L}^2} = \sum_{n\in\mathbb{N}} |\, \langle v, \psi_n \rangle_0 \,|^2$$

so that the mapping $v \mapsto (\langle v, \psi_n \rangle_0)_{n\in\mathbb{N}}$ is an isometry from $\mathbb{L}^2(0,1)$ to the space of square summable sequences

$$l^2 = \{\, (c_n)_{n\in\mathbb{N}} \, : \, \sum_{n\in\mathbb{N}} |c_n|^2 < \infty \,\}.$$

which is a Hilbert space with the scalar product $\langle c, d \rangle_{l^2} = \sum_{n\in\mathbb{N}} c_n d_n$. Using the cosine basis

$$\phi_0 \equiv 1, \quad \phi_n(x) = \sqrt{2}\cos(n\pi x) \qquad n \in \mathbb{N}$$

we get with the same argument that

$$(\langle v, \phi_n \rangle_0)_{n\in\mathbb{N}} \in l^2 \qquad \forall v \in \mathbb{L}^2(0,1).$$

If we now take $u \in H^1(0,1) \subset \mathbb{L}^2(0,1)$, we observe

$$n\pi \,\langle u, \phi_n \rangle_0 = \langle u, \psi_n' \rangle_0 = -\,\langle u', \psi_n \rangle_0 + |u(x)\psi_n(x)|_0^1 \,.$$

Since ψ_n vanishes at the end points of the interval we conclude

$$(n\,\langle u, \phi_n \rangle_0)_{n\in\mathbb{N}} = -\frac{1}{\pi}(\langle u', \psi_n \rangle_0)_{n\in\mathbb{N}} \in l^2$$

Hence, for $u \in H^1(0,1)$ the coefficients of the cosine basis are square summable even if they are multiplied by the increasing sequence $(n)_{n\in\mathbb{N}}$. In other words, the high frequency parts of H^1 functions (i.e. the coefficients corresponding to highly oscillatory basis functions) are much smaller than those of $\mathbb{L}^2$ functions. Introducing the space

$$h^1 = \{\, c \in l^2 \; : \; \sum_{n\in\mathbb{N}} n^2 |c_n|^2 < \infty \,\}$$

we conclude that

$$u \in H^1(0,1) \qquad \Rightarrow \qquad (\langle u, \phi_n \rangle_0)_{n\in\mathbb{N}} \in h^1$$

(Note that h^1 is also a Hilbert space based on the scalar product $\langle c, d \rangle_{h^1} = \sum_{n\in\mathbb{N}} n^2 c_n d_n$.) The compact embedding of $H^1(0,1)$ in $\mathbb{L}^2(0,1)$ now easily follows from the compact embedding $h^1 \hookrightarrow l^2$. (The l^2 elements just have to be interpreted as coefficients for the cosine basis of $\mathbb{L}^2(0,1)$. More precisely, if $(u_m)_{m\in\mathbb{N}}$ is a uniformly bounded sequence in $H^1(0,1)$ then

$$\sum_{n\in\mathbb{N}} n^2 |\, \langle u_m, \phi_n \rangle_0 \,|^2 = \frac{1}{\pi^2} \sum_{n\in\mathbb{N}} |\, \langle u'_m, \psi_n \rangle_0 \,|^2 = \frac{1}{\pi^2} \|u'_m\|^2_{\mathbb{L}^2} \le \frac{1}{\pi^2} \|u_m\|^2_{H^1}$$

so that $(c_n^{(m)})_{n\in\mathbb{N}}(\langle u_m, \phi_n \rangle_0)_{n\in\mathbb{N}}$ gives rise to a uniformly bounded sequence in h^1. Due to compactness, we can extract an l^2 convergent subsequence which corresponds to a sequence in $\mathbb{L}^2(0,1)$ with the desired properties.) To show the compactness of $h^1 \hookrightarrow l^2$ we take a uniformly bounded sequence $(c^{(m)})_{m\in\mathbb{N}} \in h^1$, i.e. $\|c^{(m)}\|_{h^1} \le C$ for all $m \in \mathbb{N}$. This implies, that each component $c_n^{(m)}$ is a bounded sequence in $\mathbb{R}$. Using a diagonal process, we can therefore find a subsequence $(c^{(m_k)})_{k\in\mathbb{N}}$ which converges in each component n to some $c_n \in \mathbb{R}$. The resulting sequence is again in h^1 since

$$\|c\|^2_{h^1} = \sum_{n\in\mathbb{N}} n^2 |c_n|^2 = \sum_{n\in\mathbb{N}} \lim_{k\to\infty} n^2 |c_n^{(m_k)}|^2 \le \liminf_{k\to\infty} \sum_{n\in\mathbb{N}} n^2 |c_n^{(m_k)}|^2 \le C^2$$

where the discrete version of Fatou's Lemma has been used. To show that the extracted subsequence converges in l^2 we use the observation that for any $d \in h^1$ we can estimate the tail $\mathcal{X}_{n>N} d$ according to

$$\|\mathcal{X}_{n>N} d\|_{l^2} = \left(\sum_{n>N} |d_n|^2 \right)^{\frac{1}{2}} \le \frac{1}{N} \left(\sum_{n>N} n^2 |d_n|^2 \right)^{\frac{1}{2}} \le \frac{\|d\|_{h^1}}{N}.$$

($\mathcal{X}_{n>N}$ is a sequence which is equal to one for $n > N$ and zero otherwise. The product $\mathcal{X}_{n>N} d$ refers to a component wise product of two sequences.) To a given $\epsilon > 0$ we then pick N such that $C/N < \epsilon$. After that, the leading parts $\mathcal{X}_{n\le N} c^{(m_k)}$ can be replaced by $\mathcal{X}_{n\le N} c$ with an error of at most ϵ if k is sufficiently large because $c_n^{(m_k)} \to c_n$ for each n. Altogether,

$$\|c^{(m_k)} - c\|_{l^2} \le \|\mathcal{X}_{n\le N}(c^{(m_k)} - c)\|_{l^2} + \|\mathcal{X}_{n>N} c^{(m_k)}\|_{l^2} + \|\mathcal{X}_{n>N} c\|_{l^2} \le 3\epsilon.$$

which shows that the subsequence converges in l^2 and thus completes the proof. For more general results in the case of sequence spaces we refer to Exercise C.5.

C.2. Generalized traces

It turns out that for elements of Sobolev spaces with $m \geq 1$ the restriction to the boundary $\partial\Omega$ (the so called *trace*) can be defined in a natural way provided $\partial\Omega$ is sufficiently smooth. One reasons for this is that differentiable functions are dense in the Sobolev space and for these functions the trace is obtained in the usual way. Second, the trace operation turns out to be a bounded linear mapping into certain $\mathbb{L}^q$ spaces on $\partial\Omega$ so that the natural trace for smooth functions can be extended to all elements of the Sobolev space. More precisely, the space $C^m(\bar{\Omega})$ is dense in $W^{m,p}(\Omega)$ if Ω has the so called *segment property* (to each $x \in \partial\Omega$ there exists a neighborhood U_x and a vector $y \neq 0$ such that with $z \in U_x \cap \bar{\Omega}$ the whole segment $z + ty$, $t \in (0, 1)$ belongs to Ω; the segment property rules out that Ω lies on both sides on parts of its boundary).

Theorem C.3. *If $\Omega \subset \mathbb{R}^n$ has the segment property, then the set of restrictions to Ω of functions in $C_0^\infty(\mathbb{R}^n)$ is dense in $W^{m,p}(\Omega)$ for $1 \leq p < \infty$. If $u \in C(\bar{\Omega}) \cap W^{m,p}(\Omega)$ then u can be approximated in $W^{m,p}$ and locally uniformly by $C_0^\infty(\mathbb{R}^n)$-functions.*

Proof: The proof is based on a partition of unity argument. Locally at the boundary, the $W^{m,p}$ function is slightly moved outside the domain (along the vector y provided by the segment property) and afterwards smoothed by mollification. The resulting sequence of functions is easily seen to converge in $W^{m,p}$ due to the continuity of the translation operator (in $\mathbb{L}^p$) and the convergence of mollification (in $\mathbb{L}^p$). For more details, we refer to [**1**]. The part concerning continuous u is contained in the argument but not explicitly mentioned. ∎

Due to Theorem C.3, the trace operation

$$Bf = f|_{\partial\Omega}$$

makes sense at least for a dense subset of $W^{m,p}(\Omega)$. To extend the trace operator B to the whole space $W^{m,p}(\Omega)$ we need some boundedness property with respect to $\|\cdot\|_{m,p}$. Indeed, if we have

$$\|Bf\|_X \leq C\|f\|_{m,p,\Omega} \qquad \forall f \in C_0^\infty(\mathbb{R}^n) \tag{C.1}$$

for some Banach space X then we can define for any $f \in W^{m,p}(\Omega)$

$$Bf := \lim_{n\to\infty} Bf_k \tag{C.2}$$

where $(f_k)_{k\in\mathbb{N}}$ is a sequence in $C_0^\infty(\mathbb{R}^n)$ which converges to f in $W^{m,p}(\Omega)$ (such sequences exist due to Theorem C.3). The limit in (C.2) exists since X is a Banach space and Bf_k is, in fact, a Cauchy sequence

$$\|Bf_k - Bf_l\|_X \leq C\|f_k - f_l\|_{m,p,\Omega}.$$

Also, the definition (C.2) is independent of the chosen sequence. If $(g_k)_{k\in\mathbb{N}}$ is another sequence with the same properties as $(f_k)_{k\in\mathbb{N}}$ then

$$\|Bf_k - Bg_k\|_X \le C\|f_k - g_k\|_{m,p} \le C(\|f_k - f\|_{m,p} + \|g_k - f\|_{m,p}) \xrightarrow[n\to\infty]{} 0.$$

The crucial estimate (C.1) can be obtained for $X = \mathbb{L}^q(\partial\Omega)$ with a suitable $q \in [1,\infty)$ if we assume that Ω is a bounded domain with a Lipschitz continuous boundary and if $m \ge 1$. In general, it turns out that the trace bf of a $W^{m,p}$–function f with $m \ge 1$ has less regularity than f itself. A detailed investigation of this issue gives rise to Sobolev spaces of fractional order $X = W^{s,q}(\partial\Omega)$ with $s \in \mathbb{R}^+$. Under suitable assumptions on the smoothness of $\partial\Omega$ one can show, for example, that B maps $W^{1,p}(\Omega)$ onto $W^{1-1/p,p}(\partial\Omega)$ (see [**12**]). We will not go into these details but content ourselves with the case $X = \mathbb{L}^q(\Omega)$.

The boundary of Ω is assumed to be at least Lipschitz continuous in the following. That means, locally $\partial\Omega$ can be described as the graph of a Lipschitz continuous function with Ω lying on one side of the surface (if $\partial\Omega$ is unbounded, additional uniform bounds are required for example for the Lipschitz constants of the functions describing the surface). For a detailed discussion we refer to [**12**]. The general idea is to reduce the investigations to half space problems by suitable transformations and partition of unity arguments. Since these auxiliary steps are only of technical nature we focus our attention solely on the case

$$\Omega = \{\, x \in \mathbb{R}^n \ : \ x_n > 0 \,\}$$

to exhibit the main arguments. (Observe that Ω is not bounded but $\partial\Omega$ is so simple that the ideas necessary for bounded domains also apply.)

Theorem C.4. *Let $\Omega \subset \mathbb{R}^n$ be a bounded $C^{0,1}$ domain. Then there exists a uniquely determined continuous linear mapping $B : W^{1,p}(\Omega) \mapsto \mathbb{L}^q(\partial\Omega)$ such that $Bu = u|_{\partial\Omega}$ for all $u \in C(\bar{\Omega}) \cap W^{1,p}(\Omega)$. If $1 \le p < n$ then $1 \le q \le (np-p)/(n-p)$. Otherwise, any $q \in [1,\infty)$ is possible. In particular, B maps $W^{1,p}(\Omega)$ into $\mathbb{L}^p(\partial\Omega)$ and the image is dense.*

Proof: First, we consider the case $1 \le p < n$ and $q = (np-p)/(n-p)$. For a detailed proof we refer to [**12**] and consider only the case where Ω is a half space. We will show for $u \in C_0^\infty(\mathbb{R}^n)$ that

$$\|Bu\|_{0,q,\partial\Omega} \le C\|u\|_{1,p,\Omega}.$$

In view of our considerations above, this suffices to prove the existence of the trace operator B on $W^{1,p}(\Omega)$. We first note that

$$q = \frac{n-1}{n-p}p \ge p \ge 1.$$

If $u \in C_0^\infty(\mathbb{R}^n)$, there exists $R > 0$ such that $u(\hat{x}, R) = 0$ for all $\hat{x} \in \mathbb{R}^{n-1}$. Applying the fundamental theorem of calculus, we easily get

$$|u(\hat{x}, 0)|^q \leq \int_0^R |u|^{q-1} \left| \frac{\partial u}{\partial x_n} \right| (\hat{x}, x_n)\, dx_n. \tag{C.3}$$

Integrating over $\partial\Omega$ and applying Hölder's inequality, we obtain

$$\|Bu\|_{0,q,\partial\Omega}^q \leq q \|u\|_{0,\frac{np}{n-p},\Omega}^{q-1} \|u\|_{1,p,\Omega}$$

where we have used the relation

$$\frac{q-1}{1-\frac{1}{p}} = \frac{np}{n-p}.$$

For $q = 1 = p$ the norm estimate is already finished. In the other cases we use the Sobolev embedding theorem C.2 which shows that $W^{1,p}(\Omega)$ can be viewed as a subspace of $\mathbb{L}^{\frac{np}{n-p}}(\Omega)$ with

$$\|u\|_{0,\frac{np}{n-p},\Omega} \leq C\|u\|_{1,p,\Omega}$$

so that we can conclude

$$\|Bu\|_{0,q,\partial\Omega} \leq qC^{\frac{q-1}{q}} \|u\|_{1,p,\Omega}.$$

The case $1 \leq p < n$ and $1 \leq q \leq (np-p)/(n-p)$ (now with bounded Ω)is reduced to the previous one with the remark that $\mathbb{L}^{q_2}(\partial\Omega)$ is continuously embedded into $\mathbb{L}^{q_1}(\partial\Omega)$ if $q_1 \leq q_2$ since $\partial\Omega$ is bounded, i.e.

$$\|Bu\|_{0,q_1,\partial\Omega} \leq C\|Bu\|_{0,q_2,\partial\Omega} \qquad \text{for } q_1 \leq q_2.$$

Similarly, $p \geq n$ is reduced to the case above. Indeed, picking any $q \in [1,\infty)$ we can find $\hat{p} \in [1, n)$ such that $q = (n\hat{p} - \hat{p})/(n - \hat{p})$. Consequently, B is a bounded linear operator from $W^{1,\hat{p}}(\Omega)$ to $\mathbb{L}^q(\partial\Omega)$. Since $\partial\Omega$ is bounded and $\hat{p} < p$ we also have

$$\|u\|_{1,\hat{p},\Omega} \leq C\|u\|_{1,p,\Omega}.$$

In both cases we see that B maps $W^{1,p}(\Omega)$ into $\mathbb{L}^p(\partial\Omega)$. To show that the image is dense we consider again the simple (but crucial) case where Ω is the half space. For a given $\phi \in C_0^\infty(\mathbb{R}^{n-1})$ we can easily find a function $\phi \in C_0^\infty(\mathbb{R}^n)$ such that $B\phi = \psi$. Indeed, if $\theta \in C_0^\infty(\mathbb{R})$ is a function which is equal to one at the origin, we simply set $\phi(\hat{x}, x_n) = \psi(\hat{x})\theta(x_n)$. Since $\phi|_\Omega \in W^{1,p}(\Omega)$ we find that $C_0^\infty(\mathbb{R}^{n-1})$ is contained in $B(W^{1,p}(\Omega))$ which is therefore dense in $\mathbb{L}^p(\partial\Omega)$. Finally, if $u \in C(\bar{\Omega}) \cap W^{1,p}(\Omega)$ then the $W^{1,p}$–approximating sequence of $C_0^\infty(\mathbb{R}^n)$ functions is also locally uniformly convergent (w.l.o.g. due to Theorem C.3). Thus, $u_n(x) \to u(x)$ for all $x \in \partial\Omega$. We know already that $Bu_n \to Bu$ in an $\mathbb{L}^q$ space on $\partial\Omega$ so that in particular a subsequence of Bu_n converges a.e. on $\partial\Omega$ to Bu which implies $Bu = u|_{\partial\Omega}$ and the proof is complete. ∎

Using the trace of $W^{1,p}$ functions, we can immediately define traces of derivatives. For $u \in W^{2,p}(\Omega)$, for example, the first derivatives are in $W^{1,p}(\Omega)$ and thus $B\frac{\partial u}{\partial x_i}$ is well defined. In particular, if n is the normal vector field on $\partial\Omega$ then the normal derivative $\partial u/\partial n$ is well defined for $u \in W^{2,p}(\Omega)$.

Another immediate consequence of Theorem C.4 is that the integration by parts formula can be extended to $W^{1,p}$ functions. If $\phi \in C_0^\infty(\mathbb{R}^n)$ and $u \in W^{1,p}(\Omega)$ with $1 \le p < \infty$ then

$$\int_\Omega u \frac{\partial \phi}{\partial x_i}\, dx = \int_{\partial\Omega} (Bu)\phi n_i \, d\sigma - \int_\Omega \phi \frac{\partial u}{\partial x_i}\, dx. \tag{C.4}$$

(Note that for $C^{0,1}$ domains the outer unit normal field n is defined almost everywhere on $\partial\Omega$.) To prove (C.4) we approximate u by $C_0^\infty(\mathbb{R}^n)$ functions u_k. Then (C.4) holds with u_k in place of u due to Gauß theorem. Going over to the $W^{1,p}$ limit and noting that Bu_k converges in some $\mathbb{L}^q(\partial\Omega)$ to Bu so that $(Bu_k)\phi \to (Bu)\phi$ in $\mathbb{L}^1(\partial\Omega)$ we obtain (C.4). If we now take an approximating sequence $(\phi_k)_{k\in\mathbb{N}} \subset C_0^\infty(\mathbb{R}^n)$ which approximates $v \in W^{1,q}(\Omega)$, we can pass to the limit in (C.4) with ϕ_k in place of ϕ which yields

$$\int_\Omega u \frac{\partial v}{\partial x_i}\, dx = -\int_\Omega v \frac{\partial u}{\partial x_i}\, dx + \int_{\partial\Omega} (Bu)(Bv) n_i \, d\sigma$$

for $u \in W^{1,p}(\Omega)$ and $v \in W^{1,q}(\Omega)$. Similarly, one can show Green's formula for $u \in W^{1,p}(\Omega)$ and $v \in W^{2,q}(\Omega)$

$$\int_\Omega u \Delta v \, dx = -\int_\Omega \nabla u \cdot \nabla v \, dx + \int_{\partial\Omega} (Bu) n \cdot B(\nabla v)\, d\sigma.$$

In order to treat Dirichlet boundary problems, the closed subspace of functions in $W^{1,p}(\Omega)$ which have zero trace on $\partial\Omega$ (the kernel $\ker B$ of B) is especially important. Obviously, $C_0^\infty(\Omega) \subset \ker B$ and one can even show that the compactly supported C^∞ functions on Ω are dense in $\ker B$.

Lemma C.5. *Let Ω be a bounded $C^{0,1}$–domain in $\mathbb{R}^n$. Then, the closure of $C_0^\infty(\Omega)$ in the $W^{1,p}(\Omega)$ norm contains exactly those $W^{1,p}(\Omega)$ functions which vanish on $\partial\Omega$.*

Proof: Again, we refer to [**12**] for a detailed proof of this result and present the main idea only for the half space. We have already mentioned that the closure of $C_0^\infty(\Omega)$ certainly gives rise to $W^{1,p}(\Omega)$ functions which vanish on the boundary. Conversely, let $u \in W^{1,p}(\Omega)$ with $Bu = 0$. We define

$$v(x) = \begin{cases} u(x) & x \in \Omega \\ 0 & x \in \Omega^c \end{cases}$$

Then, $v \in \mathbb{L}^p(\mathbb{R}^n)$ and the distributional x_i–derivative is given by the zero extension of $\partial u/\partial x_i$. Indeed, for $\phi \in C_0^\infty(\mathbb{R}^n)$

$$\left\langle \frac{\partial v}{\partial x_i}, \phi \right\rangle = -\left\langle v, \frac{\partial \phi}{\partial x_i} \right\rangle = -\int_\Omega u \frac{\partial \phi}{\partial x_i}\, dx.$$

Using (C.4) we get

$$\left\langle \frac{\partial v}{\partial x_i}, \phi \right\rangle = -\int_{\partial\Omega} (Bu) \frac{\partial \phi}{\partial x_i} n_i \, d\sigma + \int_{\Omega} \frac{\partial u}{\partial x_i} \phi \, dx$$

and since $Bu = 0$ we get

$$\frac{\partial v}{\partial x_i}(x) = \begin{cases} \frac{\partial u}{\partial x_i} & x \in \Omega \\ 0 & x \in \Omega^c \end{cases}$$

In particular, the zero extension of a traceless $W^{1,p}(\Omega)$ function is contained in $W^{1,p}(\mathbb{R}^n)$. Since the translation operator

$$T_h v(x) = v(x + he_n)$$

is continuous on $\mathbb{L}^p(\mathbb{R}^n)$ (see [**1**]) it is also continuous on $W^{1,p}(\mathbb{R}^n)$ because

$$\frac{\partial}{\partial x_i}(T_h v) = T_h \frac{\partial v}{\partial x_i}.$$

The main idea of the proof is now to approximate v in $W^{1,p}(\mathbb{R}^n)$ by $T_{-h}v$ for a small $h > 0$. The shifted function $T_{-h}v$ is then supported inside the half space (a distance h away from the boundary) and can thus be approximated by $C_0^\infty(\Omega)$ functions (for example by cut–off and mollification). Consequently, v can be obtained as the $W^{1,p}(\Omega)$ limit of $C_0^\infty(\Omega)$ functions. ■

This characterization of $\ker B$ leads to a natural generalization. Even on domains Ω where a reasonable trace operator cannot be constructed, the closure of $C_0^\infty(\Omega)$ is well defined and contains, in some sense, functions which vanish on the boundary (because they can be approximated by functions which are zero on $\partial\Omega$). This gives rise to

Definition C.6. *Let $\Omega \subset \mathbb{R}^n$ be open and let $1 \le p \le \infty$, $m \in \mathbb{N}_0$. Then, $W_0^{m,p}(\Omega)$ denotes the closure of $C_0^\infty(\Omega)$ in $W^{m,p}(\Omega)$.*

Lemma C.5 now has the following counterpart.

Theorem C.7. *Let Ω be a bounded $C^{0,1}$–domain in $\mathbb{R}^n$, let $m \ge 1$ and $1 \le p < \infty$. Then,*

$$W_0^{m,p}(\Omega) = \{\, u \in W^{m,p}(\Omega) \,:\, B(\nabla^\alpha u) = 0, \quad \forall |\alpha| < m \,\}.$$

Proof: For $m = 1$ this is just Lemma C.5 so let us assume that the result is correct for $m - 1$. If $u \in W_0^{m,p}$, then $\nabla^\beta u \in W_0^{1,p}(\Omega)$ if $|\beta| = m - 1$. Thus $B(\nabla^\beta u) = 0$ according to Lemma C.5. By induction assumption and since $W_0^{m,p}(\Omega) \subset W_0^{m-1,p}(\Omega)$ we already have for all lower derivatives $B(\nabla^\alpha u) = 0$ for $|\alpha| < m - 1$. Altogether,

$$u \in \{\, u \in W^{m,p}(\Omega) \,:\, B(\nabla^\alpha u) = 0, \quad \forall |\alpha| < m \,\}.$$

The converse direction is shown in a similar way based on Lemma C.5. ■

We remark that $Bu = 0$ implies automatically that all tangential derivatives of u vanish on the boundary (if the boundary is smooth enough to allow for tangential derivatives). Thus, the statement $Bu = 0$ and $B(\nabla u) = 0$, which holds for $W_0^{2,p}(\Omega)$ functions, is equivalent to $Bu = 0$ and $B(\partial u/\partial n) = 0$ since the derivatives in the remaining $n-1$ directions are tangential. Introducing the l^{th} normal derivative

$$\frac{\partial^l \phi}{\partial n^l}(x) := \left(\frac{d}{dt}\right)^l \phi(x + tn(x))\bigg|_{t=0} = \sum_{|\alpha|=l} \frac{l!}{\alpha!} n^\alpha \nabla^\alpha \phi \qquad \phi \in C^\infty.$$

(observe that the t derivative is half sided with $t \leq 0$) we thus have

Lemma C.8. *Let $1 \leq p < \infty$, $m \in \mathbb{N}$ and let Ω be a bounded $C^{m,1}$ domain. Then $u \in W^{m,p}(\Omega)$ is contained in $W_0^{m,p}(\Omega)$ if and only if $\partial^l u/\partial n^l = 0$ a.e. on $\partial\Omega$ (in the sense of trace) for $l = 0, \dots, m-1$.*

Proof: The idea is similar to the one of Lemma C.5 (see the reference cited in [**12**]). ■

C.3. Exercises

Exercise C.1. Show that $u(x) = \sin\frac{1}{x}$ is contained in $\mathbb{L}^p(0,1)$ but not in $W^{1,p}(0,1)$ for any $p \in [1,\infty]$. To which Sobolev space $W^{m,p}(0,1)$ does $v(x) = x^s$ belong for different values of $s \in \mathbb{R}$?

Exercise C.2. Let $\Omega = \{x \in \mathbb{R}^n : x_n > 0\}$ be the half space and let $u \in C_0^\infty(\mathbb{R}^n)$. Show (directly without using Sobolev embedding) that the $\mathbb{L}^2(\partial\Omega)$ norm of the boundary value of u can be estimated by the $W^{1,2}$ norm on Ω

$$\|Bu\|_{0,2,\partial\Omega} \leq \|u\|_{1,2,\Omega}.$$

Exercise C.3. Let $\Omega \subset \mathbb{R}^n$ be open and $u \in W^{m,p}(\Omega)$ with $m \in \mathbb{N}$ and $p \in [1,\infty)$. Show that the zero extension of u

$$\tilde{u}(x) := \begin{cases} u(x) & x \in \Omega \\ 0 & x \in \Omega^c \end{cases}$$

is an element of $W^{m,p}(\mathbb{R}^n)$ if u and its derivatives up to order $m-1$ vanish on the boundary (i.e. $u \in W_0^{m,p}(\Omega)$).

Exercise C.4.

a) Let $\phi_0(x) = 1$, $\phi_n(x) = \sqrt{2}\cos(n\pi x)$ for $n \in \mathbb{N}$. Show that $\{\phi_n : n \in \mathbb{N}\}$ is an orthonormal basis of $\mathbb{L}^2(0,1)$.
(Hint: Check orthonormality; use the mapping $\cos(\pi\cdot) : [0,1] \mapsto [-1,1]$ and the Weierstraß approximation theorem to show that any $f \in C([0,1])$ can be uniformly approximated by expressions $\sum_{n=0}^N a_n(\cos(\pi x))^n$; conclude the proof with elementary trigonometric identities and the density of $C([0,1])$ in $\mathbb{L}^2(0,1)$.)

b) Let $\psi_n(x) = \sqrt{2}\sin(n\pi x)$ for $n \in \mathbb{N}$. Show that $\{\psi_n : n \in \mathbb{N}\}$ is an orthonormal basis of $\mathbb{L}^2(0,1)$.
(Hint: After checking orthonormality, use that $f(x)/\sin(\pi x)$ is continuous on $[0,1]$ for any $f \in \mathscr{D}(0,1)$ together with the result in a), an elementary trigonometric identity, and the density of $\mathscr{D}(0,1)$ in $\mathbb{L}^2(0,1)$.)

c) Show that $\{e^{in\pi x}/\sqrt{2} : n \in \mathbb{Z}\}$ is a basis of $\mathbb{L}^2(-1,1)$.
(Hint: Use a) and b) and the decomposition of any function on $(-1,1)$ into an odd and even part.)

d) Use an appropriate normalization of the functions ψ_n in b) to construct an orthonormal basis of $H_0^1(0,1)$.

Exercise C.5. Let $s = \{(\gamma_n)_{n\in\mathbb{N}} : \gamma_n \in \mathbb{R}\}$ be the space of all real valued sequences. The subset of all terminating (or compactly supported) sequences is denoted

$$d = \{\gamma \in s : \gamma_n \neq 0 \quad \text{for finitely many } n\}.$$

A sequence $(\gamma^{(m)})_{m\in\mathbb{N}} \subset d$ is said to be convergent to γ in d if there exists $N \in \mathbb{N}$ such that $\gamma_n^{(m)} = 0$ for all $m \in \mathbb{N}$ and $n > N$ and $\gamma_n^{(m)} \to \gamma_n$ for $m \to \infty$ and $n \le N$.

- Show that the dual d' can be identified with s based on the dual product
$$\langle \Lambda, \phi\rangle = \sum_{n\in\mathbb{N}} \Lambda_n \phi_n \qquad \forall \Lambda \in s, \quad \phi \in d.$$
(Note the similarity between $d, d' = s$ and $\mathscr{D}, \mathscr{D}'$.)

For $s \in \mathbb{R}$, $1 \le p \le \infty$ and $\gamma \in s$ we define

$$\|\gamma\|_{s,p} := \begin{cases} \left(\sum_{n\in\mathbb{N}} n^{sp}|\gamma_n|^p\right)^{\frac{1}{p}} & 1 \le p < \infty \\ \max_{n\in\mathbb{N}} n^s|\gamma_n| & p = \infty \end{cases}$$

and set $w^{s,p} = \{\gamma \in s : \|\gamma\|_{s,p} < \infty\}$.

- Show that d is dense in $w^{s,p}$ for any $1 \le p < \infty$ and $s \in \mathbb{R}$ but not if $p = \infty$.
- Show that $w^{s,p_1} \hookrightarrow w^{s,p_2}$ whenever $p_1 \le p_2$ (i.e. $w^{s,p_1} \subset w^{s,p_2}$ and $\|\cdot\|_{s,p_2} \le C\|\cdot\|_{s,p_1}$).
- Show that $w^{s_1,p_1} \hookrightarrow w^{s_2,p_2}$ is a compact embedding whenever $p_1 \le p_2$ and $s_1 > s_2$.
- Show that $w^{s,\infty} \hookrightarrow w^{t,p}$ for $(s-t)p > 1$.

Bibliography

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975

[2] D. Braess, *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*, Springer, 1997

[3] W. Bunse, A. Bunse–Gerstner, *Numerische lineare Algebra*, Teubner, 1985

[4] Gen. Edts. P.G. Ciarlet, J.L. Lions, *Handbook of numerical analysis Vol II: Finite Element Method (part 1)*, Amsterdam : North-Holland,1991

[5] R. Dautray, J.L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2, Springer 1988

[6] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, 1997

[7] P. Deuflhard, A. Hohmann, *Numerische Mathematik*, Springer, 1991

[8] N. Dunford, J. T. Schwartz, *Linear Operators Part I: General Theory*, Interscience, 1957

[9] D. Gilbarg, N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, 1977

[10] W. Hackbusch, *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner, 1986

[11] L. Hörmander, *The Analysis of Linear Partial Differential Operators I*, Springer, 1983

[12] A. Kufner, O. John, S. Fucik, *Function spaces*, Nordhoff International Publishing, 1977

[13] P. K. Kythe, *Fundamental Solutions for Differential OPerators and Applications*, Birkhäuser, 1996

[14] M. Renardy, R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer, 1993

[15] W. Rudin, *Functional Analysis*, Mc Graw Hill, 1991

[16] L. N. Trefethen, D. Bau, *Numerical Linear Algebra*, SIAM, 1997

[17] F. Treves, *Basic Linear Partial Differential Equations*, Academic Press, 1975