MA643 - Numerical Analysis of Partial Differential Equations
Lecture Notes

Abhinav Jha
Indian Institute of Technology, Gandhinagar
Summer Semester 2025

Preface

These notes are designed to provide a structured and comprehensive understanding of the course content. They will cover key topics, concepts, and computational techniques that are fundamental to numerical analysis of partial differential equations. Please note that this is the first iteration (Version 1.0.0) of the notes and hence there is a chance that some of the content is incorrect. If you find some flaws, please email me at `abhinav.jha@iitgn.ac.in`.

Just because someone stumbles and loses their path, doesn't mean they're lost forever.
—Prof. Charles Xavier

Contents

Chapter 1

Partial Differential Equations

In real life, many physical phenomena are governed by differential equations. From the steam rising from your morning coffee to the spread of pollution in the air, these processes can often be modeled using partial differential equations (PDEs).

Figure 1.1: Smoke coming out of an industrial chimney.

Let us consider a simple example. Imagine smoke rising from an industrial chimney (see Fig. 1.1). What do you observe? First, the smoke is carried in the direction of the wind. Then, it begins to disperse into the surrounding air. Finally, although not directly visible, chemical reactions take place between the smoke and components of the atmosphere. The equation that models such behavior is known as the convection–diffusion–reaction equation—one among many PDEs used to describe natural phenomena.

The origin of differential calculus can be traced back to the foundational work of Isaac Newton and Gottfried Wilhelm Leibniz. The study of PDEs, however, is typically credited to Jean le Rond d'Alembert, who investigated the wave equation in the 18th century. Subsequent major contributions came from Leonhard Euler, who formulated what are now called the Euler equations for fluid dynamics.

This chapter provides a brief introduction to partial differential equations: how they are derived, how they are classified, and what challenges arise when attempting to solve them analytically. For a deeper understanding, the reader is encouraged to consult textbooks dedicated to this topic. Some recommended references include [14, 18].

Figure 1.2: From left to right: Isaac Newton: 4 January 1643 – 31 March 1727, Gottfried Wilhelm Leibniz: 1 July 1646 – 14 November 1716, Jean le Rond d'Alembert : 16 November 1717 – 29 October 1783, and Leonhard Euler : 15 April 1707 – 18 September 1783.

1.1 Heat Equation

Engineers and physicists are often interested in properties of materials that vary continuously in space and time. Let $\mathbf{x} := (x, y, z)$ denote the spatial coordinates and t denote time. One fundamental concept in this context is that of *conservation*.

So, what is a *conservation law*? At its core, a conservation law is concerned with tracking changes in a material property within a control volume V, which is enclosed by a surface S. It relates the change of that property inside V to its production within the volume and its flow across the boundary.

There are three key components involved in formulating a conservation law:

1. The quantity Q of the property per unit volume,
2. The rate F at which the property is produced or destroyed (i.e., net production) per unit volume, and
3. The net *flux* of the property through the surface S, i.e., the flow of the property into or out of the volume.

If $\boldsymbol{q}$ denotes the flux vector (i.e., flow rate per unit area) at a point inside V, then the flux across the boundary at a point on S is given by $\boldsymbol{q} \cdot \mathbf{n}$, where $\mathbf{n}$ is the outward unit normal vector (see Fig. 1.3).

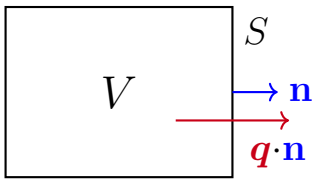


Figure 1.3: Flux vector $\boldsymbol{q}$ flowing through the boundary S enclosing volume V.

A conservation law can be succinctly stated as:

The rate of change of Q inside the volume is equal to the net production within the volume minus the net outward flux through its boundary.

Mathematically, the conservation principle can be expressed as

$$\frac{d}{dt}\int_V Q\,dV = \int_V F\,dV - \int_S \boldsymbol{q}\cdot\mathbf{n}\,dS. \tag{1.1}$$

Using the *divergence theorem* (also known as Gauss's theorem), the surface integral can be converted into a volume integral:

$$\int_S \boldsymbol{q}\cdot\mathbf{n}\,dS = \int_V \nabla\cdot\boldsymbol{q}\,dV.$$

Furthermore, if the control volume V is fixed (i.e., independent of time), we can interchange the order of time differentiation and volume integration in the first term of Eq. (1.1), yielding

$$\int_V \left(\frac{dQ}{dt} + \nabla\cdot\boldsymbol{q} - F\right) dV = 0,$$

where $\nabla\cdot\boldsymbol{q}$ denotes the divergence of the flux vector $\boldsymbol{q}$.

Since V is arbitrary, the integrand must vanish pointwise (see [19, Proposition 6.3.3]). This gives the local form of the conservation law:

$$\frac{dQ}{dt} + \nabla\cdot\boldsymbol{q} = F. \tag{1.2}$$

The quantities Q, $\boldsymbol{q}$, and F generally depend on the unknown variable u (e.g., temperature or concentration), the spatial position $\mathbf{x}$, time t, and physical properties of the medium such as conductivity or density.

We now turn to a classical example: the heat equation, where the quantity of interest is temperature, denoted by $u(t,\mathbf{x})$. Joseph Fourier is credited with the derivation and solution of the heat equation, which he introduced in his 1822 work *Théorie analytique de la chaleur*.

Figure 1.4: Joseph Fourier: 21 March 1768 – 16 May 1830

Let $C(u)$ be the specific heat capacity, i.e., the amount of heat energy required to raise the temperature of a unit mass by one degree. If u_0 is a reference (or base) temperature, then the thermal energy per unit mass is

$$E = \int_{u_0}^{u} C(u)\,du.$$

The thermal energy per unit volume is then ρE, where ρ is the material density. Thus, the conserved quantity is

$$Q = \rho E = \rho \int_{u_0}^{u} C(u)\, du.$$

If ρ and C are constant, this simplifies to

$$Q = \rho C(u - u_0).$$

The heat flux $\boldsymbol{q}$ is given by *Fourier's law*:

$$\boldsymbol{q} = -\lambda \nabla u,$$

where λ is the thermal conductivity—a material property that quantifies the ability to conduct heat. The negative sign indicates that heat flows from regions of high temperature to low temperature.

Assuming no other modes of heat transfer and letting $F(u, \mathbf{x}, t)$ represent the internal heat sources (or sinks) per unit volume, the energy conservation law (1.2) becomes

$$\rho C u_t - \nabla \cdot (\lambda \nabla u) = F.$$

If λ is constant, this further simplifies to the classical heat equation:

$$u_t - \varepsilon \Delta u = f, \tag{1.3}$$

where the thermal diffusivity is $\varepsilon = \frac{\lambda}{\rho C}$ and the source term is $f = \frac{F}{\rho C}$.

We assume that Eq. (1.3) is posed over a spatial domain Ω with boundary Γ, and the time variable t ranges over the interval $[0, T]$.

The heat equation is not limited to thermal processes. Equations of the form (1.3) arise in many other diffusive phenomena, such as chemical diffusion, pollutant dispersion, and image smoothing.

1.1.1 Boundary Conditions

Another important aspect of a partial differential equation (PDE) is the specification of appropriate *boundary conditions*. There are three major types of boundary conditions commonly encountered in physical problems:

1. **Dirichlet Boundary Condition** Named after Peter Gustav Lejeune Dirichlet, this condition prescribes the value of the solution itself on a portion of the boundary. In the context of heat transfer, it specifies the temperature $u(t, \mathbf{x})$ on a part of the boundary:

   $$u = g, \quad \text{on } (0, T) \times \Gamma_{\mathrm{D}},$$

   where $\Gamma_{\mathrm{D}} \subset \Gamma$. These are also referred to as *essential boundary conditions*.

2. **Neumann Boundary Condition** Named after Carl Neumann, this condition prescribes the heat flux across the boundary. It specifies the normal derivative of the solution on a portion of the boundary:
$$-\frac{\partial u}{\partial \mathbf{n}} = g, \quad \text{on } (0,T) \times \Gamma_{\mathrm{N}},$$
where $\Gamma_{\mathrm{N}} \subset \Gamma$. These are also called *natural boundary conditions.*

3. **Robin Boundary Condition** Named after Victor Gustav Robin, this condition models convective heat exchange at the boundary according to Newton's law of cooling. It is a combination of Dirichlet and Neumann conditions:
$$\frac{\partial u}{\partial \mathbf{n}} + h(u - u_{\mathrm{env}}) = 0, \quad \text{on } (0,T) \times \Gamma_{\mathrm{R}},$$
where $\Gamma_{\mathrm{R}} \subset \Gamma$, h is the heat transfer coefficient, and u_{env} is the ambient temperature.

In addition to boundary conditions, an initial condition is required to fully determine the solution. This specifies the state of the system at the initial time $t = 0$:

$$u(0, \mathbf{x}) = u_0(\mathbf{x}), \quad \text{for } \mathbf{x} \in \Omega.$$

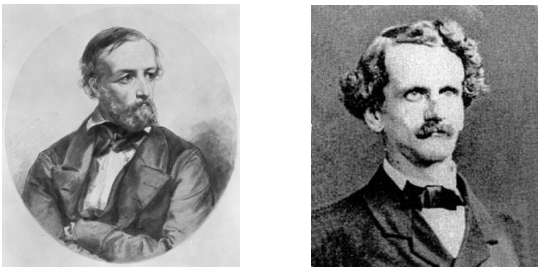


Figure 1.5: Peter Gustav Lejeune Dirichlet : 13 February 1805 – 5 May 1859 (left) and Carl Neumann : 7 May 1832 – 27 March 1925 (right).

1.1.2 Poisson Equation

If the temperature remains constant in time, we obtain a special case of the heat equation:

$$-\varepsilon \Delta u = f, \quad \text{in } \Omega, \tag{1.4}$$

which is known as the *Poisson equation.* It is named after Siméon Denis Poisson, who published it in 1823. When the source term vanishes, i.e., $f(\mathbf{x}) = 0$, the equation reduces to the *Laplace equation*, named after Pierre-Simon Laplace, who studied it in 1786.

Solutions to the Laplace equation are known as *harmonic functions.* These functions play a central role in an area of mathematics known as *harmonic analysis*, which connects PDEs with Fourier analysis, potential theory, and other fields.

The focus of this course will be on developing numerical methods for solving the Poisson equation.


Figure 1.6: Siméon Denis Poisson : 21 June 1781 – 25 April 1840 (left) and Pierre-Simon Laplace : 23 March 1749 – 5 March 1827 (right).

1.1.3 Non-Dimensional Form

Until now, we have discussed PDEs from the perspective of engineers and physicists, deriving equations based on physical laws. However, one important aspect we have not yet addressed is the role of *units* or *dimensions* in these equations.

In the natural sciences, units play a fundamental role. The most widely used and standardized unit system is the *International System of Units (SI)*. For example, the SI unit of energy is the joule. More fundamentally, the SI system is built upon seven *base quantities*, each associated with a specific unit (see Table 1.1). Other units can be expressed in terms of these base units. For instance, one joule is equivalent to kg m^2 s^{-2} in base units.

Symbol	**Name**	**Base Quantity**
s	second	time
m	metre	length
kg	kilogram	mass
A	ampere	electric current
K	kelvin	thermodynamic temperature
mol	mole	amount of substance
cd	candela	luminous intensity

Table 1.1: SI base units corresponding to the seven fundamental physical quantities.

We now examine whether the heat equation (1.3) is *dimensionally consistent*, i.e., whether all terms in the equation have the same physical units. From this point onward, we restrict ourselves to the SI base unit system.

Recall that $u(t, \mathbf{x})$ represents temperature, which is measured in kelvins (K). The thermal diffusivity $\varepsilon = \lambda/(\rho C)$ is a material parameter composed of:

- Thermal conductivity λ with units kg m s^{-3} K^{-1},
- Density ρ with units kg m^{-3},
- Specific heat capacity C with units m^2 s^{-2} K^{-1}.

Thus, the thermal diffusivity ε has units:

$$\varepsilon = \frac{\lambda}{\rho C} \quad \Rightarrow \quad \mathrm{m}^2\ \mathrm{s}^{-1}.$$

The source term is given by $f = F/(\rho C)$, where F has units of heat generated per unit volume per unit time: kg m^{-1} s^{-3}. Consequently,

$$f = \frac{F}{\rho C} \quad \Rightarrow \quad \mathrm{K\ s}^{-1}.$$

On the left-hand side of the heat equation, the time derivative u_t has units K s^{-1}, while the Laplacian Δu has units K m^{-2}, and multiplying it by ε yields K s^{-1}. Therefore, all three terms in the heat equation have consistent SI base units, confirming that the equation is dimensionally valid.

In numerical analysis, it is often beneficial to work with a *dimensionless form* of the equation. This allows for a formulation that is independent of measurement systems—for example, temperature could be measured in either Kelvin or Celsius.

Let L, U, and T^* represent the characteristic length, temperature, and time scales of the problem, respectively. We define the dimensionless variables (denoted by primes) as:

$$\mathbf{x}' = \frac{\mathbf{x}}{L}, \quad u' = \frac{u}{U}, \quad t' = \frac{t}{T^*}.$$

Using the chain rule, the derivatives transform as:

$$u_t = \frac{U}{T^*} u'_{t'}, \quad \Delta u = \frac{U}{L^2} \Delta' u'.$$

Substituting these into the heat equation gives:

$$\begin{aligned} \frac{U}{T^*} u'_{t'} - \varepsilon \frac{U}{L^2} \Delta' u' = f, \quad \text{in } \left(0, \frac{T}{T^*}\right) \times \Omega', \\ u'_{t'} - \varepsilon \frac{T^*}{L^2} \Delta' u' = \frac{T^*}{U} f, \quad \text{in } \left(0, \frac{T}{T^*}\right) \times \Omega', \end{aligned}$$

where Ω' is the dimensionless spatial domain. Note that both $\varepsilon \frac{T^*}{L^2}$ and $\frac{T^*}{U} f$ are dimensionless, so all terms in the rescaled equation are nondimensional.

By abuse of notation, we drop the primes and obtain the *dimensionless heat equation*:

$$u_t - \varepsilon \Delta u = f, \quad \text{in } (0, T) \times \Omega,$$

where ε and f are now interpreted as nondimensional parameters. A similar procedure can be used to obtain the dimensionless form of the Poisson equation.

1.2 Classification of PDEs

After exploring the modeling and physical derivation of partial differential equations (PDEs), we now shift to a more abstract and mathematical perspective. In this course, we are primarily interested in a specific class of PDEs known as *second-order partial differential equations*. To make this precise, we begin by defining the notion of the *order* of a PDE.

Definition 1.1 (**Order of a PDE**). The *order* of a partial differential equation is the order of the highest derivative of the unknown function appearing in the equation. For example, if the highest derivative is a second derivative, such as $\partial^2 u/\partial x^2$, then the equation is said to be of *second order*.

Second-order PDEs play a central role in modeling a wide range of physical phenomena. These equations are broadly classified into three types: *elliptic*, *parabolic*, and *hyperbolic*.

Let $\Omega \subset \mathbb{R}^d$, where $d \in \mathbb{N}$. A general linear second-order PDE can be written in the form

$$\sum_{j,k=1}^{d} a_{jk}(\mathbf{x})\, \partial_j \partial_k u(\mathbf{x}) + F(\mathbf{x}, u, \partial_1 u, \ldots, \partial_d u) = 0,$$

or equivalently,

$$\nabla \cdot (\mathbf{A}(\mathbf{x}) \nabla u) + F(\mathbf{x}, u, \partial_1 u, \ldots, \partial_d u) = 0,$$

where $\mathbf{A}(\mathbf{x}) = [a_{jk}(\mathbf{x})]$ is a matrix-valued function defined on Ω.

If $u(\mathbf{x})$ is sufficiently smooth, then by Schwarz's theorem, the mixed second partial derivatives commute: $\partial_j \partial_k u = \partial_k \partial_j u$. This allows us to assume that the coefficient matrix $\mathbf{A}(\mathbf{x})$ is symmetric. Even if $\mathbf{A}(\mathbf{x})$ is not symmetric initially, we can rewrite the equation in symmetric form:

$$\begin{aligned}
\sum_{j,k=1}^{d} a_{jk}(\mathbf{x})\, \partial_j \partial_k u(\mathbf{x}) &= \sum_{j<k} a_{jk}(\mathbf{x})\, \partial_j \partial_k u + \sum_{j>k} a_{jk}(\mathbf{x})\, \partial_j \partial_k u + \sum_{j=1}^{d} a_{jj}(\mathbf{x})\, \partial_{jj} u \\
&= \sum_{j<k} a_{jk}(\mathbf{x})\, \partial_j \partial_k u + \sum_{j>k} a_{jk}(\mathbf{x})\, \partial_k \partial_j u + \sum_{j=1}^{d} a_{jj}(\mathbf{x})\, \partial_{jj} u \\
&= \sum_{j<k} (a_{jk}(\mathbf{x}) + a_{kj}(\mathbf{x}))\, \partial_j \partial_k u + \sum_{j=1}^{d} a_{jj}(\mathbf{x})\, \partial_{jj} u \\
&= \sum_{j,k=1}^{d} \left(\frac{a_{jk}(\mathbf{x}) + a_{kj}(\mathbf{x})}{2} \right) \partial_j \partial_k u(\mathbf{x}).
\end{aligned}$$

Hence, without loss of generality, we may take $\mathbf{A}(\mathbf{x})$ to be symmetric. As a result, all eigenvalues of $\mathbf{A}(\mathbf{x})$ are real.

The classification of a second-order PDE depends on the signs of the eigenvalues of the matrix $\mathbf{A}(\mathbf{x})$. Let:

- α be the number of positive eigenvalues,
- β the number of negative eigenvalues,
- γ the number of zero eigenvalues.

We say the PDE is of type (α, β, γ), and classify it as:

- **Elliptic PDE** : Type $(d, 0, 0)$ or $(0, d, 0)$.
- **Parabolic PDE** : Type $(d-1, 0, 1)$ or $(0, d-1, 1)$.
- **Hyperbolic PDE** : Type $(d-1, 1, 0)$ or $(1, d-1, 0)$.

This matrix-based definition may be unfamiliar to those who have previously studied PDEs only in two variables. To reconcile the two perspectives, consider the classical form of a second-order PDE in two dimensions:

$$\mathbf{A}u_{xx} + \mathbf{B}u_{xy} + \mathbf{C}u_{yy} + \mathbf{D}u_x + \mathbf{E}u_y + \mathbf{F}u = \mathbf{G}.$$

Focusing only on the second-order part, the associated coefficient matrix is:

$$\begin{bmatrix} \mathbf{A} & \frac{\mathbf{B}}{2} \\ \frac{\mathbf{B}}{2} & \mathbf{C} \end{bmatrix},$$

with determinant

$$\mathbf{AC} - \frac{\mathbf{B}^2}{4}.$$

From linear algebra, the determinant of a 2×2 symmetric matrix equals the product of its eigenvalues (see [25]). Therefore, the sign of the discriminant $\mathbf{B}^2 - 4\mathbf{AC}$ determines the PDE type:

- Elliptic if $\mathbf{B}^2 - 4\mathbf{AC} < 0$,
- Parabolic if $\mathbf{B}^2 - 4\mathbf{AC} = 0$,
- Hyperbolic if $\mathbf{B}^2 - 4\mathbf{AC} > 0$.

This classical definition coincides with the eigenvalue-based classification and extends naturally to higher dimensions.

Examples:

- The **Poisson equation** is elliptic.
- The **Heat equation** is parabolic.
- The **Wave equation** is hyperbolic.

1.3 Analytical Solution to Elliptic PDEs

Until now, we have derived partial differential equations (PDEs), classified them, and discussed various boundary conditions. We now turn to the important question of the *existence and uniqueness* of solutions, followed by how such solutions can be computed.

Before presenting the main result, we introduce some useful notations and definitions.

Definition 1.2 (**Multi-Index**). A d-dimensional multi-index is a tuple $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, where each $\alpha_i \in \mathbb{N} \cup \{0\}$. The corresponding partial derivative is defined as

$$\partial^{\boldsymbol{\alpha}} := \partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_d^{\alpha_d}, \quad \text{where} \quad \partial_i^{\alpha_i} := \frac{\partial^{\alpha_i}}{\partial x_i^{\alpha_i}}.$$

We denote the order by $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i$.

Example 1.3. Let $|\boldsymbol{\alpha}| = 2$ If $d = 2$ then $\boldsymbol{\alpha} = (1,1), (2,0)$, and $(0,2)$ and if $d = 3$ then $\boldsymbol{\alpha} = (1,1,0), (0,1,1), (1,0,1), (2,0,0), (0,2,0)$, and $(0,0,2)$.

Definition 1.4 (**Hölder Continuous Functions**). Let Ω be an open subset of $\mathbb{R}^d$, $0 < \alpha \leq 1$, and $k \geq 0$ an integer. The space of Hölder continuous functions is defined as

$$\mathcal{C}^{k,\alpha}(\overline{\Omega}) := \left\{ f \in \mathcal{C}^k(\overline{\Omega}) : \sup_{\mathbf{x} \neq \mathbf{y} \in \overline{\Omega}} \frac{|\partial^{\boldsymbol{\beta}} f(\mathbf{x}) - \partial^{\boldsymbol{\beta}} f(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|^{\alpha}} < \infty \text{ for all multi-index } \boldsymbol{\beta} \text{ with } |\boldsymbol{\beta}| = k \right\}.$$

Remark 1.5. If $\alpha = 1$ and $k = 0$, then $\mathcal{C}^{0,1}(\overline{\Omega})$ corresponds to the space of Lipschitz continuous functions. For $\alpha = 0$ and $k = 0$, we recover the space of bounded continuous functions.

Remark 1.6. A domain $\Omega \subset \mathbb{R}^d$ is said to have a $\mathcal{C}^{k,\alpha}$ boundary if, near every boundary point, the domain can be locally represented as the graph of a $\mathcal{C}^{k,\alpha}$ function. Intuitively, when you zoom in near any boundary point, the boundary looks smooth up to order k with Hölder-continuous derivatives of order k, and the domain lies entirely on one side of this surface.. For example:

- The set $\Omega = \{(x,y) \in \mathbb{R}^2 : y^2 > x\}$ has a $\mathcal{C}^{1,1}$ boundary, since near each boundary point the curve can be written as $y = \pm\sqrt{x}$, whose first derivatives are Lipschitz continuous.
- In contrast, $\Omega = \{(x,y) \in \mathbb{R}^2 : y > |x|\}$ does *not* have a $\mathcal{C}^1$ boundary, because the boundary $y = |x|$ has a corner at the origin where the derivative does not exist.

Theorem 1.7 (**Existence and Uniqueness of Solution** [17, Theorem 6.14]). *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with $\mathcal{C}^{2,\alpha}$ boundary. Consider the elliptic boundary value problem:*

$$\begin{aligned} -\varepsilon \Delta u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \Gamma. \end{aligned}$$

If $f \in \mathcal{C}^{0,\alpha}(\overline{\Omega})$ and $g \in \mathcal{C}^{2,\alpha}(\overline{\Omega})$, then there exists a unique solution $u \in \mathcal{C}^{2,\alpha}(\overline{\Omega})$.

Let us now consider some examples of the Poisson equation and their solutions:

1. On $\Omega = [0,1] \times [0,1]$, with $g = 0$ and

$$f(x,y) = 2\pi^2 \sin(\pi x) \sin(\pi y),$$

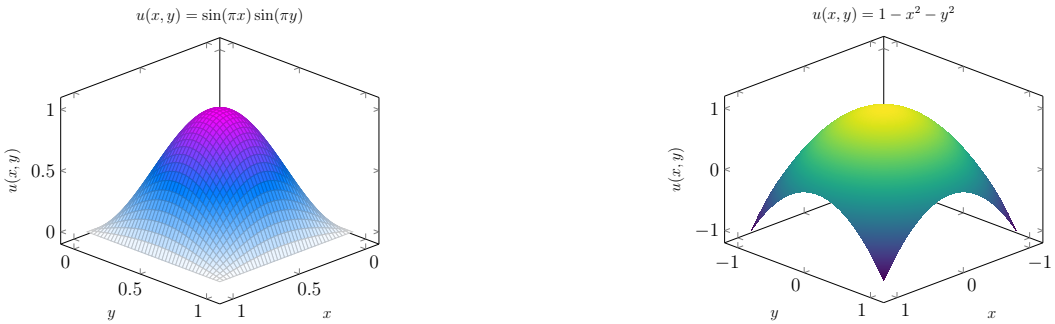
the exact solution (see Fig. 1.7a) is

$$u(x, y) = \sin(\pi x) \sin(\pi y).$$

2. On $\Omega = \{(x, y) : x^2 + y^2 < 1\}$ (the unit disk), with $g = 0$ and $f(x, y) = 4$, the solution (see Fig. 1.7b) in polar coordinates (r, θ) is

$$u(r, \theta) = 1 - r^2.$$

3. On $\Omega = [-1, 1]^2 \setminus [0, 1] \times [-1, 0]$, with $g = 0$ and $f(x, y) = 1$, the problem does not admit a classical solution. This is due to the re-entrant corner at $(0, 0)$ where the domain fails to have a $\mathcal{C}^{2,\alpha}$ boundary.



(a) Solution on $[0, 1]^2$.

(b) Solution on the unit disk.

Figure 1.7: Solutions to the Poisson equation in two simple domains.

The first two examples can be solved analytically using the *method of separation of variables*. In general, if a PDE admits a unique classical solution and the domain is regular (e.g., a rectangle or a disk), then separation of variables can often be applied. However, this approach fails for irregular domains such as that shown in Figure 1.8.

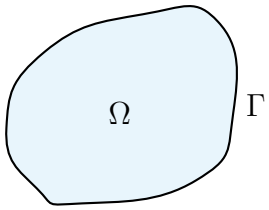


Figure 1.8: A domain Ω with smooth, curved boundary Γ.

We observe that even for relatively simple domains, classical solutions may fail to exist. For more complex domains, even when existence and uniqueness are guaranteed theoretically, computing the solution explicitly can become intractable.

This leads us to two major motivations:

1. To reconsider and generalize the notion of a "solution" (e.g., to weak solutions).

2. To employ numerical methods to approximate solutions.

The next chapter will focus on Point 2: numerical methods. Point 1 will be revisited later when we discuss the variational and weak formulations of PDEs.

Chapter 2

Finite Difference Methods

In the previous chapter, we provided a general overview of partial differential equations (PDEs), including their physical motivation, classification, boundary conditions, and the question of existence and uniqueness of classical solutions. We saw that analytical solutions are often limited to highly idealized cases, and that even simple domains can present significant challenges. This motivates the use of numerical methods, which will be the focus of this chapter.

We now turn our attention to the numerical solution of PDEs, specifically the Poisson equation, which serves as a prototype for elliptic problems. This chapter will build on ideas from the course `MA637: Numerical Analysis and Computing`, where we studied boundary value problems in one dimension. In particular, we considered equations of the form

$$-y''(x) = f\left(x, y(x), y'(x), y''(x)\right), \quad \text{for } a \leq x \leq b, \tag{2.1}$$

with Dirichlet boundary conditions $y(a) = \alpha$ and $y(b) = \beta$ (see [15, Chapter 11]).

The methods developed for such one-dimensional problems extend naturally to two-dimensional settings. By the end of this chapter, we aim to understand how to discretize elliptic PDEs using finite difference methods, how to incorporate various types of boundary conditions, and how to analyze the resulting schemes in terms of consistency, stability, and convergence. We will also address practical aspects of solving the resulting linear systems efficiently.

2.1 Dirichlet Problem in Square Domain

We begin our study of numerical methods with the Poisson equation subject to Dirichlet boundary conditions. The goal is to find $u \in \mathcal{C}^4(\Omega)$ such that

$$-\Delta u = f \quad \text{in } \Omega, \tag{2.2}$$
$$u = g \quad \text{on } \Gamma, \tag{2.3}$$

where $\Omega \subset \mathbb{R}^2$ is a domain with boundary $\Gamma := \partial\Omega$, and $f \in \mathcal{C}(\Omega)$, $g \in \mathcal{C}(\Gamma)$. We begin by considering the case $\Omega = (0,1) \times (0,1)$, which allows us to adapt ideas from finite difference methods (FDM) for boundary value problems in one dimension, similar to Eq. (2.1).

The first step in solving the Poisson equation is to discretize the domain Ω using a uniform grid of size $h \times h$ where $h = 1/M$ for some integer M. The internal grid points are defined by

$$\Omega_h = \{(x_i, y_j) : x_i = ih,\ y_j = jh,\ i, j = 1, 2, \ldots, M-1\}.$$

We denote the boundary grid points by $\partial\Omega_h$ and the complete set of grid points by $\overline{\Omega}_h = \Omega_h \cup \partial\Omega_h$ (see Fig. 2.1).

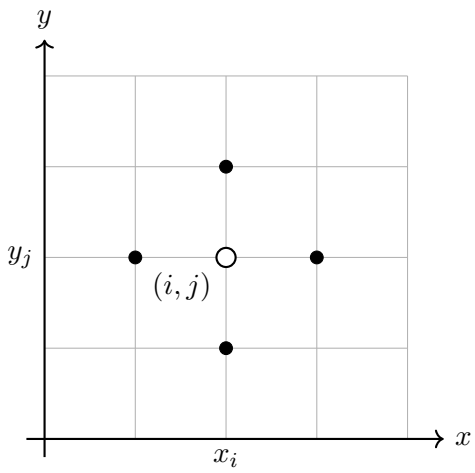


Figure 2.1: A uniform grid on the domain $\Omega = (0, 1) \times (0, 1)$. The white point represents an interior grid node (i, j), and the black points are its four nearest neighbors used in the five-point stencil.

Definition 2.1 (**Grid Function**). A vector $\mathbf{u} \in \mathbb{R}^{(M+1)\times(M+1)}$ that assigns a function value to each grid point is called a *grid function*.

For a grid point $(x_i, y_j) \in \Omega_h$, where $x_i = ih$ and $y_j = jh$, we denote the numerical approximation of the solution by $\mathbf{u}_{i,j}$, and the exact analytical solution by $u_{i,j} := u(x_i, y_j)$. While the boundary values of u are known from the prescribed Dirichlet condition, the values at interior points are computed by solving a system of equations.

Remark 2.2. Whenever we use bold notation such as $\mathbf{u}$ or $\mathbf{f}$, it refers to the vector representation over the grid. The subscripted form $\mathbf{u}_{i,j}$ refers to the scalar value at the grid point (i, j).

Returning to Eq. (2.2), we recall that the Laplacian is given by

$$\Delta u = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right).$$

We use a Taylor expansion to approximate the second derivatives. At the grid point (x_i, y_j), we have

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{i,j} = \frac{1}{h^2}\left(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}\right) + \mathcal{O}(h^2).$$

Letting $\delta_x^2 u_{i,j} := u_{i-1,j} - 2u_{i,j} + u_{i+1,j}$, we get

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{i,j} = \frac{\delta_x^2 u_{i,j}}{h^2} + \mathcal{O}(h^2).$$

Similarly, in the y-direction:

$$\left.\frac{\partial^2 u}{\partial y^2}\right|_{i,j} = \frac{\delta_y^2 u_{i,j}}{h^2} + \mathcal{O}(h^2), \quad \text{where } \delta_y^2 u_{i,j} := u_{i,j-1} - 2u_{i,j} + u_{i,j+1}.$$

This gives the approximation:

$$-\Delta u\big|_{i,j} \approx -\frac{1}{h^2}\left(\delta_x^2 u_{i,j} + \delta_y^2 u_{i,j}\right) + \mathcal{O}(h^2).$$

The finite difference equations that approximate the PDE are obtained by ignoring the truncation error and replacing $u_{i,j}$ with the grid function value $\mathbf{u}_{i,j}$. This leads to the algebraic equation

$$-\frac{1}{h^2}\left(\delta_x^2 \mathbf{u}_{i,j} + \delta_y^2 \mathbf{u}_{i,j}\right) = f_{i,j},$$

where $f_{i,j} := f(x_i, y_j)$. Substituting the definitions of δ_x^2 and δ_y^2, we get:

$$4\mathbf{u}_{i,j} - \mathbf{u}_{i+1,j} - \mathbf{u}_{i-1,j} - \mathbf{u}_{i,j+1} - \mathbf{u}_{i,j-1} = h^2 f_{i,j}, \tag{2.4}$$

for each $(x_i, y_j) \in \Omega_h$.

Since the values of $\mathbf{u}$ at the boundary grid points are known from the Dirichlet condition, we obtain a system of $(M-1)^2$ linear equations for the unknowns at the interior points. Each equation involves the grid point and its four immediate neighbors, hence this is referred to as the *five-point stencil*.

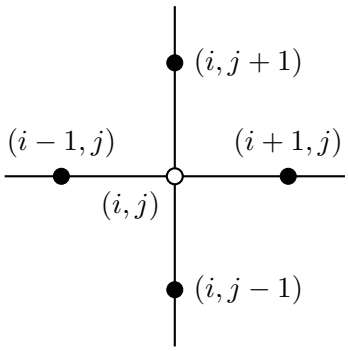


Figure 2.2: Five-point stencil: the central white node (i, j) is surrounded by its four nearest neighbors $(i \pm 1, j)$ and $(i, j \pm 1)$.

In the special case where $f \equiv 0$ (i.e., Laplace's equation), Eq. (2.4) simplifies to

$$\mathbf{u}_{i,j} = \frac{1}{4}\left(\mathbf{u}_{i+1,j} + \mathbf{u}_{i-1,j} + \mathbf{u}_{i,j+1} + \mathbf{u}_{i,j-1}\right).$$

That is, the value at an interior grid point is the arithmetic mean of its four immediate neighbors (see Fig. 2.2). This is reminiscent of the mean value property satisfied by harmonic functions in complex analysis. Remarkably, our discrete solution satisfies a similar property, highlighting the consistency of the method with the underlying theory.

2.1.1 Implementation of Five-Point Stencil

We now examine the organization of the unknowns $\mathbf{u}_{i,j}$, which are indexed by two indices. To convert them into a single column vector suitable for matrix-based computations, we arrange the interior grid values into a matrix and then stack the columns. This yields

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{M-1} \end{bmatrix}^\top,$$

where each $\mathbf{u}_i \in \mathbb{R}^{M-1}$ represents a column vector of values along the vertical grid line x_i, i.e.,

$$\mathbf{u}_i = \begin{bmatrix} \mathbf{u}_{i,1} & \mathbf{u}_{i,2} & \cdots & \mathbf{u}_{i,M-1} \end{bmatrix}^\top.$$

Next, consider the matrix structure corresponding to the finite difference equation (2.4). Each row in the matrix corresponds to a grid point $(i, j) \in \Omega_h$ and involves at most five non-zero entries due to the five-point stencil. At boundary-adjacent interior nodes, some terms involve known values from the Dirichlet boundary condition and are moved to the right-hand side.

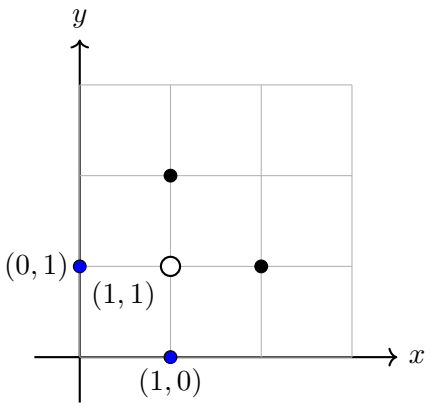


Figure 2.3: Stencil at the bottom-left corner of the grid. The interior node $(1, 1)$ (white) depends on neighboring nodes. The Dirichlet boundary values at $(0, 1)$ and $(1, 0)$ (blue) are known from boundary conditions, while the others are unknown.

For instance, consider $(i, j) = (1, 1)$. The finite difference equation is

$$\begin{aligned} 4\mathbf{u}_{1,1} - \mathbf{u}_{2,1} - \mathbf{u}_{1,2} - \mathbf{u}_{0,1} - \mathbf{u}_{1,0} &= h^2 f_{1,1}, \\ \Rightarrow \quad 4\mathbf{u}_{1,1} - \mathbf{u}_{2,1} - \mathbf{u}_{1,2} &= h^2 f_{1,1} + g_{0,1} + g_{1,0}, \end{aligned}$$

where $\mathbf{u}_{0,1} = g_{0,1}$ and $\mathbf{u}_{1,0} = g_{1,0}$ are known from the Dirichlet data. Hence, only three unknowns remain on the left-hand side.

We now focus on the matrix representation column by column. Fix a vertical index i and consider the values $\mathbf{u}_{i,j}$ for $j = 1, 2, \ldots, M-1$. Each $\mathbf{u}_i$ is influenced by its neighbors $\mathbf{u}_{i-1}$ and $\mathbf{u}_{i+1}$, as shown below.

For $j = 1$:

$$\frac{1}{h^2}\left(4\mathbf{u}_{i,1} - \mathbf{u}_{i-1,1} - \mathbf{u}_{i+1,1} - \mathbf{u}_{i,2} - \mathbf{u}_{i,0}\right) = f_{i,1}.$$

Since $\mathbf{u}_{i,0} = g_{i,0}$ is known, this becomes:

$$4\mathbf{u}_{i,1} - \mathbf{u}_{i-1,1} - \mathbf{u}_{i+1,1} - \mathbf{u}_{i,2} = h^2 f_{i,1} + g_{i,0}.$$

For $j = 2$:

$$4\mathbf{u}_{i,2} - \mathbf{u}_{i-1,2} - \mathbf{u}_{i+1,2} - \mathbf{u}_{i,1} - \mathbf{u}_{i,3} = h^2 f_{i,2}.$$

For $j = M - 1$:

$$4\mathbf{u}_{i,M-1} - \mathbf{u}_{i-1,M-1} - \mathbf{u}_{i+1,M-1} - \mathbf{u}_{i,M-2} = h^2 f_{i,M-1} + g_{i,M},$$

where $\mathbf{u}_{i,M} = g_{i,M}$ is again known from the boundary.

Writing these equations in matrix form for fixed i, we observe a tridiagonal structure:

$$\begin{aligned}
\begin{bmatrix}-1\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i-1,1}\end{bmatrix} + \begin{bmatrix}4 & -1\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i,1}\\ \mathbf{u}_{i,2}\end{bmatrix} + \begin{bmatrix}-1\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i+1,1}\end{bmatrix} &= h^2 f_{i,1} + g_{i,0},\\
\begin{bmatrix}-1\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i-1,2}\end{bmatrix} + \begin{bmatrix}-1 & 4 & -1\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i,1}\\ \mathbf{u}_{i,2}\\ \mathbf{u}_{i,3}\end{bmatrix} + \begin{bmatrix}-1\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i+1,2}\end{bmatrix} &= h^2 f_{i,2},\\
&\vdots\\
\begin{bmatrix}-1 & 4\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i,M-2}\\ \mathbf{u}_{i,M-1}\end{bmatrix} + \begin{bmatrix}-1\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i+1,M-1}\end{bmatrix} &= h^2 f_{i,M-1} + g_{i,M}.
\end{aligned}$$

Each column vector $\mathbf{u}_i$ satisfies a linear system involving a tridiagonal matrix $\mathbf{D} \in \mathbb{R}^{(M-1)\times(M-1)}$:

$$\mathbf{D} = \begin{bmatrix} 4 & -1 & 0 & \cdots & 0\\ -1 & 4 & -1 & \cdots & 0\\ 0 & -1 & 4 & \cdots & 0\\ \vdots & \vdots & \ddots & \ddots & -1\\ 0 & 0 & \cdots & -1 & 4\end{bmatrix}. \tag{2.5}$$

Each $\mathbf{u}_i$ is also coupled with neighboring columns $\mathbf{u}_{i-1}$ and $\mathbf{u}_{i+1}$ through identity matrices (denoted by $\mathbb{I}$). The full system for $i = 1, 2, \ldots, M-1$ becomes:

$$-\mathbb{I}\mathbf{u}_{i-1} + \mathbf{D}\mathbf{u}_i - \mathbb{I}\mathbf{u}_{i+1} = h^2\mathbf{f}_i + \mathbf{g}_i,$$

where

$$\mathbf{f}_i = \begin{bmatrix} f_{i,1} & f_{i,2} & \cdots & f_{i,M-1}\end{bmatrix}^\top, \quad \text{and} \quad \mathbf{g}_i = \begin{bmatrix} g_{i,0} & 0 & \cdots & 0 & g_{i,M}\end{bmatrix}^\top.$$

The values $\mathbf{u}_0 = \mathbf{g}_0$ and $\mathbf{u}_M = \mathbf{g}_M$ correspond to the Dirichlet boundary and are moved to the right-hand side.

Stacking the equations for all $i = 1, \ldots, M-1$, we obtain a linear system of the form

$$\mathbf{A}\mathbf{u} = \mathbf{f},$$

where $\mathbf{A} \in \mathbb{R}^{(M-1)^2\times(M-1)^2}$, and both $\mathbf{u}$ and $\mathbf{f}$ are vectors in $\mathbb{R}^{(M-1)^2}$. The matrix $\mathbf{A}$ has a block tridiagonal structure given by

$$\mathbf{A} = \frac{1}{h^2}\begin{bmatrix} \mathbf{D} & -\mathbb{I} & & \\ -\mathbb{I} & \mathbf{D} & \ddots & \\ & \ddots & \ddots & -\mathbb{I}\\ & & -\mathbb{I} & \mathbf{D}\end{bmatrix}.$$

The right-hand side takes the form

$$\mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_{M-1} \end{bmatrix} + \frac{1}{h^2} \begin{bmatrix} \mathbf{g}_0 + \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_{M-1} + \mathbf{g}_M \end{bmatrix},$$

which accounts for the contributions from the boundary conditions.

Let us now examine the structure of the matrix $\mathbf{A}$ in more detail. Recall that $\mathbf{A} \in \mathbb{R}^{(M-1)^2 \times (M-1)^2}$, so it contains approximately M^4 entries in total. However, if we inspect the structure of $\mathbf{A}$, particularly the second block row, we find that each block row contains roughly $5M$ non-zero entries—due to contributions from the main diagonal block and the two neighboring identity blocks. Consequently, the total number of non-zero entries in $\mathbf{A}$ is approximately $5M^2$.

This indicates that the matrix $\mathbf{A}$ is *sparse*, meaning that the majority of its entries are zero. Fig. 2.4 illustrates the sparsity pattern of $\mathbf{A}$ for $M = 10$, with non-zero entries shown in black.

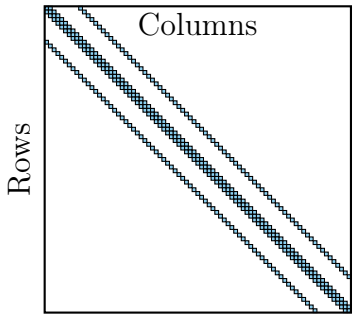


Figure 2.4: Sparsity pattern for matrix $\mathbf{A}$ for $M = 10$.

Definition 2.3 (**Sparse Matrix**). A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is called a *sparse matrix* if a significant fraction of its entries are zero. The *sparsity* of a matrix is typically quantified as the ratio

$$\text{sparsity}(\mathbf{A}) = 1 - \frac{\text{number of non-zero entries in } \mathbf{A}}{n^2}.$$

Solving linear systems involving sparse matrices requires methods that exploit their structure to reduce computational cost and memory usage. Among the most widely used direct methods is LU decomposition, which factorizes $\mathbf{A}$ into a product of a lower triangular matrix $\mathbf{L}$ and an upper triangular matrix $\mathbf{U}$.

Remark 2.4. We might revisit sparse matrix storage schemes and efficient solution techniques—both direct and iterative—later in the course when we focus on linear solvers.

Let us now look at the implementation of the matrix $\mathbf{A}$ and the right-hand side vector $\mathbf{f}$ for the finite difference method (FDM).

Definition 2.5 (Kronecker Product). Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$. The *Kronecker product* $\mathbf{A} \otimes \mathbf{B}$ is defined as the block matrix:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \in \mathbb{R}^{mp\times nq}.$$

In component form, this means that each scalar entry a_{ij} of matrix $\mathbf{A}$ is multiplied by the entire matrix $\mathbf{B}$. For example, if $\mathbf{A}$ is 2×2 and $\mathbf{B}$ is 2×2, then:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

then

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{bmatrix}.$$

For the 2D finite difference discretization of the Poisson equation, the global stiffness matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2,$$

where

$$\mathbf{A}_1 = \mathbb{I} \otimes \mathbf{D}, \quad \text{and} \quad \mathbf{A}_2 = \bar{\mathbb{I}} \otimes \mathbb{I}.$$

Here $\bar{\mathbb{I}}$ is the tridiagonal matrix with zero diagonal and -1 on its first sub- and super-diagonals:

$$\bar{\mathbb{I}} = \begin{bmatrix} 0 & -1 & 0 & \cdots & 0 \\ -1 & 0 & -1 & \cdots & 0 \\ 0 & -1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 0 \end{bmatrix}.$$

Thus, the global matrix becomes:

$$\mathbf{A} = \mathbb{I} \otimes \mathbf{D} + \bar{\mathbb{I}} \otimes \mathbb{I}.$$

In Python, the Kronecker products can be constructed using the function `np.kron`$(\mathbf{A}, \mathbf{B})$ from the NumPy library. Algorithm 1 presents a basic implementation of this method for the 2D Poisson equation with Dirichlet boundary conditions.

Algorithm 1 Finite Difference Method for 2D Poisson Equation with Dirichlet BCs

Given: Domain $\Omega = (0,1)^2$, Number of nodes per dimension: M, right-hand side: $f(x,y)$, boundary data: $g(x,y)$
Find: Approximate solution $\mathbf{u}$ on interior grid

Step 1: Meshing
Initialize $h = \frac{1}{M}$
Define interior grid points $x_i = ih$, $y_j = jh$, for $i, j = 1, 2, \ldots, M-1$
Total number of interior nodes $N = M - 1$

Step 2: Assemble matrix A
Define matrix $\mathbf{D}, \bar{\mathbb{I}} \in \mathbb{R}^{N \times N}$ and by
for $i = 0$ **to** $N - 1$ **do**
  $\mathbf{D}_{i,i} = 4$, $\bar{\mathbb{I}}_{i,i} = 0$
  if $i > 0$ **then**
    $\mathbf{D}_{i,i-1} = -1$, $\bar{\mathbb{I}}_{i,i-1} = -1$
  end if
  if $i < N - 1$ **then**
    $\mathbf{D}_{i,i+1} = -1$, $\bar{\mathbb{I}}_{i,i+1} = -1$
  end if
end for
Define identity matrix $\mathbb{I} \in \mathbb{R}^{N \times N}$
Define $\mathbf{A} = \left(\mathbb{I} \otimes \mathbf{D} + \bar{\mathbb{I}} \otimes \mathbb{I}\right)/h^2$

Step 3: Assemble the RHS
Initialize load vector $\mathbf{F} \in \mathbb{R}^{N^2}$
for $i = 0$ **to** $N - 1$ **do**
  for $j = 0$ **to** $N - 1$ **do**
    Compute grid point: $x = (i+1)h$, $y = (j+1)h$
    Compute linear index: $\ell = i \cdot N + j$
    $\mathbf{F}_\ell = f(x,y)$
    Apply Dirichlet BCs:
    if $i = 0$ **then**
      $\mathbf{F}_\ell += g(0,y)/h^2$
    end if
    if $i = N - 1$ **then**
      $\mathbf{F}_\ell += g(1,y)/h^2$
    end if
    if $j = 0$ **then**
      $\mathbf{F}_\ell += g(x,0)/h^2$
    end if
    if $j = N - 1$ **then**
      $\mathbf{F}_\ell += g(x,1)/h^2$
    end if
  end for
end for

Step 4: Solve the linear system
Solve $\mathbf{Au} = \mathbf{F}$ for $\mathbf{u} \in \mathbb{R}^{N^2}$

---

return u

2.1.2 Convergence Theory

Implementation of a numerical scheme is one aspect that primarily concerns engineers. Another important aspect, which concerns mathematicians, is the study of its convergence properties. We say a numerical solution $\mathbf{u}$ is a good approximation of the analytical solution u if the error between them reduces as the discretization parameter h tends to zero. One important question is how to measure the closeness of $\mathbf{u}$ and u, as $\mathbf{u}$ consists of discrete values, whereas u is a continuous function.

We re-write Eq. (2.2) as $\mathcal{L}u = \mathcal{F}$, where

$$\mathcal{L}u = \begin{cases} -\Delta u & \text{in } \Omega, \\ u & \text{on } \Gamma, \end{cases} \quad \text{and} \quad \mathcal{F} = \begin{cases} f & \text{in } \Omega, \\ g & \text{on } \Gamma. \end{cases} \tag{2.6}$$

We define the five-point stencil operator by $\mathcal{L}_h\mathbf{u} = \mathcal{F}_h$, where

$$\mathcal{L}_h\mathbf{u} = \begin{cases} \mathfrak{L}_h\mathbf{u} & \text{in } \Omega_h, \\ \mathfrak{B}_h\mathbf{u} & \text{on } \Gamma_h, \end{cases} \quad \text{and} \quad \mathcal{F}_h = \begin{cases} \mathbf{f} & \text{in } \Omega_h, \\ \mathbf{g} & \text{on } \Gamma_h. \end{cases} \tag{2.7}$$

Here, $\mathfrak{L}_h\mathbf{u}_{i,j} := -h^{-2}\left(\delta_x^2 + \delta_y^2\right)\mathbf{u}_{i,j}$ and $\mathfrak{B}_h\mathbf{u}_{i,j} := \mathbf{u}_{i,j}$, i.e., the identity operator.

The operator $\mathcal{L}_h$ approximates the continuous operator $\mathcal{L}$. To quantify how well it does so, we introduce the notion of *local truncation error*.

Definition 2.6. (**Local Truncation Error**) The local truncation error, denoted by $\mathfrak{R}_h$, is defined as the residual when the exact solution u of Eq. (2.6) is substituted into the discrete equation Eq. (2.7), i.e.,

$$\mathfrak{R}_h := \mathcal{L}_h u - \mathcal{F}_h.$$

Definition 2.7. (**Consistency**) The approximation $\mathcal{L}_h\mathbf{u} = \mathcal{F}_h$ is said to be *consistent* with $\mathcal{L}u = \mathcal{F}$ if $\mathfrak{R}_h \to 0$ as $h \to 0$. It is said to be consistent of order p if $\mathfrak{R}_h = \mathcal{O}(h^p)$ with $p > 0$.

The first step is to verify whether the FDM is consistent. Since Dirichlet boundary conditions are implemented exactly, we only need to analyze $\mathfrak{L}_h\mathbf{u}$. Thus, the local truncation error reduces to $\mathfrak{R}_h = \mathfrak{L}_h u - \mathbf{f}$. Using Taylor series expansion for $\delta_x^2 u_{i,j}$ and $\delta_y^2 u_{i,j}$ at the internal node (ih, jh), we get

$$\begin{aligned} \mathfrak{R}_h|_{i,j} &= -h^{-2}\left(\delta_x^2 + \delta_y^2\right)u_{i,j} - \mathbf{f}_{i,j} \\ &= -\left(u_{xx}|_{i,j} + u_{yy}|_{i,j}\right) - \mathbf{f}_{i,j} - \frac{h^2}{12}\left(u_{xxxx} + u_{yyyy}\right)|_{i,j} + \mathcal{O}(h^4) \\ &= -\frac{h^2}{12}\left(u_{xxxx} + u_{yyyy}\right)|_{i,j} + \mathcal{O}(h^4), \end{aligned}$$

since $-\Delta u = f$. Hence, the method is consistent of order 2, provided $u \in \mathcal{C}^4(\Omega)$.

Consistency tells us how closely $\mathcal{L}_h$ approximates $\mathcal{L}$. The numerical solution $\mathbf{u}$ satisfies $\mathcal{L}_h\mathbf{u} = \mathcal{F}_h$, while the analytical solution u satisfies $\mathcal{L}_h u = \mathcal{F}_h + \mathfrak{R}_h$. To study their difference, let $\mathbf{e} = u - \mathbf{u}$, where $\mathbf{e}_{i,j} = u(x_i, y_j) - \mathbf{u}_{i,j}$. Then,

$$\mathcal{L}_h\mathbf{e} = \mathcal{L}_h\mathbf{u} - \mathcal{L}_h u = \mathfrak{R}_h.$$

We want the error $\mathbf{e}$ to go to zero as $h \to 0$. This happens if $\mathfrak{R}_h \to 0$ and $\mathcal{L}_h$ is invertible. This leads to the notion of *stability*.

Definition 2.8. (ℓ_∞ **Stability**) The discrete operator is said to be stable (with respect to the maximum norm $\|\cdot\|_{h,\infty}$) if there exists a constant $C > 0$, independent of h, such that the solution of the equation $\mathcal{L}_h\mathbf{u} = \mathcal{F}_h$ satisfies

$$\|\mathbf{u}\|_{h,\infty} \leq C\|\mathcal{F}_h\|_{h,\infty},$$

where C is known as the *stability constant*.

A few things to note: first, the stability of $\mathcal{L}_h$ makes no reference to the original operator $\mathcal{L}$; second, stability depends on the particular norm, in this case the maximum norm $\|\cdot\|_{h,\infty}$. This notion extends to other norms such as ℓ_p norms for $1 \leq p < \infty$. As a refresher, the ℓ_∞ norm is defined as

$$\|u - \mathbf{u}\|_{h,\infty} = \max_{i,j} |u(x_i, y_j) - \mathbf{u}_{i,j}|.$$

As mentioned in the introduction, this is how we measure the closeness of the numerical solution.

Remark 2.9. The notion of ℓ_∞ stability mirrors that of a well-posed problem. A boundary value problem with a unique solution is said to be well-posed if the solution varies continuously with respect to the input data. That is, if δf is a perturbation in the source or boundary terms and δu is the resulting change in the solution, then $\|\delta u\|_a \leq C\|\delta f\|_a$ for some norm $\|\cdot\|_a$.

But why do we need both consistency and stability? To answer that, we now formally define convergence.

Definition 2.10. (**Convergence**) A numerical method is said to *converge* if $\|u - \mathbf{u}\|_{h,\infty} \to 0$ as $h \to 0$. It is said to be convergent of order p if $\|u - \mathbf{u}\|_{h,\infty} = \mathcal{O}(h^p)$ for some $p > 0$.

We introduced consistency and stability first because there is a fundamental relationship among the three:

Theorem 2.11. (**Convergence**) *Suppose that the discrete boundary value problem $\mathcal{L}_h\mathbf{u} = \mathcal{F}_h$ is a consistent approximation of $\mathcal{L}u = \mathcal{F}$ and that $\mathcal{L}_h$ is ℓ_∞ stable. Then $\|u - \mathbf{u}\|_{h,\infty} \to 0$ as $h \to 0$. Moreover, if the order of consistency is $p > 0$, then the order of convergence is also p.*

Proof. Since $\mathcal{L}_h$ is ℓ_∞ stable and the error $\mathbf{e}$ satisfies

$$\mathcal{L}_h\mathbf{e} = \mathfrak{R}_h,$$

we have $\|\mathbf{e}\|_{h,\infty} \leq C\|\mathfrak{R}_h\|_{h,\infty}$. Because $\mathcal{L}_h$ is consistent of order p, we conclude that $\|\mathbf{e}\|_{h,\infty} \to 0$ as $h \to 0$ with the same order, provided C is independent of h. □

We have already seen that our method is consistent, and hence, if it is stable, then it will also be convergent due to Theorem 2.11. Before we prove stability, we need to introduce a few concepts. We first mention certain properties of the operator $\mathcal{L}$.

- $\mathcal{L}$ is said to be linear if $\mathcal{L}(u + \alpha v) = \mathcal{L}(u) + \alpha\mathcal{L}(v)$ for some scalar α.
- $\mathcal{L}$ is said to be *inverse monotone* if $\mathcal{L}u \geq 0$ implies that $u \geq 0$.

To prove stability, we also require a bounded, non-negative function $\varphi(x)$, called a *comparison function*, such that $\mathcal{L}\varphi(x) \geq 1$ for all $x \in [0, 1]$.

Lemma 2.12. (**Stability**) *Suppose that the operator $\mathcal{L}_h$ is linear and inverse monotone, and that there exists a comparison function $\Phi > 0$ such that $\mathcal{L}_h\Phi \geq 1$. Then $\mathcal{L}_h$ is stable with stability constant $C = \max_{i,j} \Phi_{i,j}$, provided Φ is bounded independently of h.*

Proof. Let $\mathbf{u}$ be a solution of $\mathcal{L}_h\mathbf{u} = \mathcal{F}_h$. We need to show that $\|\mathbf{u}\|_{h,\infty} \leq C\|\mathcal{F}_h\|_{h,\infty}$. For a particular (i, j)-th entry:

$$\begin{aligned} \mathcal{L}_h\mathbf{u}_{i,j} = (\mathcal{F}_h)_{i,j} &\leq \|\mathcal{F}_h\|_{h,\infty} \\ &= \|\mathcal{F}_h\|_{h,\infty} \times 1 \\ &\leq \|\mathcal{F}_h\|_{h,\infty}\mathcal{L}_h\Phi_{i,j} \\ &= \mathcal{L}_h\left(\|\mathcal{F}_h\|_{h,\infty}\Phi\right)_{i,j}, \end{aligned}$$

which implies $0 \leq \mathcal{L}_h\left(\|\mathcal{F}_h\|_{h,\infty}\Phi - \mathbf{u}\right)_{i,j}$. Hence, by inverse monotonicity, $\mathbf{u}_{i,j} \leq \|\mathcal{F}_h\|_{h,\infty}\Phi_{i,j}$. Similarly, we obtain $-\mathbf{u}_{i,j} \leq \|\mathcal{F}_h\|_{h,\infty}\Phi_{i,j}$. Since $-\mathbf{u}_{i,j}, \mathbf{u}_{i,j} \leq |\mathbf{u}_{i,j}|$, we get $|\mathbf{u}_{i,j}| \leq \|\mathcal{F}_h\|_{h,\infty}\Phi_{i,j}$.

Now, taking the maximum over all i, j:

$$\|\mathbf{u}\|_{h,\infty} = \max_{i,j} |\mathbf{u}_{i,j}| \leq C\|\mathcal{F}_h\|_{h,\infty}, \quad \text{where } C = \max_{i,j} \Phi_{i,j}.$$

□

Remark 2.13. In Lemma 2.12, we consider a grid function $\mathbf{u}$. When we say $\mathbf{u} > 0$, it means $\mathbf{u}_{i,j} > 0$ for all $i, j = 1, 2, \ldots, M - 1$.

We now prove the stability of our FDM operator $\mathcal{L}_h$.

Theorem 2.14. *The difference operator $\mathcal{L}_h$ defined in Eq.* (2.7) *is stable.*

Proof. We need to show two things for $\mathcal{L}_h$ to be stable. First is inverse monotonicity, and second, we need to compute a comparison function Φ such that $\mathcal{L}_h\Phi \geq 1$.

Inverse Monotonicity: We need to show that $\mathcal{L}_h\mathbf{u} \geq 0$ implies $\mathbf{u} \geq 0$.

We proceed by contradiction. Suppose this is not true. Then there exists a grid function $\mathbf{u}$ such that $\mathcal{L}_h\mathbf{u} \geq 0$ but $\mathbf{u} < 0$, i.e., there exists a point (i, j) such that

$$\mathbf{u}_{i,j} < 0. \tag{2.8}$$

Let (i_0, j_0) be the point where the minimum occurs.

Now,

$$\mathcal{L}_h\mathbf{u}_{i_0,j_0} = \frac{1}{h^2}\left(4\mathbf{u}_{i_0,j_0} - \mathbf{u}_{i_0+1,j_0} - \mathbf{u}_{i_0-1,j_0} - \mathbf{u}_{i_0,j_0+1} - \mathbf{u}_{i_0,j_0-1}\right).$$

Since $\mathbf{u}_{i_0,j_0}$ is the minimum, we have $\mathbf{u}_{i_0,j_0} \leq \mathbf{u}_{i_0\pm1,j_0}$ and $\mathbf{u}_{i_0,j_0} \leq \mathbf{u}_{i_0,j_0\pm1}$, which implies

$$\mathbf{u}_{i_0,j_0} - \mathbf{u}_{i_0\pm1,j_0} \leq 0 \quad \text{and} \quad \mathbf{u}_{i_0,j_0} - \mathbf{u}_{i_0,j_0\pm1} \leq 0.$$

Hence,

$$\begin{aligned}\mathcal{L}_h\mathbf{u}_{i_0,j_0} &= \frac{1}{h^2}\Big(\mathbf{u}_{i_0,j_0} - \mathbf{u}_{i_0+1,j_0} + \mathbf{u}_{i_0,j_0} - \mathbf{u}_{i_0-1,j_0} + \mathbf{u}_{i_0,j_0} - \mathbf{u}_{i_0,j_0+1} + \mathbf{u}_{i_0,j_0} - \mathbf{u}_{i_0,j_0-1}\Big)\\ &\leq 0.\end{aligned}$$

But we also assumed that $\mathcal{L}_h\mathbf{u}_{i_0,j_0} \geq 0$. Therefore, we must have $\mathcal{L}_h\mathbf{u}_{i_0,j_0} = 0$, and all terms in the sum must be zero:

$$\mathbf{u}_{i_0,j_0} = \mathbf{u}_{i_0\pm1,j_0} = \mathbf{u}_{i_0,j_0\pm1}.$$

That is, the value at (i_0, j_0) equals the values at its four neighbors. Now, pick one of these neighbors, say $(i_0 + 1, j_0)$, and apply the same argument. Again, you will conclude that all its neighbors must have the same value. Repeating this process across the domain implies that $\mathbf{u}$ is constant on $\overline{\Omega}_h$.

Moreover, from (2.8), this constant is strictly negative: $\mathbf{u} < 0$ on $\overline{\Omega}_h$. But on the boundary $\partial\Omega_h$, we use Dirichlet conditions: $\mathcal{L}_h\mathbf{u} = \mathfrak{B}_h\mathbf{u} = \mathbf{u} = g$, and from $\mathcal{L}_h\mathbf{u} \geq 0$ this implies $g \geq 0$ on $\partial\Omega_h$. So we obtain a contradiction since $\mathbf{u} < 0$ on the boundary. Hence, our assumption is false, and we conclude that $\mathcal{L}_h$ is inverse monotone.

Comparison Function: Next, we need to find a comparison function $\Phi(x, y) > 0$ and show that $\mathcal{L}_h\Phi \geq 1$. Let us consider the function

$$\Phi(x, y) = 1 + \frac{1}{2}x(1 - x),$$

which is bounded and non-negative on $[0, 1]^2$ (see Figure 2.5).

Now, compute $\mathcal{L}_h\Phi_{i,j}$. Since Φ depends only on x, we can write:

$$\begin{aligned}\mathcal{L}_h\Phi_{i,j} &= -\frac{1}{h^2}\left(-4\Phi_{i,j} + \Phi_{i+1,j} + \Phi_{i-1,j} + \Phi_{i,j-1} + \Phi_{i,j+1}\right)\\ &= -\frac{1}{h^2}\Big(-4\left(1 + \frac{1}{2}x_i(1 - x_i)\right) + \left(1 + \frac{1}{2}x_{i+1}(1 - x_{i+1})\right)\\ &\quad + \left(1 + \frac{1}{2}x_{i-1}(1 - x_{i-1})\right) + \left(1 + \frac{1}{2}x_i(1 - x_i)\right) + \left(1 + \frac{1}{2}x_i(1 - x_i)\right)\Big).\end{aligned}$$

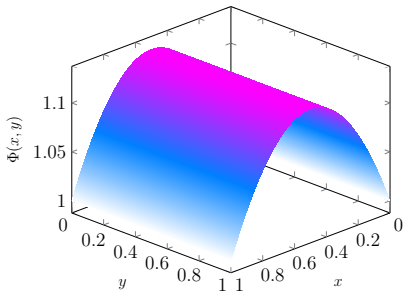


Figure 2.5: Surface plot of the comparison function $\Phi(x, y) = 1+\frac{1}{2}x(1-x)$, which is independent of y.

Using $x_{i\pm1} = x_i \pm h$, compute:

$$x_{i+1}(1 - x_{i+1}) = (x_i + h)(1 - x_i - h) = x_i(1 - x_i) - h(1 - 2x_i) - h^2,$$
$$x_{i-1}(1 - x_{i-1}) = (x_i - h)(1 - x_i + h) = x_i(1 - x_i) + h(1 - 2x_i) - h^2.$$

Adding both:

$$x_{i-1}(1 - x_{i-1}) + x_{i+1}(1 - x_{i+1}) = 2x_i(1 - x_i) - 2h^2.$$

Substitute this into the expression for $\mathcal{L}_h\Phi_{i,j}$:

$$\begin{aligned}\mathcal{L}_h\Phi_{i,j} &= -\frac{1}{h^2}\left(-4\left(1 + \frac{1}{2}x_i(1 - x_i)\right) + 2\left(1 + \frac{1}{2}x_i(1 - x_i)\right) + \frac{1}{2}(2x_i(1 - x_i) - 2h^2)\right) \\ &= -\frac{1}{h^2}\left(-2 - x_i(1 - x_i) + x_i(1 - x_i) - h^2\right) = 1.\end{aligned}$$

Hence, $\mathcal{L}_h\Phi = 1$, and the comparison function condition is satisfied.

Therefore, both required properties are fulfilled, and $\mathcal{L}_h$ is ℓ_∞ stable. □

We have the consistency and stability of our numerical method. Therefore we have the following result.

Corollary 2.15. *The five-point approximation to the Poisson equation with a Dirichlet boundary condition on the unit square, Ω, is second-order convergent if the fourth derivatives of the exact solution are bounded.*

Remark 2.16. (**Experimental Order of Convergence (E.O.C.)**) From theory, we expect second-order convergence for the five-point scheme. To verify this numerically, we define the order α of a method by

$$\|u - \mathbf{u}_h\|_{h,\infty} = \mathcal{O}(h^\alpha), \quad \text{for some } \alpha > 0, \tag{2.9}$$

where $\mathbf{u}_h$ denotes the solution obtained on a grid of width h. If we use a coarser grid of size $2h$, then

$$\|u - \mathbf{u}_{2h}\|_{2h,\infty} = \mathcal{O}(2^\alpha h^\alpha).$$

Therefore,

$$\frac{\|u - \mathbf{u}_h\|_{h,\infty}}{\|u - \mathbf{u}_{2h}\|_{2h,\infty}} \approx \frac{1}{2^\alpha}$$

which gives the estimate

$$\alpha \approx \frac{1}{\log(2)} \log\left(\frac{\|u - \mathbf{u}_{2h}\|_{2h,\infty}}{\|u - \mathbf{u}_h\|_{h,\infty}}\right).$$

This quantity can be computed directly from numerical results and is called the *experimental order of convergence* (E.O.C.).

Another way to estimate α is via a log-log plot. Taking the logarithm of Eq. (2.9) gives

$$\log\left(\|u - \mathbf{u}_h\|_{h,\infty}\right) \approx \alpha \log(h).$$

Defining $Y = \log\left(\|u - \mathbf{u}_h\|_{h,\infty}\right)$ and $X = \log(h)$, we obtain a linear relation $Y \approx \alpha X$. The slope of the Y vs. X curve (for varying h) then provides an estimate of α.

Example 2.17. Consider Eq. (2.2) with

$$f = 2\pi^2 \sin(\pi x)\cos(\pi y), \quad g = 0.$$

The analytical solution is

$$u(x, y) = \sin(\pi x)\cos(\pi y).$$

We compute the numerical solution using the five-point stencil and report the results in Table 2.1. We observe that the computed E.O.C. matches the theoretical prediction of second-order accuracy. We also plot the log-log plot in Fig. 2.6.

h	$\|u - \mathbf{u}\|_{h,\infty}$	**E.O.C.**
0.250000	0.053029	-
0.125000	0.012951	2.033754
0.062500	0.003219	2.008367
0.031250	0.000804	2.002087
0.015625	0.000201	2.000522
0.007812	0.000050	2.000130

Table 2.1: Example 2.17: Experimental order of convergence.

2.1.3 Discrete Maximum Principle

The analytical solution of Eq. (2.2) satisfies certain important properties. One such property is the *maximum principle*.

Theorem 2.18. (**Maximum Principle for the Poisson Problem**) *Suppose that the function $u(x, y)$ satisfies the inequality*

$$-u_{xx} - u_{yy} \leq 0,$$

for all $(x, y) \in \Omega$. Then $u(x, y)$ is either constant or attains its maximum value on $\partial\Omega$.

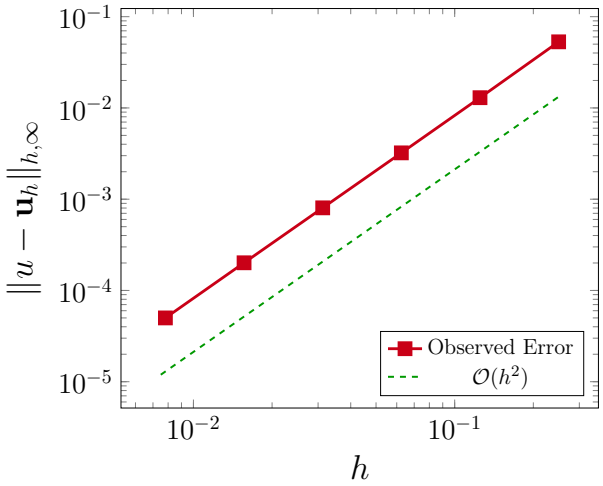


Figure 2.6: Log-log plot of error vs. mesh size h.

Remark 2.19. In some references (see [14]), maximum principles are distinguished as the *local maximum principle* and the *global maximum principle*. Here, we present a unified version encompassing both.

Remark 2.20. The maximum principle can also be understood from a physical point of view. Consider the steady-state heat distribution on a metal plate governed by the Poisson equation with no internal heat sources. Then, the maximum temperature must occur on the boundary of the plate. Intuitively, this means that heat cannot spontaneously concentrate in the interior if it is not being generated there, which aligns with our everyday physical experience.

The above theorem states that the maximum of the solution can occur only on the boundary. If it is attained in the interior, then the solution must be constant.

An important aspect in designing numerical methods is ensuring that the numerical solution also satisfies a discrete analogue of the analytical properties. In this case it will be the *discrete maximum principle*.

Theorem 2.21. (**Discrete Maximum Principle**) *Suppose that the discrete operator $\mathcal{L}_h\mathbf{u}$ is defined as in Eq. (2.7) and satisfies $\mathcal{L}_h\mathbf{u} \leq 0$ at all grid points in Ω_h. Then, either $\mathbf{u}$ is constant in Ω_h or it attains its maximum on the boundary $\partial\Omega_h$.*

Proof. We proceed by contradiction. The proof follows the same structure as the proof of ℓ_∞-stability of $\mathcal{L}_h$. Suppose $\mathbf{u}$ attains a non-constant maximum at an interior point (x_i, y_j), where $0 < i, j < M$. Then, since $\mathcal{L}_h\mathbf{u}_{i,j} \leq 0$, we have

$$\mathbf{u}_{i,j} \leq \frac{1}{4}\left(\mathbf{u}_{i+1,j} + \mathbf{u}_{i-1,j} + \mathbf{u}_{i,j+1} + \mathbf{u}_{i,j-1}\right).$$

But since $\mathbf{u}_{i,j}$ is the maximum, it satisfies

$$\mathbf{u}_{i,j} \geq \mathbf{u}_{i\pm1,j},\ \ \mathbf{u}_{i,j\pm1},$$

and hence

$$\mathbf{u}_{i,j} \geq \frac{1}{4}\left(\mathbf{u}_{i+1,j} + \mathbf{u}_{i-1,j} + \mathbf{u}_{i,j+1} + \mathbf{u}_{i,j-1}\right).$$

Combining both inequalities, we conclude

$$\mathbf{u}_{i,j} = \frac{1}{4}\left(\mathbf{u}_{i+1,j} + \mathbf{u}_{i-1,j} + \mathbf{u}_{i,j+1} + \mathbf{u}_{i,j-1}\right),$$

which can only happen if $\mathbf{u}_{i,j} = \mathbf{u}_{i\pm1,j} = \mathbf{u}_{i,j\pm1}$ as $\mathbf{u}_{i,j}$ is the maximum. That is, all five values are equal. Therefore, $\mathbf{u}$ attains the same maximum at the five-point stencil

$$\{(i,j),\ (i\pm1,j),\ (i,j\pm1)\}.$$

Repeating the same argument for the four neighboring points and propagating this reasoning through the domain, we find that either the solution is constant on $\overline{\Omega}_h$ or the maximum eventually appears on the boundary. This contradicts the assumption that the maximum occurs strictly in the interior. □

2.2 Advanced Topics

Until now, our focus has been on a simple problem, Eq. (2.2). A natural question that arises is: why solve Eq. (2.2) numerically when analytical solutions exist? One reason is to better understand the method, validate results computationally, and develop more advanced methods for problems where analytical solutions are not available. In this section, we explore such extensions.

2.2.1 Neumann Problem in Square Domain

Consider the Poisson problem with Neumann boundary conditions defined on the unit square $\Omega := (0,1)^2$,

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ \partial_{\mathbf{n}} u &= g \qquad \text{on } \Gamma, \end{aligned} \tag{2.10}$$

where $\partial_{\mathbf{n}} u := \nabla u \cdot \mathbf{n}$ denotes differentiation in the direction of the outward normal. From PDE theory, the solution to this problem is not well-posed unless f and g satisfy the compatibility condition

$$\int_\Omega f \, d\Omega + \int_\Gamma g \, ds = 0. \tag{2.11}$$

To develop a finite difference method for Eq. (2.10), we follow the same procedure as in the previous section. We will consider two distinct cases: first, the case $i = 0$ and $0 < j < M$ (inner boundary grid points); and second, the case $i = 0$ with $j = 0$ or $j = M$ (corner boundary grid points).

Inner Boundary Grid Points. Consider the grid point $(0, jh)$ for $0 < j < M$, i.e., a point lying on the boundary $x = 0$ (see Fig. 2.7).

Here, the outward normal is $\mathbf{n} = (-1, 0)$. Hence, the Neumann boundary condition becomes

$$\partial_{\mathbf{n}} u = (-1,0)\cdot(u_x, u_y) = -u_x = g(0,y), \quad \text{for } 0 < y < 1. \tag{2.12}$$

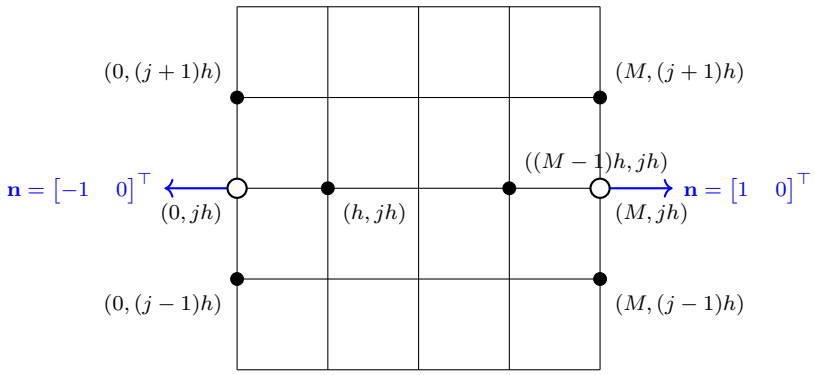


Figure 2.7: Stencil at the grid point $i = 0$ and $0 < j < M$. The interior node $(0, jh)$ (white) depends on neighboring nodes.

We recall the standard finite difference operators. Let Δ^+ denote the forward difference operator, and Δ^- the backward difference operator:

$$\Delta^+ u_i = u_{i+1} - u_i = hu_x|_{ih} + \frac{h^2}{2}u_{xx}|_{ih} + \mathcal{O}(h^3),$$
$$\Delta^- u_i = u_i - u_{i-1} = hu_x|_{ih} - \frac{h^2}{2}u_{xx}|_{ih} + \mathcal{O}(h^3).$$

Let Δ_x^+ denote the forward difference operator in the x-direction. Applying it to $u_{0,j}$, we get

$$\Delta_x^+ u_{0,j} = u_{1,j} - u_{0,j} = hu_x|_{0,j} + \frac{h^2}{2}u_{xx}|_{0,j} + \mathcal{O}(h^3). \tag{2.13}$$

Rewriting,

$$-u_x(0, jh) = -\frac{1}{h}\Delta_x^+ u_{0,j} + \frac{h}{2}u_{xx}|_{0,j} + \mathcal{O}(h^2). \tag{2.14}$$

Now, since the point $(0, jh)$ lies on the boundary, we cannot directly apply a second-order centered difference for u_{xx} (as it would require $u_{-1,j}$). However, assuming u is smooth, we use the PDE itself to substitute:

$$u_{xx}|_{0,j} = -f_{0,j} - u_{yy}|_{0,j} = -f_{0,j} - \frac{1}{h^2}\delta_y^2 u_{0,j} + \mathcal{O}(h^2). \tag{2.15}$$

Substituting this into the previous expression:

$$-u_x(0, jh) = -\frac{1}{h}\Delta_x^+ u_{0,j} - \frac{h}{2}f_{0,j} - \frac{1}{2h}\delta_y^2 u_{0,j} + \mathcal{O}(h^2). \tag{2.16}$$

Replacing u_x with the Neumann condition and dropping higher-order terms gives the discrete boundary condition:

$$-\frac{1}{h}\Delta_x^+ \mathbf{u}_{0,j} - \frac{1}{2h}\delta_y^2 \mathbf{u}_{0,j} = \frac{h}{2}\mathbf{f}_{0,j} + \mathbf{g}_{0,j}. \tag{2.17}$$

Expanding the operators, this becomes:

$$-2(\mathbf{u}_{1,j} - \mathbf{u}_{0,j}) + (2\mathbf{u}_{0,j} - \mathbf{u}_{0,j+1} - \mathbf{u}_{0,j-1}) = h^2\mathbf{f}_{0,j} + 2h\mathbf{g}_{0,j}, \tag{2.18}$$

$$\Rightarrow 4\mathbf{u}_{0,j} - 2\mathbf{u}_{1,j} - \mathbf{u}_{0,j+1} - \mathbf{u}_{0,j-1} = h^2\mathbf{f}_{0,j} + 2h\mathbf{g}_{0,j}. \tag{2.19}$$

The corresponding stencil is shown in Fig. 2.7. A similar strategy can be applied at the opposite boundary ($i = M$), using a backward difference in the x-direction instead. The resulting system is slightly modified and is left as an exercise.

Let us now consider the consistency error. The local truncation error at the boundary point $(0, jh)$ is given by

$$\mathfrak{R}_h|_{0,j} := -\frac{1}{h}\Delta_x^+\mathbf{u}_{0,j} - \frac{1}{2h}\delta_y^2\mathbf{u}_{0,j} - \mathbf{g}_{0,j} - \frac{h}{2}\mathbf{f}_{0,j}. \tag{2.20}$$

Using Taylor expansions, we get

$$\begin{aligned}\mathfrak{R}_h|_{0,j} &= -\left(u_x + \frac{h}{2}u_{xx} + \mathcal{O}(h^2)\right)_{0,j} - \frac{h}{2}\left(u_{yy} + \mathcal{O}(h^2)\right)_{0,j} - \mathbf{g}_{0,j} - \frac{h}{2}\mathbf{f}_{0,j} \\ &= (-u_x - \mathbf{g})_{0,j} + \frac{h}{2}\left(-u_{xx} - u_{yy} - \mathbf{f}\right)_{0,j} + \mathcal{O}(h^2).\end{aligned}$$

Since $u_x = -g$ and $-\Delta u = f$ at $(0, jh)$, the truncation error vanishes up to $\mathcal{O}(h^2)$, confirming second-order consistency.

Corner Boundary Grid Point

Now, let us examine the corner case. Consider the point $(0, 1)$, corresponding to $i = 0$ and $j = M$ (see Fig. 2.8). Along the edge $x = 0$, the Neumann condition is $-u_x = g$ since the outward normal is $\mathbf{n} = \begin{bmatrix}-1 & 0\end{bmatrix}^\top$. Along $y = 1$, the condition becomes $u_y = g$ since $\mathbf{n} = \begin{bmatrix}0 & 1\end{bmatrix}^\top$. At the corner, the normal derivative is not well-defined, so we use the average of the two directional conditions.

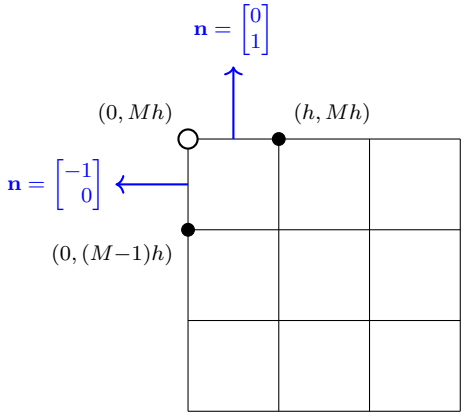


Figure 2.8: Stencil at the grid point $i = 0$ and $j = M$. The corner node $(0, Mh)$ (white) depends on neighboring nodes.

Let Δ_y^- denote the backward difference in the y-direction. Then,

$$\frac{1}{h}\Delta_y^- u_{0,M} = u_y|_{0,M} - \frac{h}{2}u_{yy}|_{0,M} + \mathcal{O}(h^2) \quad \Rightarrow \quad u_y|_{0,M} = \frac{1}{h}\Delta_y^- u_{0,M} + \frac{h}{2}u_{yy}|_{0,M} + \mathcal{O}(h^2). \tag{2.21}$$

Similarly, using the forward difference Δ_x^+ in the x-direction:

$$\frac{1}{h}\Delta_x^+ u_{0,M} = u_x|_{0,M} + \frac{h}{2}u_{xx}|_{0,M} + \mathcal{O}(h^2) \quad \Rightarrow \quad -u_x|_{0,M} = -\frac{1}{h}\Delta_x^+ u_{0,M} + \frac{h}{2}u_{xx}|_{0,M} + \mathcal{O}(h^2). \tag{2.22}$$

Averaging both expressions gives:

$$2g_{0,M} = \frac{1}{h}\left(\Delta_y^- u_{0,M} - \Delta_x^+ u_{0,M}\right) + \frac{h}{2}\left(u_{xx} + u_{yy}\right)_{0,M} + \mathcal{O}(h^2). \tag{2.23}$$

Replacing the second derivatives using the PDE $-\Delta u = f$ and dropping higher-order terms:

$$2\mathbf{g}_{0,M} = \frac{1}{h}\left(\mathbf{u}_{0,M} - \mathbf{u}_{0,M-1} - \mathbf{u}_{1,M} + \mathbf{u}_{0,M}\right) - \frac{h}{2}\mathbf{f}_{0,M}. \tag{2.24}$$

Simplifying,

$$2\mathbf{u}_{0,M} - \mathbf{u}_{0,M-1} - \mathbf{u}_{1,M} = \frac{h^2}{2}\mathbf{f}_{0,M} + 2h\mathbf{g}_{0,M}. \tag{2.25}$$

The local truncation error at the corner point is given by:

$$\begin{aligned}\mathfrak{R}_h|_{0,M} &= -\left(u_x + \frac{h}{2}u_{xx} + \mathcal{O}(h^2)\right)_{0,M} + \left(u_y - \frac{h}{2}u_{yy} + \mathcal{O}(h^2)\right)_{0,M} - \frac{h}{2}f_{0,M} - 2g_{0,M} \\ &= -\frac{h}{2}\left(u_{xx} + u_{yy} + f\right)_{0,M} + \left(u_y - g\right)_{0,M} - \left(u_x + g\right)_{0,M} + \mathcal{O}(h^2).\end{aligned}$$

Again, using the PDE and the Neumann boundary conditions at $(0, Mh)$, we conclude that the scheme is consistent with order $\mathcal{O}(h^2)$.

So far, we have not discussed the stability of the method. As seen in the previous section, the notion of stability is closely tied to the well-posedness of the underlying PDE. However, Eq. (2.10) does not have a unique solution: if u is a solution, then so is $u + c$ for any constant $c \in \mathbb{R}$.

Introducing a notion of stability without using function space theory is challenging. We may revisit this later from a more analytical perspective. In the meantime, it is worth pondering: with respect to which norm does the scheme exhibit stability?

Now, let us discuss the implementation of the method. One of the key differences from the Dirichlet case is that our number of unknowns increases, since the boundary values are also treated as unknowns. Thus, if

$$\mathbf{A}\mathbf{u} = \mathbf{f}$$

is the system of equations, then $\mathbf{A} \in \mathbb{R}^{(M+1)^2 \times (M+1)^2}$ and $\mathbf{u}, \mathbf{f} \in \mathbb{R}^{(M+1)^2}$.

For the interior grid points, the structure of the matrix remains the same. The additional rows and columns arise from the Neumann boundary conditions at $i = 0$, $i = M$, $j = 0$, and $j = M$. Looking at Figs. 2.7 and 2.8, we observe that:

- The unknowns at $i = 0$ depend on values at $i = 1$ (via forward difference).

– Similarly, values at $i = M$ depend on $i = M - 1$ (via backward difference).

Let the vector of unknowns at $i = 0$ be

$$\mathbf{u}_0 = \begin{bmatrix} \mathbf{u}_{0,0} & \mathbf{u}_{0,1} & \dots & \mathbf{u}_{0,M} \end{bmatrix}^\top,$$

and similarly define the right-hand side vector $\mathbf{f}_0$. The Neumann boundary stencil gives, for $j = 1, \dots, M-1$,

$$4\mathbf{u}_{0,j} - 2\mathbf{u}_{1,j} - \mathbf{u}_{0,j+1} - \mathbf{u}_{0,j-1} = h^2\mathbf{f}_{0,j} + 2h\mathbf{g}_{0,j}.$$

At the corners:

- For $(i, j) = (0, 0)$:

$$2\mathbf{u}_{0,0} - \mathbf{u}_{1,0} - \mathbf{u}_{0,1} = h^2/2\mathbf{f}_{0,0} + 2h\mathbf{g}_{0,0},$$

- For $(i, j) = (0, M)$:

$$2\mathbf{u}_{0,M} - \mathbf{u}_{1,M} - \mathbf{u}_{0,M-1} = h^2/2\mathbf{f}_{0,M} + 2h\mathbf{g}_{0,M}.$$

We can now collect the stencil coefficients for $\mathbf{u}_0$ into a matrix:

$$\mathbf{D}_{\mathrm{N},1} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{(M+1)\times(M+1)},$$

and the corresponding right-hand side:

$$\mathbf{f}_0 = \begin{bmatrix} h^2/2\mathbf{f}_{0,0} + 2h\mathbf{g}_{0,0} \\ h^2\mathbf{f}_{0,1} + 2h\mathbf{g}_{0,1} \\ \vdots \\ h^2\mathbf{f}_{0,M-1} + 2h\mathbf{g}_{0,M-1} \\ h^2/2\mathbf{f}_{0,M} + 2h\mathbf{g}_{0,M} \end{bmatrix}.$$

The coupling to the neighboring column $\mathbf{u}_1$ comes from:

$$\mathbf{D}_{\mathrm{N},2} = \mathrm{diag}\,(-1, -2, \dots, -2, -1) \in \mathbb{R}^{(M+1)\times(M+1)}.$$

So, the boundary system at $i = 0$ becomes:

$$\mathbf{D}_{\mathrm{N},2}\mathbf{u}_1 + \mathbf{D}_{\mathrm{N},1}\mathbf{u}_0 = \mathbf{f}_0.$$

The analogous structure holds at the right boundary:

$$\mathbf{D}_{\mathrm{N},1}\mathbf{u}_M + \mathbf{D}_{\mathrm{N},2}\mathbf{u}_{M-1} = \mathbf{f}_M.$$

Now, comes the inner grid points, i.e., $i = 1, 2, \ldots, M-1$. For the grid point $(i, 0)$ and (i, M) we need to use *Inner Grid Points* method. Here the equation would look like

$$4\mathbf{u}_{i,0} - 2\mathbf{u}_{i,1} - \mathbf{u}_{i-1,0} - \mathbf{u}_{i+1,0} = h^2\mathbf{f}_{i,0} + 2h\mathbf{g}_{i,0},$$

and similarly for $j = M$

$$4\mathbf{u}_{i,M} - 2\mathbf{u}_{i,M-1} - \mathbf{u}_{i-1,M} - \mathbf{u}_{i+1,M} = h^2\mathbf{f}_{i,M} - 2h\mathbf{g}_{i,M}.$$

For $j = 1, 2, \ldots, M-1$ we use the five-point stencil and hence for $\mathbf{u}_i$ we get the system of equations as

$$-\mathbb{I}\mathbf{u}_{i-1} + \mathbf{D}_{N,3}\mathbf{u}_i - \mathbb{I}\mathbf{u}_{i+1} = \mathbf{f}_i,$$

where $\mathbb{I}$ is a $(M+1) \times (M+1)$ identity matrix and

$$\mathbf{D}_{\mathrm{N},3} = \begin{bmatrix} 4 & -2 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -2 & 4 \end{bmatrix} \in \mathbb{R}^{(M+1)\times(M+1)}.$$

The full block structure of the matrix $\mathbf{A}$ is:

$$\mathbf{A} = \begin{bmatrix} \mathbf{D}_{\mathrm{N},1} & \mathbf{D}_{\mathrm{N},2} & & & \\ -\mathbb{I} & \mathbf{D}_{\mathrm{N},3} & -\mathbb{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbb{I} & \mathbf{D}_{\mathrm{N},3} & -\mathbb{I} \\ & & & \mathbf{D}_{\mathrm{N},2} & \mathbf{D}_{\mathrm{N},1} \end{bmatrix}.$$

Let us now express $\mathbf{A}$ using Kronecker products:

$$\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3 + \mathbf{A}_4,$$

where: $\mathbf{A}_1 = \mathbf{A}_{1,1} \otimes \mathbf{D}_{\mathrm{N},3}$, $\mathbf{A}_2 = \mathbf{A}_{2,2} \otimes \mathbb{I}$, $\mathbf{A}_3 = \mathbf{A}_{3,3} \otimes \mathbf{D}_{\mathrm{N},1}$, and $\mathbf{A}_4 = \mathbf{A}_{4,4} \otimes \mathbf{D}_{\mathrm{N},2}$.

The matrix blocks $\mathbf{A}_{i,i}$ are defined as follows (all of size $(M+1) \times (M+1)$):

$$\mathbf{A}_{1,1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \mathbf{A}_{2,2} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ -1 & 0 & -1 & \cdots & 0 \\ 0 & -1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 0 \end{bmatrix}, \quad \mathbf{A}_{3,3} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix},$$

and

$$\mathbf{A}_{4,4} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

Compatibility condition. Since the Poisson problem with pure Neumann boundary conditions is solvable only if the right-hand side satisfies a compatibility condition, the discrete system must fulfill

$$\sum_{i=0}^{M}\sum_{j=0}^{M} h^2 \mathbf{f}_{i,j} \;=\; \sum_{\text{boundary nodes}} \big(2h\,\mathbf{g}_{i,j}\big),$$

where at the corners both boundary flux contributions are included. Equivalently, $\mathbf{A}$ is singular with a one-dimensional nullspace (the constant vector), and one must either enforce a mean-zero condition or fix one degree of freedom to obtain a unique solution.

2.2.2 Non-Rectangular Domains

Until now we have dealt with "good" domains—those having a nice geometrical structure such as a square. The ideas can, of course, be extended to rectangular domains as well. However, in practice we may encounter more complicated domains (see Fig. 2.9). In such cases, the main difficulty arises in computing the values at grid points located near the boundary.

Consider the Dirichlet problem, Eq. (1.4), defined on a simply connected domain[1] Ω.

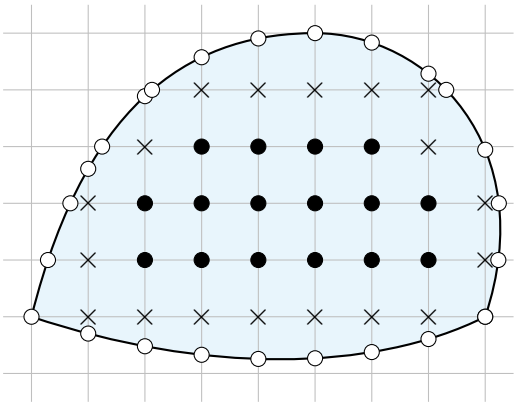


Figure 2.9: Non-square domain. Boundary nodes are denoted by $\circ$, nodes adjacent to boundary nodes are denoted by $\times$, and inner nodes by $\bullet$.

If we cover the domain with a grid of mesh size h, then the solution is known at the Dirichlet nodes from the boundary condition $g(x, y)$. We denote these boundary nodes by white dots ($\circ$).

The grid points marked with a cross ($\times$) are those for which at least one neighboring point belongs to $\partial\Omega_h$. Clearly, we cannot have the case when all neighbors belong to $\partial\Omega_h$. We denote this set by $\Omega_h^{\times}$.

The remaining interior grid points are denoted by black dots ($\bullet$), with the corresponding set denoted $\Omega_h^{\bullet}$. Thus, we have the decomposition

$$\overline{\Omega}_h = \Omega_h^{\bullet} \cup \Omega_h^{\times} \cup \partial\Omega_h.$$

The solution at $\partial\Omega_h$ is known from the Dirichlet condition; the solution at $\Omega_h^{\bullet}$ can be approximated using the standard five-point stencil. Hence, the only remaining task is to compute the solution at $\Omega_h^{\times}$.

[1]Informally, a simply connected domain is an open connected set with "no holes."

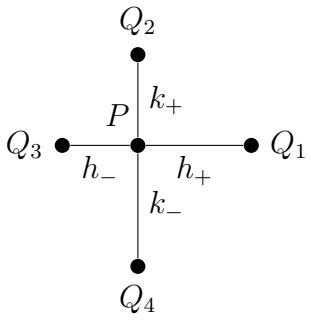


Figure 2.10: Point P with uneven spacing to its neighbors.

Let us take a general point $P \in \Omega_h^\times$, and let $\{Q_i\}_{i=1}^4$ be its neighboring nodes, labeled counterclockwise, with Q_1 being the right neighbor. The distances from P to its neighbors are denoted h_+, h_-, k_+, k_- (see Fig. 2.10). Here, $h_\pm$ correspond to x-direction distances and $k_\pm$ to y-direction distances.

To approximate the Laplacian Δu at P, we approximate u_{xx} and u_{yy} separately.

Approximation of u_{xx}: Using Taylor expansion about P in the x-direction:

$$\begin{aligned} u(x+h_+, y) &= u(x,y)|_P + h_+ u_x|_P + \frac{h_+^2}{2} u_{xx}|_P + \mathcal{O}(h^3), \\ u(x-h_-, y) &= u(x,y)|_P - h_- u_x|_P + \frac{h_-^2}{2} u_{xx}|_P + \mathcal{O}(h^3), \end{aligned}$$

where h is the maximum grid spacing in the x-direction.

Multiplying the first equation by h_-, the second by h_+, and adding eliminates u_x:

$$\begin{aligned} h_- u|_{Q_1} &= h_- u|_P + h_- h_+ u_x|_P + \frac{h_- h_+^2}{2} u_{xx}|_P + \mathcal{O}(h^4), \\ h_+ u|_{Q_3} &= h_+ u|_P - h_- h_+ u_x|_P + \frac{h_-^2 h_+}{2} u_{xx}|_P + \mathcal{O}(h^4). \end{aligned}$$

Adding and simplifying gives:

$$u_{xx}|_P = \frac{2}{h_- + h_+} \left(\frac{u|_{Q_1} - u|_P}{h_+} + \frac{u|_{Q_3} - u|_P}{h_-} \right) + \mathcal{O}(h), \tag{2.26}$$

with a formal truncation error of $\mathcal{O}(h)$.

Approximation of u_{yy}: By the same reasoning in the y-direction:

$$u_{yy}|_P = \frac{2}{k_- + k_+} \left(\frac{u|_{Q_2} - u|_P}{k_+} + \frac{u|_{Q_4} - u|_P}{k_-} \right) + \mathcal{O}(k), \tag{2.27}$$

where k is the maximum grid spacing in the y-direction.

Discrete PDE at P: Combining (2.26) and (2.27), the PDE can be approximated at P by:

$$-\frac{2}{h_- + h_+} \left(\frac{u|_{Q_1} - u|_P}{h_+} + \frac{u|_{Q_3} - u|_P}{h_-} \right) - \frac{2}{k_- + k_+} \left(\frac{u|_{Q_2} - u|_P}{k_+} + \frac{u|_{Q_4} - u|_P}{k_-} \right) = f|_P. \tag{2.28}$$

If $P \in \Omega_h^\bullet$ and $h_- = h_+ = k_- = k_+ = h$, then we recover the standard second-order finite difference method. For Neumann boundary conditions, similar approximations can be derived using the ideas presented in Sec. 2.2.1.

Now, we turn to the convergence of the method. For this, we require the notions of *stability* and *consistency*.

For stability, we require that the finite difference operator $\mathcal{L}_h$ is stable. Stability of $\mathcal{L}_h$ means that it must be: Linear, Inverse-monotone, and satisfy the comparison principle. Linearity is straightforward to verify. We therefore move on to *inverse-monotonicity*. To discuss this property, we first recall the concept of operators of *general positive type*.

Definition 2.22 (**General Positive Type Operators**). Suppose that P is a grid point and $\{Q_i\}_{i=1}^{\nu}$ are its neighboring grid points. A finite difference operator of the form

$$\mathcal{L}_h u|_P \;=\; \alpha_0\, u|_P - \sum_{j=1}^{\nu} \alpha_j\, u|_{Q_j}, \tag{2.29}$$

is said to be of *general positive type* if the coefficients satisfy:

$$\alpha_j \geq 0, \quad j = 0, 1, \ldots, \nu, \qquad \text{and} \qquad \alpha_0 \geq \sum_{j=1}^{\nu} \alpha_j. \tag{2.30}$$

Theorem 2.23 (**Inverse Monotonicity**). *Suppose that the difference operator $\mathcal{L}_h$ is defined by*

$$\mathcal{L}_h \mathbf{u}_{i,j} = \begin{cases} \mathbf{g}_{i,j}, & \text{for } i = 0 \text{ or } M, \text{ or } j = 0 \text{ or } M, \\ \mathfrak{L}_h \mathbf{u}_{i,j}, & \text{otherwise}, \end{cases}$$

where $\mathfrak{L}_h$ is of positive type and $\alpha_j > 0$ for $j = 0, 1, \ldots, \nu$. Then $\mathcal{L}_h$ is inverse monotone.

Proof. We proceed by contradiction. Suppose there exists $\mathbf{u}$ such that $\mathcal{L}_h \mathbf{u} \geq 0$ but $\mathbf{u} \not\geq 0$. Then there exists a grid point (i_0, j_0) such that $\mathbf{u}_{i_0,j_0} < 0$.

Since $\mathcal{L}_h \mathbf{u} \geq 0$ and the Dirichlet boundary conditions are applied directly (implying $\mathbf{u}_{i,j} \geq 0$ for $i, j \in \{0, M\}$), the point (i_0, j_0) must be in the interior. Let $\{Q_j\}_{j=1}^{\nu_0}$ be the neighboring nodes of (i_0, j_0). By minimality, we have

$$\mathbf{u}_{i_0,j_0} \leq \mathbf{u}|_{Q_j}, \quad j = 1, 2, \ldots, \nu_0.$$

From the definition of $\mathfrak{L}_h$,

$$\begin{aligned}\mathfrak{L}_h\mathbf{u}_{i_0,j_0} &= \alpha_0\mathbf{u}_{i_0,j_0} - \sum_{j=1}^{\nu_0}\alpha_j\mathbf{u}|_{Q_j} \\ &\leq \alpha_0\mathbf{u}_{i_0,j_0} - \sum_{j=1}^{\nu_0}\alpha_j\mathbf{u}_{i_0,j_0} \\ &= \mathbf{u}_{i_0,j_0}\left(\alpha_0 - \sum_{j=1}^{\nu_0}\alpha_j\right).\end{aligned}$$

From (2.30), $\alpha_0 \geq \sum_{j=1}^{\nu_0}\alpha_j$, and since $\mathbf{u}_{i_0,j_0} < 0$, it follows that

$$\mathfrak{L}_h\mathbf{u}_{i_0,j_0} \leq 0.$$

But we also have $\mathfrak{L}_h\mathbf{u}_{i_0,j_0} \geq 0$ by assumption, hence

$$\mathfrak{L}_h\mathbf{u}_{i_0,j_0} = 0.$$

Equality can hold only if $\alpha_0 = \sum_{j=1}^{\nu_0}\alpha_j$. Substituting into the definition of $\mathfrak{L}_h$ gives

$$\sum_{j=1}^{\nu_0}\alpha_j\left(\mathbf{u}_{i_0,j_0} - \mathbf{u}|_{Q_j}\right) = 0.$$

Since $\alpha_j > 0$ and $\mathbf{u}_{i_0,j_0} - \mathbf{u}|_{Q_j} \leq 0$ for all j, the above can hold only if

$$\mathbf{u}_{i_0,j_0} = \mathbf{u}|_{Q_j}, \quad j = 1, 2, \ldots, \nu_0.$$

Thus, the same strict minimum value is attained at all neighboring nodes. Repeating this argument propagates the minimum to the boundary, where $\mathbf{u} \geq 0$, leading to a contradiction. Therefore, $\mathcal{L}_h$ is inverse monotone. □

Now, we check whether (2.28) can be written in the form (2.29). Let $\mathbf{u}|_P = \mathbf{u}_0$ and $\mathbf{u}|_{Q_i} = \mathbf{u}_i$ for $i = 1, 2, 3, 4$. Then

$$\begin{aligned}\mathfrak{L}_h\mathbf{u}_0 &= -\frac{2}{h_- + h_+}\left(\frac{\mathbf{u}_1 - \mathbf{u}_0}{h_+} + \frac{\mathbf{u}_3 - \mathbf{u}_0}{h_-}\right) - \frac{2}{k_- + k_+}\left(\frac{\mathbf{u}_2 - \mathbf{u}_0}{k_+} + \frac{\mathbf{u}_4 - \mathbf{u}_0}{k_-}\right) \\ &= \left[\frac{2}{h_+ + h_-}\left(\frac{1}{h_+} + \frac{1}{h_-}\right) + \frac{2}{k_+ + k_-}\left(\frac{1}{k_+} + \frac{1}{k_-}\right)\right]\mathbf{u}_0 \\ &\quad - \frac{2}{h_+(h_+ + h_-)}\mathbf{u}_1 - \frac{2}{h_-(h_+ + h_-)}\mathbf{u}_3 - \frac{2}{k_+(k_+ + k_-)}\mathbf{u}_2 - \frac{2}{k_-(k_+ + k_-)}\mathbf{u}_4.\end{aligned}$$

The coefficients clearly satisfy the conditions for general positive type operators. Hence, $\mathcal{L}_h$ is inverse monotone. It is then stable by the comparison principle, as shown next.

Corollary 2.24. *The operator $\mathcal{L}_h$ comprising $\mathfrak{L}_h$ at points of $\Omega_h^\bullet \cup \Omega_h^\times$ together with the Dirichlet operator is stable.*

Proof. Without loss of generality, suppose $\Omega \subset \{(x, y) \mid 0 \leq x \leq a\}$. Then the comparison function

$$\Phi(x, y) = 1 + \frac{1}{2}x(a - x)$$

is non-negative on Ω and satisfies $\mathcal{L}_h\Phi \geq 1$. Hence, by the comparison principle, $\mathcal{L}_h$ is stable. □

Now, the only thing left to show is the second–order convergence of the method. We have already seen that it is consistent with $\mathcal{O}(h^2)$ in $\Omega_h^\bullet$ and $\mathcal{O}(h)$ in $\Omega_h^\times$.

Corollary 2.25. *The five-point approximation* (2.28) *to the Poisson equation with Dirichlet boundary conditions in a general, simply connected domain Ω is second–order convergent, provided the fourth derivatives of the exact solution are bounded in Ω.*

Proof. We split the global error $\mathbf{e}$ into two parts:

$$\mathbf{e} = \mathbf{e}^\bullet + \mathbf{e}^\times,$$

where $\mathbf{e}^\bullet$ is the error on $\Omega_h^\bullet$ and $\mathbf{e}^\times$ is the error on $\Omega_h^\times$. They satisfy the discrete problems

$$\mathcal{L}_h\mathbf{e}^\bullet = \begin{cases} 0, & \text{on } \partial\Omega_h, \\ 0, & \text{on } \Omega_h^\times, \\ \mathfrak{R}_h^\bullet, & \text{on } \Omega_h^\bullet, \end{cases} \qquad \mathcal{L}_h\mathbf{e}^\times = \begin{cases} 0, & \text{on } \partial\Omega_h, \\ \mathfrak{R}_h^\times, & \text{on } \Omega_h^\times, \\ 0, & \text{on } \Omega_h^\bullet, \end{cases} \tag{2.31}$$

where $\mathfrak{R}_h^\bullet = \mathfrak{L}_h u - f$ on $\Omega_h^\bullet$ is $\mathcal{O}(h^2)$ and $\mathfrak{R}_h^\times = \mathfrak{L}_h u - f$ on $\Omega_h^\times$ is $\mathcal{O}(h)$.

To control $\mathbf{e}^\times$ we introduce a non-negative discrete function Ψ defined by

$$\Psi|_P = \begin{cases} 0, & \text{on } \partial\Omega_h, \\ h^2, & \text{on } \Omega_h, \end{cases}$$

so that $\mathfrak{L}_h\Psi|_P = 0$ for $P \in \Omega_h^\bullet$ (since Ψ is constant there).

If $P \in \Omega_h^\times$, then at least one of its neighbors lies on $\partial\Omega_h$. Let Q_1 be such a boundary neighbor, so that $\Psi|_{Q_1} = 0$. Using (2.30) we have

$$\begin{aligned} \mathfrak{L}_h\Psi|_P &= \alpha_0\Psi|_P - \sum_{j=1}^{4} \alpha_j\Psi|_{Q_j} \\ &\geq h^2\left(\alpha_0 - \sum_{j=2}^{4}\alpha_j\right) \\ &= h^2\left(\alpha_0 - \sum_{j=1}^{4}\alpha_j\right) + \alpha_1 h^2 \\ &\geq \alpha_1 h^2. \end{aligned}$$

Since $\alpha_1 = \frac{2}{h_+(h_+ + h_-)}$ and $h_+, h_- \leq h$, it follows that $\alpha_1 h^2 \geq 1$. If P has more than one boundary neighbor, this lower bound is even larger. Thus $\mathfrak{L}_h\Psi|_P \geq 1$ for all $P \in \Omega_h^\times$.

From (2.31) and the above, we compute

$$\mathcal{L}_h \left(\mathbf{e}^{\times} - \|\mathfrak{R}_h^{\times}\|_{h,\infty} \Psi \right) \leq 0,$$

because on $\Omega_h^{\times}$ we have

$$\mathfrak{L}_h \mathbf{e}^{\times} - \|\mathfrak{R}_h^{\times}\|_{h,\infty} \mathfrak{L}_h \Psi \leq \mathfrak{R}_h^{\times} - \|\mathfrak{R}_h^{\times}\|_{h,\infty} \leq 0.$$

By inverse monotonicity of $\mathcal{L}_h$,

$$\mathbf{e}^{\times} \leq \|\mathfrak{R}_h^{\times}\|_{h,\infty} \Psi.$$

A similar argument applied to $\mathbf{e}^{\times}$ yields

$$|\mathbf{e}^{\times}| \leq \|\mathfrak{R}_h^{\times}\|_{h,\infty} h^2.$$

Since $\|\mathfrak{R}_h^{\times}\|_{h,\infty} = \mathcal{O}(h)$, this gives $\mathbf{e}^{\times} = \mathcal{O}(h^3)$ in the maximum norm.

Finally, $\mathbf{e}^{\bullet}$ is governed by a residual $\mathcal{O}(h^2)$ and the stable operator $\mathcal{L}_h$, hence $\mathbf{e}^{\bullet} = \mathcal{O}(h^2)$. Since $\mathbf{e} = \mathbf{e}^{\bullet} + \mathbf{e}^{\times}$ and $\mathbf{e}^{\bullet}$ dominates, we conclude

$$\mathbf{e} = \mathcal{O}(h^2),$$

establishing second–order convergence. □

2.2.3 Higher-Order Discretisation

In all the examples we have studied until now, all of them are second-order convergent. But can we achieve a higher-order discretisation, say of order four?

To answer this question, we start with the Taylor expansion formula in two dimensions.

Theorem 2.26. *Suppose that $u(x,y)$ and all its partial derivatives of order less than or equal to $n+1$ are continuous on $D = \{(x,y) : a \le x \le b,\ c \le y \le d\}$ and let $(x_0, y_0) \in D$. For every $(x,y) \in D$, there exist ξ between x and x_0 and μ between y and y_0 such that*

$$u(x,y) = P_n(x,y) + R_n(x,y),$$

where

$$\begin{aligned} P_n(x,y) &= u(x_0,y_0) + \left[(x-x_0)\frac{\partial u}{\partial x}(x_0,y_0) + (y-y_0)\frac{\partial u}{\partial y}(x_0,y_0)\right] \\ &\quad + \left[\frac{(x-x_0)^2}{2}\frac{\partial^2 u}{\partial x^2}(x_0,y_0) + (x-x_0)(y-y_0)\frac{\partial^2 u}{\partial x \partial y}(x_0,y_0)\right. \\ &\quad \left. + \frac{(y-y_0)^2}{2}\frac{\partial^2 u}{\partial y^2}(x_0,y_0)\right] \\ &\quad + \cdots + \frac{1}{n!}\sum_{j=0}^{n}\binom{n}{j}(x-x_0)^{n-j}(y-y_0)^j \frac{\partial^n u}{\partial x^{n-j}\partial y^j}(x_0,y_0), \end{aligned}$$

and

$$R_n(x,y) = \frac{1}{(n+1)!}\sum_{j=0}^{n+1}\binom{n+1}{j}(x-x_0)^{n+1-j}(y-y_0)^j \frac{\partial^{n+1} u}{\partial x^{n+1-j}\partial y^j}(\xi,\mu).$$

The function $P_n(x,y)$ is called the n^{th} Taylor polynomial in two variables for the function u about (x_0,y_0), and $R_n(x,y)$ is the remainder term associated with $P_n(x,y)$.

Now, consider the four diagonal points for (i,j) as shown in Fig. 2.11, labelled R_1, R_2, R_3, and R_4.

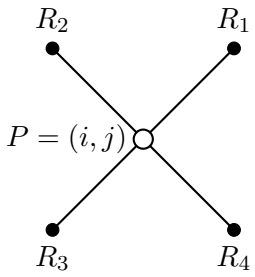


Figure 2.11: The white point represents an interior grid node (i,j), and the black points are its four nearest diagonal neighbors used in the five-point stencil.

Using the Taylor expansion at these four points gives

$$\begin{aligned}
u|_{R_1} &= u(x+h,y+h) = u_{i,j} + hu_x|_{i,j} + hu_y|_{i,j} + \frac{h^2}{2}\left(u_{xx}+2u_{xy}+u_{yy}\right)|_{i,j} + \mathcal{O}(h^3),\\
u|_{R_2} &= u(x-h,y+h) = u_{i,j} - hu_x|_{i,j} + hu_y|_{i,j} + \frac{h^2}{2}\left(u_{xx}-2u_{xy}+u_{yy}\right)|_{i,j} + \mathcal{O}(h^3),\\
u|_{R_3} &= u(x-h,y-h) = u_{i,j} - hu_x|_{i,j} - hu_y|_{i,j} + \frac{h^2}{2}\left(u_{xx}+2u_{xy}+u_{yy}\right)|_{i,j} + \mathcal{O}(h^3),\\
u|_{R_4} &= u(x+h,y-h) = u_{i,j} + hu_x|_{i,j} - hu_y|_{i,j} + \frac{h^2}{2}\left(u_{xx}-2u_{xy}+u_{yy}\right)|_{i,j} + \mathcal{O}(h^3).
\end{aligned}$$

Adding the above four equations and cancelling the third-order $\mathcal{O}(h^3)$ terms, we get

$$u_{i+1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} + u_{i-1,j+1} = 4u_{i,j} + 2h^2\left(u_{xx}+u_{yy}\right)|_{(i,j)} + \mathcal{O}(h^4).$$

Now the Laplacian $-\Delta u$ can be approximated by

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

Let us denote the new difference operator by $\mathfrak{L}_h^{\times}$, i.e.,

$$\mathfrak{L}_h^{\times}\mathbf{u}_{i,j} = \frac{1}{2h^2}\left[4\mathbf{u}_{i,j} - \mathbf{u}_{i+1,j+1} - \mathbf{u}_{i+1,j-1} - \mathbf{u}_{i-1,j-1} - \mathbf{u}_{i-1,j+1}\right].$$

Notice that this is a second-order consistent approximation. In the case of the Laplace equation, we notice that this approximation also satisfies the mean-value property. Now, to obtain a higher-order approximation we combine the newly developed five-point stencil with the one developed in Sec. 2.1.

Let us denote the five-point stencil developed in the previous section by $\mathfrak{L}_h^{+}$, i.e.,

$$\mathfrak{L}_h^{+} = \frac{\delta_x^2 + \delta_y^2}{h^2}.$$

To obtain a fourth-order approximation, consider $\lambda\mathfrak{L}_h^{\times} + (1-\lambda)\mathfrak{L}_h^{+}$ for a generic λ. The local truncation error for this method is given by

$$\begin{aligned}
\lambda\mathfrak{L}_h^{\times} + (1-\lambda)\mathfrak{L}_h^{+} &= \lambda\left[-\Delta u - \frac{h^2}{12}\left(\partial_x^4 u + 6\partial_x^2\partial_y^2 u + \partial_y^4 u\right) + \mathcal{O}(h^4)\right]\\
&\quad +(1-\lambda)\left[-\Delta u - \frac{h^2}{12}\left(\partial_x^4 u + \partial_y^4 u\right) + \mathcal{O}(h^4)\right]\\
&= -\Delta u - \frac{h^2}{12}\left(\partial_x^4 u + \partial_y^4 u\right) - \frac{6\lambda h^2}{12}\partial_x^2\partial_y^2 u + \mathcal{O}(h^4).
\end{aligned}$$

Now, for $\lambda = 1/3$ we get

$$\lambda\mathfrak{L}_h^{\times} + (1-\lambda)\mathfrak{L}_h^{+} = -\Delta u - \frac{h^2}{12}\Delta^2 u + \mathcal{O}(h^4),$$

where $\Delta^2 u = \partial_x^4 u + 2\partial_x^2\partial_y^2 u + \partial_y^4 u$. Since $-\Delta u = f$, we have $-\Delta^2 u = \Delta f$. Therefore, we have an order-4 consistency term.

Remark 2.27. We have not talked about the exact form of the local truncation error. An interested reader can use Taylor expansion to see that the $\mathcal{O}(h^4)$ term remains the leading term. However, for higher-order accuracy we require $u \in \mathcal{C}^6(\Omega)$, which demands quite high smoothness.

Nine-Point Stencil: Let $u \in \mathcal{C}^6(\Omega)$. Then the nine-point stencil formula for the Poisson equation, Eq. (2.2), is given by

$$\mathcal{L}_h^9 \mathbf{u}_{i,j} = \mathbf{f}_{i,j}^{\times},$$

where

$$\mathcal{L}_h^9 \mathbf{u}_{i,j} = \begin{cases} \mathbf{u}_{i,j}, & \text{if } i,j = 0 \text{ or } M, \\ \mathfrak{L}_h^9 \mathbf{u}_{i,j}, & \text{if } 1 \le i,j \le M-1, \end{cases} \quad \text{and} \quad \mathbf{f}_{i,j}^{\times} = \mathbf{f}_{i,j} + \frac{h^2}{12}\Delta \mathbf{f}_{i,j},$$

and

$$\begin{aligned} \mathfrak{L}_h^9 \mathbf{u}_{i,j} &= \frac{1}{3h^2}\Bigg[10\mathbf{u}_{i,j} - \frac{1}{2}\left(\mathbf{u}_{i+1,j+1} + \mathbf{u}_{i+1,j-1} + \mathbf{u}_{i-1,j+1} + \mathbf{u}_{i-1,j-1}\right) \\ &\qquad -2\left(\mathbf{u}_{i,j+1} + \mathbf{u}_{i,j-1} + \mathbf{u}_{i+1,j} + \mathbf{u}_{i-1,j}\right)\Bigg]. \end{aligned}$$

Consistency of the method is already obvious. The stability of the method is also straightforward to see, as the operator is of general positive type (see Def. 2.2.2) and hence $\mathfrak{L}_h^9$ is inverse-monotone. The next step is to show the satisfaction of the comparison lemma; for this, we can choose the same $\Phi(x, y)$ as in Fig. 2.5 to get the result. Hence, we have a convergent method of order four.

However, this method has drawbacks. The matrix obtained will not be a penta-diagonal matrix; instead, we get a nine-diagonal matrix (see Fig. 2.12). Furthermore, solving it will be more challenging as the number of non-zero entries increases.

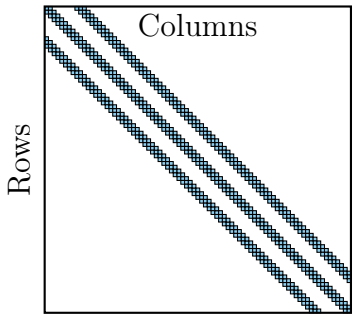


Figure 2.12: Sparsity pattern for matrix $\mathbf{A}$ developed using the nine-point stencil for $M = 10$.

Regarding the implementation of the method, the matrix $\mathbf{A}$ will have the structure

$$-\mathbf{D}_1\mathbf{u}_{i-1} + \mathbf{D}_2\mathbf{u}_i - \mathbf{D}_1\mathbf{u}_{i+1}, \quad \text{for } i = 1, 2, \ldots, M-1,$$

where

$$\mathbf{D}_1 = \begin{bmatrix} 4 & 1 & 0 & \cdots & 0 \\ 1 & 4 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 4 \end{bmatrix}, \quad \mathbf{D}_2 = \begin{bmatrix} 20 & -4 & 0 & \cdots & 0 \\ -4 & 20 & -4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 20 \end{bmatrix}.$$

The right-hand side $\mathbf{f}$ will have an extra term to incorporate $\Delta \mathbf{f}_{i,j}$. The boundary conditions can be applied in the same way as before. For Neumann boundary conditions or irregular domains, the ideas developed in the previous examples remain applicable.

2.3 Summary

In this chapter, we examined the finite difference method (FDM) for second-order elliptic equations in two dimensions. The ideas developed here extend naturally to three dimensions. The convergence theory remains unchanged, and the approach can also be generalized to Neumann boundary conditions, irregular domains, and higher-order discretizations.

If, instead of the Poisson equation, we consider a more general elliptic operator such as the convection–diffusion equation

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u = f \quad \text{in } \Omega, \tag{2.32}$$

with appropriate boundary conditions, the implementation of the FDM follows the same principles. In this case, in addition to approximating the Laplacian Δ, we must also approximate the gradient ∇u. This introduces an additional challenge. An excellent survey on this topic is given by Stynes [26].

Despite the development of modern numerical methods, the FDM remains popular in the scientific community because of its simplicity. However, it has certain limitations. One major drawback is the regularity assumption on the solution. The numerical analysis of the FDM relies on Taylor expansions of u, and achieving the expected order of accuracy typically requires $u \in \mathcal{C}^4(\Omega)$. In practice, solutions may lack such high smoothness, making it necessary to use methods that relax this assumption.

Another difficulty arises when incorporating Neumann boundary conditions or handling irregular domains. These cases require special treatment, often involving modifications to either the system matrix $\mathbf{A}$ or the right-hand side vector $\mathbf{f}$.

Also, suppose we want to find the solution at a point between the grid points, say $(i+0.5, j+0.7)$, then in this case, one either needs to do re-grid or use interpolation techniques, leading to further difficulties.

This naturally leads to the question: *do we have better methods?* In the next part of the course, we will explore approaches designed to overcome these limitations.

Chapter 3

Introduction to Functional Analysis

In the last chapter we noted that the analysis of the Finite Difference Method relies heavily on Taylor expansions, which require high differentiability of the solution, also referred to as smoothness. Let us look at some examples to see what smoothness actually means. Define

$$f_0(x) = \begin{cases} -1 & \text{if } -1 \le x < 0, \\ 1 & \text{if } 0 \le x \le 1, \end{cases} \qquad f_1(x) = |x|, \qquad f_2(x) = x|x|, \qquad f_3(x) = x^2|x|,$$

on the interval $[-1, 1]$. We now examine the behaviour of the functions $\{f_i(x)\}_{i=0}^3$ at the point $x = 0$:

1. $f_0(x)$ is not continuous at $x = 0$.
2. $f_1(x)$ is continuous at $x = 0$, but is not differentiable there.
3. $f_2(x)$ is continuous and differentiable at $x = 0$, but its second derivative does not exist at that point.
4. $f_3(x)$ is twice differentiable at $x = 0$, but not three times differentiable.

If we look at Fig. 3.1 at $(0, 0)$, we see what is meant by smoothness. As i increases, the graphs of the functions look progressively *smoother*.

But why do we need smoothness, or even continuity, in the first place? In real-world applications such as stock pricing, the price of a share may exhibit very sharp changes. For example, the value of a stock can drop from \$100 to \$0 and then rise back to \$100 within a short period of time. Like many other phenomena, this too can be modeled mathematically. The evaluation of an option price is described by the *Black–Scholes equation*, named after the American economists Fischer Sheffey Black and Myron Scholes. It is given by

$$\frac{\partial V}{\partial t} + \frac{\sigma^2}{2} S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0,$$

where $V(t, S)$ is the price of the option as a function of stock price S and time t, r is the risk-free interest rate, and σ is the volatility of the stock. As mentioned, the solution $V(t, S)$ will, in general, not be a smooth function, and hence we need to study a more general class of functions together with their properties.

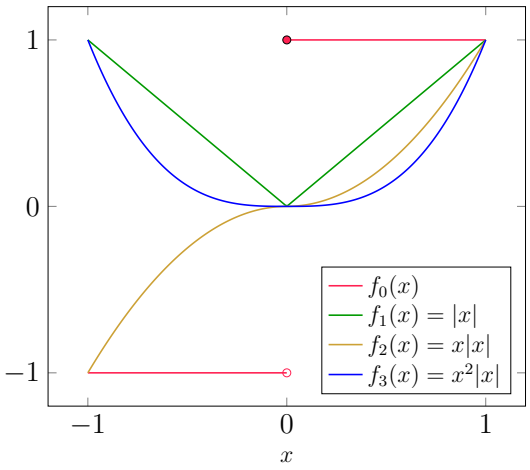


Figure 3.1: Smoothness of the functions $\{f_i(x)\}_{i=0}^3$ around $x = 0$.

Figure 3.2: Fischer Black (11 January 1938 – 30 August 1995, left) and Myron Scholes (1 July 1941–, right).

This chapter is intended as an introduction to Functional Analysis. The main topics to be covered include measurable functions, function spaces, weak derivatives, and their properties. It will present the most important concepts required for the study of PDEs, but it is by no means a complete text. For a deeper understanding of Functional Analysis, I recommend the references [20, 19].

3.1 Function Spaces

3.1.1 Banach Spaces

Until now we have looked at some very particular classes of functions such as $\mathcal{C}(\Omega)$ and $\mathcal{C}^4(\Omega)$, but now we move towards a more general notion. Let us denote such a space by X. We start with the most fundamental structure, referred to as a *vector space*.

Definition 3.1. (**Vector Space**) A *vector space* or a *linear space* over $\mathbb{R}$ (whose elements are called *scalars*) is a set X, whose elements are called *vectors*, equipped with two operations: *addition* and *scalar multiplication*, such that the following properties hold:

- **Addition**: $(x, y) \in X \times X \mapsto x + y \in X$ such that for all $x, y, z \in X$:
  - $x + y = y + x$ (Commutativity).
  - $x + (y + z) = (x + y) + z$ (Associativity).
  - There exists a unique vector $0 \in X$, called the zero vector, such that $x + 0 = x$ for all $x \in X$.
  - For all $x \in X$ there exists $-x \in X$ such that $x + (-x) = 0$.
- **Scalar Multiplication**: $(\lambda, x) \in \mathbb{R} \times X \mapsto \lambda x \in X$ such that for all $\lambda, \beta \in \mathbb{R}$ and $x, y \in X$:
  - For all $x \in X$, $1 \cdot x = x$, where 1 is the multiplicative identity of $\mathbb{R}$.
  - $\lambda(\beta x) = (\lambda\beta)x$.
  - $(\lambda + \beta)x = \lambda x + \beta x$.
  - $\lambda(x + y) = \lambda x + \lambda y$.

Example 3.2.
1. The set of real numbers $\mathbb{R}$ is a vector space over $\mathbb{R}$.
2. $\mathbb{R}^d$ is a vector space over $\mathbb{R}$.
3. The set of integers $\mathbb{Z}$ with scalars in $\mathbb{R}$ is *not* a vector space, since for $\lambda = \sqrt{3}$ and any $x \in \mathbb{Z}$, we have $\lambda x \notin \mathbb{Z}$.

After defining vector spaces, we would like to measure distances in such spaces, which leads to the notion of *normed spaces*. We have already seen the notation of a norm, but we revisit it here.

Definition 3.3. (**Normed Space**) Let X be a real vector space. A mapping $\|\cdot\| : X \to \mathbb{R}$ is called a *norm* on X if
1. $\|x\| \geq 0$ for all $x \in X$ (Non-negativity).
2. $\|x\| = 0 \Leftrightarrow x = 0$ (Definiteness).
3. $\|\lambda x\| = |\lambda|\|x\|$ for all $x \in X$ and $\lambda \in \mathbb{R}$ (Homogeneity).
4. $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (Triangle Inequality).

The pair $\big(X, \|\cdot\|\big)$ is called a *normed space*.

Example 3.4.
1. The set of real numbers $\mathbb{R}$ with the absolute value $|\cdot|$ is a normed space.
2. The space $\mathcal{C}[a, b]$ is a normed space with $\|x\| = \max_{t\in[a,b]} |x(t)|$ for $x \in \mathcal{C}[a, b]$.
3. If $\|\cdot\|$ is defined by $\|x\| = 0$ for all $x \in X$, then $\big(X, \|\cdot\|\big)$ is *not* a normed space.

We saw the importance of norms in the previous chapter, as they allow us to quantify errors. In Numerical Analysis in general, approximations arise that lead to sequences of numbers, and we hope that these sequences converge to the true solution. Before studying convergence of sequences, however, we introduce a more general concept: the notion of a *Cauchy sequence*, named after the French mathematician Augustin-Louis Cauchy.

Figure 3.3: Augustin-Louis Cauchy (21 August 1789 – 23 May 1857).

Definition 3.5. (**Cauchy Sequence**) A sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ is called a *Cauchy sequence* if for all $\varepsilon > 0$ there exists $n_0(\varepsilon) \in \mathbb{N}$ such that

$$\|x_m - x_n\| < \varepsilon \quad \forall\, m, n > n_0(\varepsilon).$$

In other words, a sequence is said to be Cauchy if for any $\varepsilon > 0$ we can find an index n_0 (depending on ε) such that all terms of the sequence beyond this index are within ε of each other.

Definition 3.6. (**Convergent Sequence**) A sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ is said to *converge* to $x \in X$ if for all $\varepsilon > 0$ there exists $n_0(\varepsilon) \in \mathbb{N}$ such that

$$\|x_n - x\| < \varepsilon \quad \forall\, n > n_0(\varepsilon).$$

In this case we write $x_n \to x$ as $n \to \infty$.

Example 3.7. In Real Analysis a classic result tells us that every convergent sequence is a Cauchy sequence, but the converse is not true in general [21, Theorem 3.11].
For instance, consider $X = \mathbb{P}[0,1]$, the space of all polynomials on $[0,1]$, equipped with the supremum norm

$$\|p\|_\infty = \max_{x\in[0,1]} |p(x)|.$$

Define

$$p_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}.$$

Then $\{p_n\}$ is a Cauchy sequence in $\big(\mathcal{C}[0,1], \|\cdot\|_\infty\big)$, since $p_n \to e^x$ uniformly on $[0,1]$. However, e^x is not a polynomial, and hence $\{p_n\}$ does not converge in $X = \mathbb{P}[0,1]$.
This illustrates that Cauchy sequences are more general than convergent sequences: a Cauchy sequence may fail to converge if the space is not *complete*. If we enlarge the space to $\mathcal{C}[0,1]$, then the same sequence $\{p_n\}$ does converge (to e^x). This observation motivates the notion of completeness.

Definition 3.8. (**Complete Space**) A normed space $\big(X, \|\cdot\|\big)$ is called *complete* if every Cauchy sequence in X converges to an element of X.

The notion of a complete space was introduced by Stefan Banach, and hence giving rise to Banach Spaces.

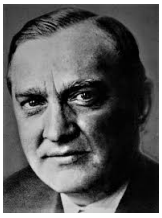


Figure 3.4: Stefan Banach (30 March 1892 – 31 August 1945).

Definition 3.9. (**Banach Space**) A complete normed space is called a *Banach space*.

Example 3.10. The Euclidean space $\mathbb{R}^d$ is a Banach space with the Euclidean norm

$$\|x\|_2 = \sqrt{\sum_{i=1}^{d} |x_i|^2}.$$

This follows since $\mathbb{R}^d$ is finite-dimensional and every finite-dimensional normed space is complete.

Lemma 3.11. *The space of continuous functions $\mathcal{C}[a,b]$ is a Banach space with*

$$\|x\|_\infty = \max_{t\in[a,b]} |x(t)|.$$

Proof. Let $\{x_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence in $\mathcal{C}[a,b]$. By definition, for every $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that

$$\|x_m - x_n\|_\infty < \varepsilon \quad \forall m, n \geq n_0.$$

This means

$$|x_m(t) - x_n(t)| < \varepsilon \quad \forall t \in [a,b],\ \forall m, n \geq n_0.$$

Fix $t_0 \in [a,b]$. Then $\{x_n(t_0)\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$, hence convergent (since $\mathbb{R}$ is complete). Define

$$x(t) = \lim_{n\to\infty} x_n(t), \quad t \in [a,b].$$

We now claim that $x_n \to x$ uniformly.

Indeed, given $\varepsilon > 0$, choose n_0 as above. Then for all $m \geq n_0$ and all $t \in [a,b]$,

$$|x_m(t) - x(t)| \leq \lim_{n\to\infty} |x_m(t) - x_n(t)| \leq \varepsilon.$$

Taking maximum over t, we get

$$\|x_m - x\|_\infty \leq \varepsilon.$$

Thus $x_n \to x$ in the $\|\cdot\|_\infty$ norm.

Finally, since x_n converges uniformly to x and each x_n is continuous, the limit x is continuous (Uniform limit theorem, [21, Theorem 7.12]). Therefore $x \in \mathcal{C}[a, b]$.

Hence $\mathcal{C}[a, b]$ is complete under $\|\cdot\|_\infty$. □

Lemma 3.12. *The space of continuous functions $\mathcal{C}[0, 2]$ is* not *a Banach space with*

$$\|x\|_1 = \int_0^1 |x(t)|\, dt.$$

Proof. We construct a Cauchy sequence in $\mathcal{C}[0, 2]$ that does not converge in $\mathcal{C}[0, 2]$ under $\|\cdot\|_1$.

Define

$$x_n(t) = \begin{cases} 0 & t \in [0, 0.5], \\ \dfrac{t - 0.5}{a_n - 0.5} & t \in [0.5, a_n], \quad a_n = 0.5 + \dfrac{1}{n}. \\ 1 & t \in [a_n, 2], \end{cases}$$

Graphically, x_n is a continuous function that rises linearly from 0 at $t = 0.5$ to 1 at $t = a_n$, and stays constant outside this interval (see Fig. 3.5).

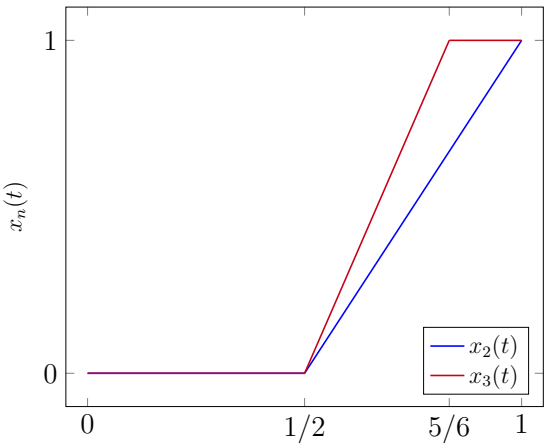


Figure 3.5: Cauchy sequence of continuous functions in $(\mathcal{C}[0, 2], \|\cdot\|_1)$.

Now consider

$$\|x_m - x_n\|_1 = \int_0^1 |x_m(t) - x_n(t)|\, dt.$$

The difference is nonzero only in the interval $[0.5, \max\{a_m, a_n\}]$. A careful computation shows

$$\|x_m - x_n\|_1 = \frac{1}{2}\left|\frac{1}{m} - \frac{1}{n}\right|.$$

Hence, given $\varepsilon > 0$, if we choose $n_0 > 1/2\varepsilon$, then for $m, n \geq n_0$, we get

$$\|x_m - x_n\|_1 < \varepsilon,$$

so $\{x_n\}$ is a Cauchy sequence in $(\mathcal{C}[0,1], \|\cdot\|_1)$.

Non-convergence: Suppose $x_n \to x$ in $\|\cdot\|_1$ for some $x \in \mathcal{C}[0,1]$. But pointwise, we see

$$x_n(t) \to \begin{cases} 0 & t \in [0, 0.5), \\ 1 & t \in (0.5, 1], \end{cases}$$

which is a discontinuous step function.

Since x would have to equal this pointwise limit almost everywhere, the candidate limit is not continuous. Hence $x \notin \mathcal{C}[0,1]$.

Therefore $\{x_n\}$ does not converge in $\mathcal{C}[0,1]$ under $\|\cdot\|_1$, and $\mathcal{C}[0,1]$ is not complete with this norm. □

In the last example, if we take the norm

$$\|x\|_2 = \sqrt{\int_0^1 |x(t)|^2 dt},$$

then the same argument will show that $\mathcal{C}[a,b]$ is not a Banach space.

From the examples that we have seen, we have noted that we have certain spaces which form a Banach space with respect to a norm but not with others. As noted we like complete spaces as they give us the limit and hence now the question is can we create a "bigger" space which is a *completion* of $\mathcal{C}[a,b]$.

3.1.2 Space of Measurable Functions

Definition 3.13. (L^p **Measurable Functions**) Let $\Omega \subset \mathbb{R}^d$ be an open and bounded domain. We denote by L^p, $1 \leq p < \infty$, the set of *measurable functions* $f : \Omega \to \mathbb{R}$ for which

$$\int_\Omega |f(x)|^p \, dx < \infty.$$

Similarly, the set of measurable functions $f : \Omega \to \mathbb{R}$ satisfying

$$\operatorname{ess\,sup} \{|f(x)| : x \in \Omega\} < \infty,$$

is denoted by $\mathrm{L}^\infty(\Omega)$. Then $\mathrm{L}^p(\Omega)$, $p \in [1, \infty]$, is a real vector space with the norm

$$\|f\|_p = \left(\int_\Omega |f(x)|^p \, dx \right)^{1/p}, \text{ for } 1 \leq p < \infty, \quad \text{and} \quad \|f\|_\infty = \operatorname{ess\,sup} \{|f(x)| : x \in \Omega\},$$

respectively.

Now, to check if $\mathrm{L}^p(\Omega)$ is a Banach space, we first need to verify whether $\|\cdot\|_p$ is a norm. The conditions of non-negativity and homogeneity are straightforward. For definiteness,

we identify functions that are equal up to a set of *measure zero*. This identification is important, since if $M \subset \Omega$ has measure zero, then

$$\int_\Omega |f(x)|^p \, dx = 0 \quad \Rightarrow \quad f(x) = 0 \ \text{ for } x \in \Omega \setminus M.$$

The last property is the triangle inequality, which in the case of L^p spaces is known as the Minkowski inequality. Its proof follows from another important result, namely Hölder's inequality.

Fun Fact: Hölder's inequality was actually proven by the British mathematician Leonard James Rogers but Hölder cites Rogers work and reproves it.


Figure 3.6: Otto Hölder (22 December 1859 – 29 August 1937, left) and Hermann Minkowski (22 June 1864–, 12 January 1909, right).

Lemma 3.14. (**Hölder's Inequality**)[19, Proposition 6.1.1] *Let $1 \leq p < \infty$ and let q be the conjugate exponent of p, i.e.,*

$$\frac{1}{p} + \frac{1}{q} = 1,$$

with $q = \infty$ when $p = 1$. If $f \in \mathrm{L}^p(\Omega)$ and $g \in \mathrm{L}^q(\Omega)$, then

$$\|fg\|_1 = \int_\Omega |f(x)g(x)| \, d\mu \leq \|f\|_p \, \|g\|_q.$$

Remark 3.15. For $p = q = 2$, we recover the well-known Cauchy–Schwarz inequality.

Lemma 3.16. (**Minkowski Inequality**)[19, Proposition 6.1.2] *Let $1 \leq p \leq \infty$ and $f, g \in \mathrm{L}^p(\Omega)$. Then $f + g \in \mathrm{L}^p(\Omega)$ and*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

One of the important properties of these L^p spaces is that they are contained in one another, a property known as embedding.

Lemma 3.17. (**Embedding Lemma**) *Let $1 \leq p < q \leq \infty$. Then*

$$\mathrm{L}^q(\Omega) \subset \mathrm{L}^p(\Omega),$$

and

$$\|f\|_p \leq \mu(\Omega)^{\frac{1}{p} - \frac{1}{q}} \, \|f\|_q.$$

Remark 3.18. The above embedding lemma states that if $p < q$, then $\mathrm{L}^q(\Omega) \subset \mathrm{L}^p(\Omega)$. In particular, if a function is L^q-integrable, then it is also L^p-integrable.

Example 3.19. In the above lemma if Ω is not bounded, then the result does not hold. Consider $f(x) = \frac{1}{1+|x|}, \quad x \in \mathbb{R}$. Then

$$\|f\|_1 = \int_{\mathbb{R}} \frac{1}{1+|x|}\, dx = 2\int_0^\infty \frac{1}{1+x}\, dx = 2\left[\log(1+x)\right]_0^\infty = \infty.$$

Hence $f \notin \mathrm{L}^1(\mathbb{R})$. However,

$$\|f\|_2^2 = \int_{\mathbb{R}} \frac{1}{(1+|x|)^2}\, dx = 2\int_0^\infty \frac{1}{(1+x)^2}\, dx = -2\left[\frac{1}{1+x}\right]_0^\infty = 2 < \infty.$$

Thus, $f \in \mathrm{L}^2(\mathbb{R})$.

Introduction to Measure Theory

In the above definition of a measurable function, we introduced the notion of *measure*. Before delving deeper into function spaces, we take a brief detour to understand what we mean by measure and measurable functions.

Remark 3.20. In mathematics, the concept of measure is a generalization of geometrical measures (length, area, volume). Since the idea of area is closely related to that of integration, measure plays a fundamental role in integration theory.

Before defining a measure, we first want to characterize the subsets of a set X that can be regarded as measurable. This leads to the notion of a σ-algebra, introduced by Émile Borel in 1898.

Figure 3.7: Émile Borel (7 January 1871 – 3 February 1956).

Definition 3.21. (σ-**algebra**) Let X be a set. A σ-algebra is a collection $\mathcal{S}$ of subsets of X such that:

1. $X \in \mathcal{S}$,
2. If $A \in \mathcal{S}$, then $A^C \in \mathcal{S}$,
3. If $A_i \in \mathcal{S}$ for $i \in \mathbb{N}$, then
$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{S}.$$

The pair $(X, \mathcal{S})$ is called a *measurable space*, and the members of $\mathcal{S}$ are called *measurable sets*.

Example 3.22.
1. On any set X, $\mathcal{S} = \{\emptyset, X\}$ is the trivial σ-algebra.
2. The power set of X is the largest possible σ-algebra.
3. The Borel σ-algebra on $\mathbb{R}$ is the smallest σ-algebra containing all open intervals of $\mathbb{R}$.

Now that we have the notion of measurable sets, we can introduce the concept of a measure.

Definition 3.23. (**Measure**) Let $(X, \mathcal{S})$ be a measurable space. A measure on X (denoted by μ) is a function
$$\mu : \mathcal{S} \to [0, \infty],$$
satisfying:

1. $\mu(\emptyset) = 0$,
2. $\mu(E) \geq 0$ for all $E \in \mathcal{S}$,
3. (Countable additivity) For any sequence $\{E_k\}_{k=1}^{\infty}$ of pairwise disjoint sets in $\mathcal{S}$,
$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k).$$

The triple $(X, \mathcal{S}, \mu)$ is called a *measure space*.

Example 3.24. 1. **Counting Measure**: Let X be a non-empty set and let $\mathcal{S}$ be the collection of all subsets of X (the power set, which is a σ-algebra). For $E \subset X$, define

$$\mu(E) = \begin{cases} \text{number of elements of } E, & \text{if } E \text{ is finite,} \\ \infty, & \text{if } E \text{ is infinite.} \end{cases}$$

This defines a measure called the *counting measure*.

2. **Dirac Measure**: Let $(X, \mathcal{S})$ be as above and let $x_0 \in X$ be a fixed point. For $E \subset X$, define

$$\mu(E) = \begin{cases} 1, & \text{if } x_0 \in E, \\ 0, & \text{otherwise.} \end{cases}$$

This defines a measure on X called the *Dirac measure*. Thus, the same measurable space can carry different measures.

After these examples of measures, we now turn to measurable functions.

Definition 3.25. (**Measurable Function**) Let $f : X \to \mathbb{R}$ be a given function. It is said to be a *measurable function* if for all $\alpha \in \mathbb{R}$ we have

$$f^{-1}\big((\alpha, \infty)\big) = \{x \in X : f(x) > \alpha\}$$

is a measurable set.

Example 3.26. 1. **Indicator Function**: Let $(X, \mathcal{S})$ be a measure space and $E \in \mathcal{S}$. The indicator function is given by

$$\chi_E(x) = \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{if } x \notin E. \end{cases}$$

Here, for $\alpha > 0$ we have $\{x \in X : \chi_E(x) > \alpha\}$ equal to E, $\emptyset$, or X.

2. **Simple Functions**: Finite linear combinations of indicator functions are called *simple functions*:

$$\varphi(x) = \sum_{i=1}^{n} a_i \chi_{E_i}(x), \quad E_i \in \mathcal{S},\ a_i \in \mathbb{R}.$$

We will discuss them later in detail.

3. **Continuous Functions**: Continuous functions are measurable since the inverse image of an open set is open, and hence belongs to the Borel σ-algebra.

One of the most important classes of measures is the *Lebesgue measure*, named after the French mathematician Henri Léon Lebesgue. This is the generalization we mentioned at the beginning of this section.

Let $I = [a, b)$ be an interval whose length is denoted by $\ell(I)$. For any subset $E \subset \mathbb{R}$,

Figure 3.8: Henri Léon Lebesgue (28 June 1875 – 26 July 1941).

the *Lebesgue outer measure*, denoted by $\mu^*(E)$, is defined as

$$\mu^*(E) = \inf \sum_{n=1}^{\infty} \ell(I_n),$$

where the infimum is taken over all countable collections of intervals with $E \subset \bigcup_{n=1}^{\infty} I_n$.

A set E is said to be *Lebesgue measurable* if it satisfies the **Carathéodory criterion**, i.e.,

$$\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c),$$

for every set $A \subset \mathbb{R}$. In this case, the outer measure μ^* restricted to measurable sets is called the *Lebesgue measure* μ.

In the case $E \subset \mathbb{R}^d$, the outer measure is defined as

$$\mu^*(E) = \inf \left\{ \sum_{B \in C} \text{vol}(B) \right\},$$

where C is a countable collection of boxes covering E, and $\text{vol}(B)$ denotes the volume of the box. If E satisfies the Carathéodory criterion, then μ^* defines the Lebesgue measure on $\mathbb{R}^d$.

Remark 3.27. A set $E \subset \mathbb{R}^d$ is said to have *measure zero* if for every $\varepsilon > 0$ there exists a countable collection of d-dimensional rectangles $\{I_k\}$ such that

1. $E \subset \bigcup_{k=1}^{\infty} I_k$,
2. $\sum_{k=1}^{\infty} \mu(I_k) < \varepsilon$.

That is, for every $\varepsilon > 0$ we can cover E with a collection of intervals whose total measure is smaller than ε.

Let us recap what we have done so far: we introduced the notion of a σ-algebra, then defined measure, and more importantly, the Lebesgue measure. The reason we studied all this is to generalize the usual notion of integration.

Definition 3.28. (**Simple Function**) Let $(X, \mathcal{S}, \mu)$ be a measure space and let $E \in \mathcal{S}$. The characteristic function of E, denoted by χ_E, is defined by

$$\chi_E(x) = \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{if } x \notin E. \end{cases}$$

A *simple function* is a function $s : X \to \mathbb{R}$ of the form

$$s(x) = \sum_{i=1}^{m} \alpha_i \chi_{A_i}(x),$$

where $\alpha_i \in \mathbb{R}$ and $A_i \in \mathcal{S}$ for $1 \leq i \leq m$.

We define the integral of a non-negative simple function $s(x)$ by

$$\int_\mu s \, d\mu := \sum_{i=1}^{m} \alpha_i \mu(A_i).$$

Lemma 3.29. *Let $(X, \mathcal{S})$ be a measurable space and let $f : X \to \mathbb{R}$ be a non-negative measurable function. Then there exists a sequence $\{s_n\}$ of non-negative simple functions such that*

$$0 \leq s_n \leq s_{n+1} \leq f, \quad \forall n,$$

and

$$\lim_{n\to\infty} s_n(x) = f(x) \quad \forall x \in X.$$

This means that every non-negative measurable function can be approximated by non-negative simple functions. Hence, we define the integral as

$$\int_X f \, d\mu := \sup \left\{ \int_X s \, d\mu : \ s \text{ simple}, \ 0 \leq s \leq f \right\}.$$

We are now in a position to define the integrability of a general measurable function f.

Definition 3.30. (**Integrable Function**) Let $(X, \mathcal{S}, \mu)$ be a measure space and let $f : X \to \mathbb{R}$ be a measurable function. Then f is said to be *integrable* if

$$\int_X |f| \, d\mu < \infty,$$

where

$$|f| = f^+ + f^-, \quad f^+ = \max\{f, 0\}, \quad f^- = -\min\{f, 0\}.$$

In this case, we define

$$\int_X f \, d\mu := \int_X f^+ \, d\mu - \int_X f^- \, d\mu.$$

Example 3.31. 1. If μ is the counting measure, then any real-valued function f on $\mathbb{N}$ is measurable and can be identified with a sequence $\{a_n\}$, where $f(n) = a_n$. In this case,

$$\int_{\mathbb{N}} f \, d\mu = \sum_{n=1}^{\infty} a_n.$$

2. If μ is the Dirac measure concentrated at $x_0 \in X$, then for any function $f : X \to \mathbb{R}$,

$$\int_X f \, d\mu = f(x_0).$$

Remark 3.32. If E is a set of measure zero, then

$$\int_E |f| \, d\mu = 0.$$

Examples of sets of measure zero include any finite set of points, or any countable set such as $\mathbb{Q} \cap [0, 1]$. This observation is important: suppose we have an integrable function defined on a domain Ω. If we replace Ω with $\Omega^* := \Omega \setminus E$, where E has measure zero, then the integral remains unchanged.

In particular, the limit obtained in Lemma 3.12 is a measurable and integrable function, since the discontinuity occurs only at finitely many points (a set of measure zero).

Finally, we recall the notion of the essential supremum, which is used in the definition of the $\|\cdot\|_\infty$ norm.

Definition 3.33. (**Essential Supremum**) Let f be a measurable function. The *essential supremum* of f, denoted by $\operatorname{ess\,sup}(f)$, is defined as

$$\operatorname{ess\,sup}(f) = \inf \{\alpha \in \mathbb{R} : f \leq \alpha \ \text{ a.e.}\}.$$

We have now completed a brief introduction to measure theory, focusing only on the results essential for our purposes. The key points included: the notion of a measure, measurable functions, and most importantly Lebesgue integration. A good introduction to Measure Theory can be found in [27].

3.1.3 Dual Space

We have defined what is a normed space, a Banach space, and even looked at a very important class of Banach spaces. Now we look at another space, a space of functions (more precisely, functionals) that act on the space X.

Definition 3.34. (**Linear Functional**) Let $(X, \|\cdot\|_X)$ be a normed space. A mapping $g : X \to \mathbb{R}$ is called *linear* if

$$g(\alpha x + \beta y) = \alpha g(x) + \beta g(y) \quad \forall \alpha, \beta \in \mathbb{R}, \quad x, y \in X.$$

A linear mapping $g : X \to \mathbb{R}$ is continuous if there exists a constant $C > 0$ such that

$$|g(x)| \leq C\|x\|_X \quad \forall x \in X.$$

Such a mapping g is referred to as a *linear functional*.

The definition of continuity might appear similar to the definition of boundedness. In fact, in Functional Analysis they are equivalent.

Remark 3.35. On finite-dimensional spaces all linear functionals are continuous, but in infinite-dimensional spaces this is not necessarily true.

Definition 3.36. (**Dual Space**) Let us define the sum of two continuous linear functionals g_1 and g_2, and the multiplication of a continuous linear functional g with a real number α, by

$$(g_1 + g_2)(x) = g_1(x) + g_2(x) \quad \text{and} \quad (\alpha g)(x) = \alpha g(x), \quad \alpha \in \mathbb{R}, \ x \in X.$$

Then the set of all continuous linear functionals forms a linear space, called the *dual space* X^*. If $g \in X^*$, we denote $g(x)$ by $\langle g, x\rangle$ for $x \in X$.

Lemma 3.37. *The set X^* of continuous linear functionals $x \mapsto \langle g, x\rangle$ on X is a Banach space with respect to the norm*

$$\|g\|_{X^*} := \sup_{0 \neq x \in X} \frac{|\langle g, x\rangle|}{\|x\|_X}.$$

Proof. We need to show two things: (1) $\|\cdot\|_{X^*}$ is a norm, and (2) $(X^*, \|\cdot\|_{X^*})$ is complete.

Step 1: $\|\cdot\|_{X^*}$ is a norm. Clearly $\|g\|_{X^*} \geq 0$ for $g \in X^*$. Also,

$$\|g\|_{X^*} = 0 \ \Leftrightarrow \ \langle g, x\rangle = 0 \ \ \forall x \in X \ \Leftrightarrow \ g = 0.$$

For homogeneity, let $\lambda \in \mathbb{R}$:

$$\begin{aligned}\|\lambda g\|_{X^*} &= \sup_{0\neq x\in X} \frac{|\langle \lambda g, x\rangle|}{\|x\|_X} \\ &= \sup_{0\neq x\in X} \frac{|\lambda\langle g, x\rangle|}{\|x\|_X} \\ &= |\lambda| \sup_{0\neq x\in X} \frac{|\langle g, x\rangle|}{\|x\|_X} \\ &= |\lambda|\|g\|_{X^*}.\end{aligned}$$

For the triangle inequality:

$$\begin{aligned}\|g_1 + g_2\|_{X^*} &= \sup_{0\neq x\in X} \frac{|\langle g_1 + g_2, x\rangle|}{\|x\|_X} \\ &= \sup_{0\neq x\in X} \frac{|\langle g_1, x\rangle + \langle g_2, x\rangle|}{\|x\|_X} \\ &\leq \sup_{0\neq x\in X} \frac{|\langle g_1, x\rangle|}{\|x\|_X} + \sup_{0\neq x\in X} \frac{|\langle g_2, x\rangle|}{\|x\|_X} \\ &= \|g_1\|_{X^*} + \|g_2\|_{X^*}.\end{aligned}$$

Hence $\|\cdot\|_{X^*}$ is a norm.

Step 2: X^* is complete. Let $\{g_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence in X^*. Then for all $\varepsilon > 0$, there exists $n_0(\varepsilon) \in \mathbb{N}$ such that

$$\|g_m - g_n\|_{X^*} < \varepsilon \quad \text{for all } m, n > n_0(\varepsilon).$$

Fix $x \in X$. Then

$$\begin{aligned}|\langle g_n, x\rangle - \langle g_m, x\rangle| &= |\langle g_n - g_m, x\rangle| \\ &\leq \|g_n - g_m\|_{X^*}\, \|x\|_X \\ &< \varepsilon\|x\|_X.\end{aligned}$$

Thus $\{\langle g_n, x\rangle\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$, hence convergent. Define

$$g(x) := \lim_{n\to\infty} \langle g_n, x\rangle \quad \forall x \in X.$$

Linearity: For $\alpha, \beta \in \mathbb{R}$ and $x, y \in X$,

$$\begin{aligned}g(\alpha x + \beta y) &= \lim_{n\to\infty} \langle g_n, \alpha x + \beta y\rangle \\ &= \alpha \lim_{n\to\infty} \langle g_n, x\rangle + \beta \lim_{n\to\infty} \langle g_n, y\rangle \\ &= \alpha g(x) + \beta g(y).\end{aligned}$$

Boundedness: Since every Cauchy sequence in a normed space is bounded, there exists $C > 0$ such that $\|g_n\|_{X^*} \leq C$ for all n. Then

$$|\langle g_n, x\rangle| \leq \|g_n\|_{X^*}\, \|x\|_X \leq C\|x\|_X.$$

Passing to the limit $n \to \infty$, we obtain

$$|g(x)| \leq C\|x\|_X, \quad \forall x \in X,$$

which shows that g is continuous.

Therefore, $g \in X^*$ and $g_n \to g$ in X^*. Hence X^* is complete, i.e., a Banach space. □

Example 3.38. For $1 < p < \infty$, the dual of $\mathrm{L}^p(\Omega)$ is $\mathrm{L}^q(\Omega)$, where q is the conjugate exponent of p, i.e.,

$$\frac{1}{p} + \frac{1}{q} = 1.$$

For $p = 1$, the dual of $\mathrm{L}^1(\Omega)$ is $\mathrm{L}^\infty(\Omega)$, but for $p = \infty$ the dual space is the space of bounded finitely additive signed measures (strictly larger than $\mathrm{L}^1(\Omega)$).
At first glance, it may appear strange that $\mathrm{L}^q(\Omega)$ is the dual of $\mathrm{L}^p(\Omega)$, since L^q is the space of q-integrable functions, whereas $(\mathrm{L}^p)^*$ is the space of continuous linear functionals. In fact, there exists an *isometric isomorphism* between $(\mathrm{L}^p)^*$ and L^q: every $g \in \mathrm{L}^q$ induces a functional $T_g \in (\mathrm{L}^p)^*$ defined by

$$T_g(f) = \int_\Omega f(x)g(x)\,dx \quad \text{for } f \in \mathrm{L}^p(\Omega),$$

and this correspondence preserves the norm. Hence we may identify elements of $(\mathrm{L}^p)^*$ with elements of L^q.

3.1.4 Hilbert Space

Until now we have discussed vector spaces and how to measure distances in them. Now, we want to see how the elements of X interact with one another. For this, the notion of an inner product comes.

Definition 3.39. (**Inner Product and Hilbert Space**) Let X be a linear space. A mapping $(\cdot,\cdot) : X \times X \to \mathbb{R}$ is called an inner product on X if

1. $(x, y) = (y, x)$ for all $x, y \in X$ (symmetry),
2. $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$ for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{R}$ (linearity in the first argument),
3. $(x, x) > 0$ for $x \neq 0$ (positive-definiteness).

The space $(X, (\cdot,\cdot))$ is referred to as an *inner product space*. An inner product also induces a norm on X denoted by

$$\|x\|_X = \sqrt{(x, x)}.$$

If $(X, \|\cdot\|_X)$ is complete, then it is referred to as a *Hilbert space*, named after the German mathematician David Hilbert.

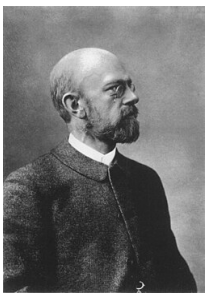


Figure 3.9: David Hilbert (23 January 1862 – 14 February 1943).

Example 3.40.

1. $X = \mathbb{R}^n$ is an inner product space with $(\cdot, \cdot)$ being the dot product, i.e.,
$$(x, y) = \sum_{i=1}^{n} x_i y_i \quad \text{for } x, y \in \mathbb{R}^n.$$
2. $X = L^2(\Omega)$. For $f, g \in X$, the inner product defined by
$$(f, g) = \int_\Omega f(x)g(x)\,dx,$$
makes $L^2(\Omega)$ into a Hilbert space.
3. $X = \mathcal{C}(\Omega)$ with the inner product
$$(f, g) = \int_\Omega f(x)g(x)\,dx$$
is an inner product space, but it is not a Hilbert space as it is not complete under the induced norm.
4. $X = L^p(\Omega)$, $p \neq 2$, is not an inner product space but merely a normed space.

Remark 3.41. One way to check if a normed space is induced by an inner product is to verify the *parallelogram law*, i.e.,

$$\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + 2\|g\|^2,$$

for all $f, g \in X$. For $(\mathrm{L}^p, \|\cdot\|_p)$ with $p \neq 2$, this law fails and hence it is not an inner product space.

For instance, consider

$$f(x) = \begin{cases} 1 & 0 \leq x < 0.5, \\ 0 & 0.5 \leq x \leq 1, \end{cases} \quad g(x) = \begin{cases} 0 & 0 \leq x < 0.5, \\ 1 & 0.5 \leq x \leq 1. \end{cases}$$

Then $\|f\|_1^2 = \|g\|_1^2 = 0.25$, but $\|f + g\|_1^2 = 1$ and $\|f - g\|_1^2 = 1$. Hence, $(\mathrm{L}^1, \|\cdot\|_1)$ is not a Hilbert space.

Remark 3.42. So far, we have looked at normed spaces, Banach spaces, inner product spaces, and Hilbert spaces.

- Normed spaces are the most general framework.
- Banach spaces are normed spaces that are complete.
- Inner product spaces are spaces equipped with an inner product, which induces a norm. Hence, every inner product space is a normed space.
- Hilbert spaces are special classes of inner product spaces that are complete in the norm induced by the inner product.

Remark 3.43. Given a normed space, if its norm comes from an inner product, then this inner product is unique (determined by the norm) via the polarization identity.

Now, what is so nice about inner product spaces? We introduced the notion of functionals in the last section. If we have a linear functional $g \in X^*$, then we can characterise its entire action on X using only an element of X and the inner product. This is called the *Riesz Representation Theorem*, named after the Hungarian mathematician Frigyes Riesz.

Figure 3.10: Frigyes Riesz (22 January 1880 – 28 February 1956).

Theorem 3.44. (**Riesz Representation Theorem**)[6, Theorem 2.4.2] *Let X be a Hilbert space. Then for every $g \in X^*$ there exists a unique $u_g \in X$ such that for all $x \in X$,*

$$(x, u_g) = \langle g, x \rangle, \qquad \|u_g\|_X = \|g\|_{X^*}.$$

Hence, we can identify the elements of the dual space X^* with the elements of the Hilbert space X itself.

We have not proved any theorems here. In the following sections, if required, we will use these results and prove them as needed.

3.2 Sobolev Spaces

Until now we have discussed integrable functions, but as we are working in a PDE setting, we are interested in differentiable functions.

The classical notion of the derivative of a function $u(x)$ that we are familiar with is

$$u'(x) = \lim_{h \to 0} \frac{u(x+h) - u(x)}{h}.$$

The above definition is "local" in the sense that it describes the derivative around the point x. However, in PDE analysis we are interested in a more global view: we want to interpret derivatives in such a way that they belong to L^2 (or more generally L^p) spaces.

To generalize the notion of derivatives, we need some preliminary notations.

Definition 3.45. (**Compact Support**) Let $f : \Omega \to \mathbb{R}$ be a real-valued function defined on a domain $\Omega \subseteq \mathbb{R}^d$. The *support* of f, written as $\text{supp}(f)$, is the closure of the set of points in Ω where f is non-zero:

$$\text{supp}(f) = \overline{\{x \in \Omega : f(x) \neq 0\}}.$$

If $\text{supp}(f)$ is compact,[a] then we say that f has *compact support*.

[a] In $\mathbb{R}^d$ a compact set is one which is closed and bounded. For example, $[0, 1]$ is compact, but $[0, \infty)$ is not compact (closed but not bounded).

Outside the support of f one can naturally extend the function by defining it to be zero. If Ω is bounded, then saying that u has compact support in Ω is equivalent to saying that u vanishes in a neighborhood of $\partial\Omega$.

Example 3.46. The function $f : [-1, 1] \to \mathbb{R}$ defined by $f(x) = 1 - x^2$ has $\text{supp}(f) = [-1, 1]$, which is compact.

Definition 3.47. Let $\Omega \subset \mathbb{R}^d$. We denote by $\mathcal{D}(\Omega)$ or $\mathcal{C}_0^\infty(\Omega)$ the set of infinitely differentiable functions with compact support in Ω.

The space $\mathcal{D}(\Omega)$ will be our "test function" space, which we use to generalize the notion of derivatives.

Definition 3.48. (**Locally Integrable Functions**) Given a domain $\Omega \subseteq \mathbb{R}^d$, the set of *locally integrable* functions, denoted by $L^1_{\text{loc}}(\Omega)$, is defined as

$$L^1_{\text{loc}}(\Omega) = \{f : f \in L^1(K) \quad \text{for all compact } K \subset \Omega\}.$$

Remark 3.49. Given $f \in L^1_{\text{loc}}(\Omega)$, one can always define a linear functional on $\mathcal{D}(\Omega)$ by

$$T_f(\varphi) = \int_\Omega f(x)\varphi(x)\,dx, \qquad \varphi \in \mathcal{D}(\Omega).$$

Such functionals are called *distributions*.

Example 3.50. 1. Any continuous function f defined on Ω belongs to $\mathrm{L}^1_{\mathrm{loc}}(\Omega)$, since continuous functions are bounded on compact sets, and compact subsets of Ω have finite measure.

2. Let $\mathbf{x} \in \mathbb{R}^d$ and define
$$\delta_{\mathbf{x}}(\varphi) = \varphi(\mathbf{x}) \quad \forall\ \varphi \in \mathcal{D}(\mathbb{R}^d).$$
This is a linear functional on $\mathcal{D}(\mathbb{R}^d)$. If $\mathbf{x} = \mathbf{0}$, then this is the *Dirac distribution* at the origin.
This functional cannot be represented by a function in $\mathrm{L}^1_{\mathrm{loc}}$, and hence it is not itself a function.
Proof of non-representability: Assume that $\delta_{\mathbf{0}}$ can be represented by a function $f \in \mathrm{L}^1_{\mathrm{loc}}$. Then, for every $\varepsilon > 0$, let $\varphi_\varepsilon \in \mathcal{D}(\mathbb{R}^d)$ be such that - $\mathrm{supp}(\varphi_\varepsilon) \subset B(\mathbf{0};\varepsilon)$, - $0 \le \varphi_\varepsilon \le 1$, - $\varphi_\varepsilon = 1$ on $B[\mathbf{0};\varepsilon/2]$.
Then
$$\delta_{\mathbf{0}}(\varphi_\varepsilon) = \varphi_\varepsilon(\mathbf{0}) = 1,$$
while, by the assumption,
$$\delta_{\mathbf{0}}(\varphi_\varepsilon) = \int_{\mathbb{R}^d} f(\mathbf{x})\varphi_\varepsilon(\mathbf{x})\,d\mathbf{x} = \int_{B(\mathbf{0};\varepsilon)} f(\mathbf{x})\varphi_\varepsilon(\mathbf{x})\,d\mathbf{x} \le \int_{B(\mathbf{0};\varepsilon)} |f(\mathbf{x})|\,d\mathbf{x}.$$
As $\varepsilon \to 0$, the last integral tends to zero by local integrability of f, leading to a contradiction. Therefore, $\delta_{\mathbf{0}}$ cannot be represented by an $\mathrm{L}^1_{\mathrm{loc}}$ function.

Definition 3.51. (**Weak Derivatives**) A function $f \in \mathrm{L}^1_{\mathrm{loc}}(\Omega)$ has a *weak derivative* $D^{\boldsymbol{\alpha}} f$ provided there exists a function $g \in \mathrm{L}^1_{\mathrm{loc}}(\Omega)$ such that
$$\int_\Omega g(x)\varphi(x)\,dx = (-1)^{|\boldsymbol{\alpha}|}\int_\Omega f(x)\,\varphi^{(\boldsymbol{\alpha})}(x)\,dx \quad \forall\ \varphi \in \mathcal{D}(\Omega).$$
If such a g exists, we define $D^{\boldsymbol{\alpha}} f := g$.

The notion of weak derivative may appear abstract at first, but let us see how it naturally arises. Suppose $f \in \mathcal{C}^1(\Omega)$ with $\Omega \subset \mathbb{R}$. Then, by integration by parts,
$$\int_\Omega f(x)\varphi'(x)\,dx = f(x)\varphi(x)\Big|_{\partial\Omega} - \int_\Omega f'(x)\varphi(x)\,dx.$$
Since $\varphi \in \mathcal{D}(\Omega)$ vanishes on $\partial\Omega$, the boundary term drops out:
$$\int_\Omega f(x)\varphi'(x)\,dx = -\int_\Omega f'(x)\varphi(x)\,dx.$$
Thus $f'(x)$ appears in the identity above.

If $f'(x)$ does not exist in the classical sense, but some $g \in \mathrm{L}^1_{\mathrm{loc}}(\Omega)$ satisfies the same identity for all test functions $\varphi \in \mathcal{D}(\Omega)$, then we call g the *weak derivative* of f.

The reason this works is that $\mathcal{D}(\Omega)$ is dense[1] in $\mathrm{L}^p(\Omega)$ for $1 \le p < \infty$, so such g (if it

[1] A subset A is *dense* in X if $\overline{A} = X$.

exists) is uniquely determined.

Example 3.52. Let $d = 1$, $\Omega = [-1, 1]$, and $f(x) = 1 - |x|$.
First weak derivative. We claim the weak derivative of f is

$$g(x) = \begin{cases} 1, & x < 0, \\ -1, & x > 0. \end{cases}$$

(its value at $x = 0$ is irrelevant). Indeed, for $\varphi \in \mathcal{D}(\Omega)$,

$$\begin{aligned} \int_{-1}^{1} f(x)\varphi'(x)\,dx &= \int_{-1}^{0} (1+x)\varphi'(x)\,dx + \int_{0}^{1} (1-x)\varphi'(x)\,dx \\ &= (1+x)\varphi(x)\Big|_{-1}^{0} - \int_{-1}^{0} \varphi(x)\,dx \\ &\quad + (1-x)\varphi(x)\Big|_{0}^{1} + \int_{0}^{1} \varphi(x)\,dx. \end{aligned}$$

Since $\varphi(\pm 1) = 0$ and φ is continuous at 0, this becomes

$$-\int_{-1}^{0} (1)\cdot\varphi(x)\,dx - \int_{0}^{1} (-1)\cdot\varphi(x)\,dx = -\int_{-1}^{1} g(x)\varphi(x)\,dx.$$

Hence g is the weak derivative of f.
Second weak derivative. Now let h be the weak derivative of g. Then for $\varphi \in \mathcal{D}([-1, 1])$,

$$\begin{aligned} \int_{-1}^{1} h(x)\varphi(x)\,dx &= -\int_{-1}^{1} g(x)\varphi'(x)\,dx \\ &= -\Big(\int_{-1}^{0} 1\cdot\varphi'(x)\,dx + \int_{0}^{1} (-1)\cdot\varphi'(x)\,dx\Big) \\ &= -\big(\varphi(0) - \varphi(-1) - (\varphi(1) - \varphi(0))\big). \end{aligned}$$

Since $\varphi(\pm 1) = 0$, this reduces to

$$2\varphi(0) = \int_{-1}^{1} 2\delta_0(x)\,\varphi(x)\,dx.$$

Thus the weak derivative is $h(x) = 2\delta_0(x)$. However, $\delta_0 \notin \mathrm{L}^1_{\mathrm{loc}}(\Omega)$, so the second weak derivative (and higher ones) do not exist as functions in $\mathrm{L}^1_{\mathrm{loc}}$.

But in the second differentiation example we obtained an object that is not a weak derivative in $\mathrm{L}^1_{\mathrm{loc}}$, but still makes sense as a *distribution*. This motivates a more general concept, the *distributional derivative*. Before that, we recall the notion of distributions.

Definition 3.53. (**Distribution**) A continuous linear functional on $\mathcal{D}(\Omega)$ is called a *distribution* on Ω. The space of all distributions is denoted by $\mathcal{D}'(\Omega)$, i.e., the (topological) dual of $\mathcal{D}(\Omega)$.

Example 3.54.

1. Let $f \in \mathrm{L}^1_{\mathrm{loc}}(\Omega)$. Define $T_f : \mathcal{D}(\Omega) \to \mathbb{R}$ by
$$T_f(\varphi) = \int_\Omega f(x)\varphi(x)\,dx.$$
Then T_f is a distribution.
2. Let $\mathbf{x} \in \mathbb{R}^d$. Define the Dirac distribution by
$$\delta_{\mathbf{x}}(\varphi) = \varphi(\mathbf{x}), \qquad \varphi \in \mathcal{D}(\mathbb{R}^d).$$

Physicist Paul Dirac introduced the notion of the Dirac δ "function," but it initially lacked rigorous mathematical meaning. Laurent Schwartz later provided the correct framework in the theory of distributions.

Figure 3.11: Paul Dirac (8 August 1902 – 20 October 1984, left) and Laurent Schwartz (5 March 1915 – 4 July 2002, right).

Definition 3.55. (**Distributional Derivative**) Let $\Omega \subset \mathbb{R}^d$ be an open set and let $T \in \mathcal{D}'(\Omega)$. Given a multi-index $\boldsymbol{\alpha}$, the *distributional derivative* of T is defined by
$$D^{\boldsymbol{\alpha}}T(\varphi) = (-1)^{|\boldsymbol{\alpha}|}T\big(D^{\boldsymbol{\alpha}}\varphi\big), \qquad \forall\ \varphi \in \mathcal{D}(\Omega).$$

Example 3.56. 1. The distribution $2\delta_0$ obtained earlier is the distributional derivative of $g(x) = \text{sign}(x)$.

2. Consider the Heaviside function on $\mathbb{R}$:

$$H(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Clearly $H \in \mathrm{L}^1_{\text{loc}}(\mathbb{R})$, hence it defines a distribution. For $\varphi \in \mathcal{D}(\mathbb{R})$,

$$T'_H(\varphi) = (-1)T_H(\varphi') = -\int_0^\infty \varphi'(x)\,dx = \varphi(0) = \delta_0(\varphi).$$

Thus, δ_0 is the distributional derivative of H.

If a function f is sufficiently smooth, then the classical, weak, and distributional derivatives all coincide. If $f \in \mathrm{L}^1_{\text{loc}}$ but is not classically differentiable, its weak derivative may still exist; if not, its distributional derivative always exists.

Having introduced weak derivatives, we can now generalize Lebesgue norms and spaces to include derivatives.

Definition 3.57. (**Sobolev Spaces**) Let $k \in \mathbb{N}_0$ and let $1 \leq p \leq \infty$. Suppose $f \in \mathrm{L}^1_{\text{loc}}(\Omega)$ and all weak derivatives $D^{\boldsymbol{\alpha}} f$ exist for $|\boldsymbol{\alpha}| \leq k$. Then we define the Sobolev norm

$$\|f\|_{\mathrm{W}^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\boldsymbol{\alpha}|\leq k} \|D^{\boldsymbol{\alpha}} f\|^p_{\mathrm{L}^p(\Omega)}\right)^{1/p}, & 1 \leq p < \infty, \\ \max_{|\boldsymbol{\alpha}|\leq k} \|D^{\boldsymbol{\alpha}} f\|_{\mathrm{L}^\infty(\Omega)}, & p = \infty. \end{cases}$$

The *Sobolev space* $\mathrm{W}^{k,p}(\Omega)$ is then defined as

$$\mathrm{W}^{k,p}(\Omega) := \left\{ f \in \mathrm{L}^1_{\text{loc}}(\Omega) : \ \|f\|_{\mathrm{W}^{k,p}(\Omega)} < \infty \right\}.$$

The name Sobolev space comes from the Russian mathematician Sergei Sobolev.

Figure 3.12: Sergei Sobolev (6 October 1908 – 3 January 1989).

It is straightforward to check that $\|\cdot\|_{\mathrm{W}^{k,p}(\Omega)}$ defines a norm, and the corresponding Sobolev space is in fact a Banach space.

When we defined a norm in Sec. 3.1.1 we forgot to mention another important concept called a *semi-norm*. A semi-norm is one which satisfies properties 1, 3, and 4 of a norm, but

may have $\|x\| = 0$ even though $x \neq 0$.

Example 3.58. Define $p : \mathbb{R}^2 \to \mathbb{R}$ by $p(\mathbf{x}) = x_1$. Then this is a semi-norm since, for $\mathbf{x} = (0, 2)$, we have $p(\mathbf{x}) = 0$ but $\mathbf{x} \neq \mathbf{0}$.

Definition 3.59. For k a non-negative integer and $f \in \mathrm{W}^{k,p}(\Omega)$ we define the *Sobolev semi-norm* by

$$|f|_{\mathrm{W}^{k,p}(\Omega)} = \left(\sum_{|\boldsymbol{\alpha}|=k} \|D^{\boldsymbol{\alpha}}_{\mathrm{w}} f\|^p_{\mathrm{L}^p(\Omega)} \right)^{1/p},$$

for $1 \leq p < \infty$, and in the case $p = \infty$

$$|f|_{\mathrm{W}^{k,\infty}(\Omega)} = \max_{|\boldsymbol{\alpha}|=k} \|D^{\boldsymbol{\alpha}}_{\mathrm{w}} f\|_{\mathrm{L}^\infty(\Omega)}.$$

Now let us look at some Sobolev spaces and how these norms look in practice.

Example 3.60.

1. For $p = 1$, $k = 1$,

$$\mathrm{W}^{1,1}(\Omega) = \left\{ f \in \mathrm{L}^1_{\mathrm{loc}}(\Omega) : \ \|f\|_{\mathrm{W}^{1,1}(\Omega)} < \infty \right\},$$

where

$$\|f\|_{\mathrm{W}^{1,1}(\Omega)} = \sum_{|\boldsymbol{\alpha}|\leq 1} \|D^{\boldsymbol{\alpha}}_{\mathrm{w}} f\|_{\mathrm{L}^1(\Omega)}.$$

For $\Omega \subset \mathbb{R}^2$, we have $\boldsymbol{\alpha} = (0, 0), (1, 0), (0, 1)$. Hence

$$\|f\|_{\mathrm{W}^{1,1}(\Omega)} = \|f\|_{\mathrm{L}^1(\Omega)} + \left\| \frac{\partial f}{\partial x} \right\|_{\mathrm{L}^1(\Omega)} + \left\| \frac{\partial f}{\partial y} \right\|_{\mathrm{L}^1(\Omega)}.$$

2. For $p = 2$, $k = 1$,

$$\mathrm{W}^{1,2}(\Omega) = \left\{ f \in \mathrm{L}^1_{\mathrm{loc}}(\Omega) : \ \|f\|_{\mathrm{W}^{1,2}(\Omega)} < \infty \right\},$$

where

$$\|f\|^2_{\mathrm{W}^{1,2}(\Omega)} = \sum_{|\boldsymbol{\alpha}|\leq 1} \|D^{\boldsymbol{\alpha}}_{\mathrm{w}} f\|^2_{\mathrm{L}^2(\Omega)} = \|f\|^2_{\mathrm{L}^2(\Omega)} + \left\| \frac{\partial f}{\partial x} \right\|^2_{\mathrm{L}^2(\Omega)} + \left\| \frac{\partial f}{\partial y} \right\|^2_{\mathrm{L}^2(\Omega)}.$$

For the special case $p = 2$, the space $\mathrm{L}^2(\Omega)$ is a Hilbert space. Similarly, the Sobolev space $\mathrm{W}^{k,2}(\Omega)$ is a Hilbert space with the inner product

$$(u, v)_{\mathrm{W}^{k,2}(\Omega)} := \sum_{|\boldsymbol{\alpha}|\leq k} (D^{\boldsymbol{\alpha}}_{\mathrm{w}} u, D^{\boldsymbol{\alpha}}_{\mathrm{w}} v)_{\mathrm{L}^2(\Omega)} = \sum_{|\boldsymbol{\alpha}|\leq k} \int_\Omega D^{\boldsymbol{\alpha}}_{\mathrm{w}} u(\mathbf{x})\, D^{\boldsymbol{\alpha}}_{\mathrm{w}} v(\mathbf{x})\, d\mathbf{x}, \quad \forall\, u, v \in \mathrm{W}^{k,2}(\Omega).$$

This space is so important that it has another notation:

$$\mathrm{W}^{k,2}(\Omega) = \mathrm{H}^k(\Omega).$$

From now on we shorten some notation. If the domain is clear we write $\|\cdot\|_{\mathrm{W}^{k,p}(\Omega)}$ simply as $\|\cdot\|_{k,p}$ (and similarly for the semi-norms). For $\mathrm{H}^k(\Omega)$ we write $\|\cdot\|_{\mathrm{H}^k(\Omega)}$ as $\|\cdot\|_k$.

We also use $D^{\boldsymbol{\alpha}}$ to denote $D^{\boldsymbol{\alpha}}_{\mathrm{w}}$, and write ∇f for the gradient, i.e.,

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d}\right).$$

Definition 3.61 ($\mathbf{W_0^{k,p}(\Omega)}$ Space). We denote by $\mathrm{W}_0^{k,p}(\Omega)$ the closure of $\mathcal{C}_0^\infty(\Omega)$ in the norm $\|\cdot\|_{k,p}$. Similarly, $\mathrm{H}_0^k(\Omega)$ is the closure of $\mathcal{C}_0^\infty(\Omega)$ in $\mathrm{H}^k(\Omega)$.

We now mention an important inequality that relates the Sobolev semi-norm and the norm: the Poincaré inequality.

Lemma 3.62 (Poincaré Inequality). [16, Lemma 1.7] *There exists a positive constant C_{P} (depending only on Ω) such that*

$$\|v\|_{0,p} \le C_{\mathrm{P}} |v|_{1,p}, \quad \forall\, v \in \mathrm{W}_0^{1,p}(\Omega),$$

where $\|v\|_{0,p} = \|v\|_{\mathrm{L}^p(\Omega)}$.

Proof. We split the proof into two steps.

Step 1: Reduction to smooth compactly supported functions. Assume the Poincaré inequality holds for every $v \in \mathcal{C}_0^\infty(\Omega)$, i.e. there is a constant C_{P} such that

$$\|w\|_{\mathrm{L}^p(\Omega)} \le C_{\mathrm{P}}\, |w|_{1,p}, \qquad \forall w \in \mathcal{C}_0^\infty(\Omega).$$

Let $v \in \mathrm{W}_0^{1,p}(\Omega)$. By definition of $\mathrm{W}_0^{1,p}(\Omega)$ there exists a sequence $\{v_n\} \subset \mathcal{C}_0^\infty(\Omega)$ with $v_n \to v$ in the $\|\cdot\|_{1,p}$-norm. Using the triangle inequality and the assumed inequality for v_n we get, for every n,

$$\begin{aligned}
\|v\|_{\mathrm{L}^p(\Omega)} &\le \|v - v_n\|_{\mathrm{L}^p(\Omega)} + \|v_n\|_{\mathrm{L}^p(\Omega)} \\
&\le \|v - v_n\|_{\mathrm{L}^p(\Omega)} + C_{\mathrm{P}}\, |v_n|_{1,p} \\
&\le \|v - v_n\|_{\mathrm{L}^p(\Omega)} + C_{\mathrm{P}}\big(|v - v_n|_{1,p} + |v|_{1,p}\big) \\
&\le (1 + C_{\mathrm{P}})\|v - v_n\|_{1,p} + C_{\mathrm{P}}\, |v|_{1,p}.
\end{aligned}$$

Letting $n \to \infty$ and using $v_n \to v$ in $\|\cdot\|_{1,p}$ gives $\|v\|_{\mathrm{L}^p(\Omega)} \le C_{\mathrm{P}}\, |v|_{1,p}$. Thus it suffices to prove the inequality for $v \in \mathcal{C}_0^\infty(\Omega)$.

Step 2: Proof for $v \in \mathcal{C}_0^\infty(\Omega)$. Extend v by zero outside Ω. Choose $a > 0$ such that $\Omega \subset [-a, a] \times \mathbb{R}^{d-1}$. For $\mathbf{x} = (x_1, \ldots, x_d)$ we have

$$v(\mathbf{x}) = \int_{-a}^{x_1} \frac{\partial v}{\partial x_1}(s, x_2, \ldots, x_d)\, ds,$$

because $v(-a, x_2, \ldots, x_d) = 0$ (extension by zero and compact support).

Let q be the conjugate exponent of p (so $1/p + 1/q = 1$). By Hölder's inequality,

$$\begin{aligned} |v(\mathbf{x})| &\le \left(\int_{-a}^{x_1} \left| \frac{\partial v}{\partial x_1}(s, x_2, \ldots, x_d) \right|^p ds \right)^{1/p} \left(\int_{-a}^{x_1} 1^q ds \right)^{1/q} \\ &\le (2a)^{1/q} \left(\int_{-a}^{a} \left| \frac{\partial v}{\partial x_1}(s, x_2, \ldots, x_d) \right|^p ds \right)^{1/p}. \end{aligned}$$

Raising to the p-th power gives

$$|v(\mathbf{x})|^p \le (2a)^{p/q} \int_{-a}^{a} \left| \frac{\partial v}{\partial x_1}(s, x_2, \ldots, x_d) \right|^p ds.$$

Integrate this inequality over all $\mathbf{x} \in \Omega$. For fixed $x_2, \ldots, x_d$ integrate in $x_1 \in [-a, a]$ and then integrate over the remaining variables; by Fubini's theorem we obtain

$$\begin{aligned} \int_\Omega |v(\mathbf{x})|^p \, d\mathbf{x} &\le (2a)^{p/q} \int_\Omega \int_{-a}^{a} \left| \frac{\partial v}{\partial x_1}(s, x_2, \ldots, x_d) \right|^p ds \, d\mathbf{x} \\ &= (2a)^{p/q} \int_{-a}^{a} \int_{\{\,(x_2,\ldots,x_d):(s,x_2,\ldots,x_d)\in\Omega\,\}} \left| \frac{\partial v}{\partial x_1}(s, x_2, \ldots, x_d) \right|^p d(x_2 \cdots x_d)\, ds \\ &\le (2a)^{p/q} \cdot (2a) \int_\Omega \left| \frac{\partial v}{\partial x_1}(\mathbf{x}) \right|^p d\mathbf{x} \\ &= (2a)^{1+p/q} \int_\Omega \left| \frac{\partial v}{\partial x_1}(\mathbf{x}) \right|^p d\mathbf{x}. \end{aligned}$$

Since $1 + p/q = p$ (because $1/q = 1 - 1/p$), we get

$$\|v\|^p_{\mathrm{L}^p(\Omega)} \le (2a)^p \left\|\partial_{x_1} v\right\|^p_{\mathrm{L}^p(\Omega)},$$

hence

$$\|v\|_{\mathrm{L}^p(\Omega)} \le 2a \left\|\partial_{x_1} v\right\|_{\mathrm{L}^p(\Omega)}.$$

Finally, recall the definition of the Sobolev seminorm:

$$|v|_{1,p} = \left(\sum_{i=1}^{d} \left\|\partial_{x_i} v\right\|^p_{\mathrm{L}^p(\Omega)} \right)^{1/p}.$$

Since each term in the sum is nonnegative, in particular $\left\|\partial_{x_1} v\right\|_{\mathrm{L}^p(\Omega)} \le |v|_{1,p}$. Combining with the previous inequality yields

$$\|v\|_{\mathrm{L}^p(\Omega)} \le 2a \, |v|_{1,p}.$$

Thus the Poincaré inequality holds for all $v \in \mathcal{C}_0^\infty(\Omega)$ with $C_{\mathrm{P}} = 2a$, and by Step 1 it holds for all $v \in \mathrm{W}_0^{1,p}(\Omega)$. □

By succeessive application of the Poincare inequality shows that the semi-norm $|\cdot|_{k,p}$ is equivalent to the norm $\|\cdot\|_{k,p}$ on $\mathrm{W}_0^{k,p}(\Omega)$, i.e.,

$$|v|_{k,p} \le \|v\|_{k,p} \le C|v|_{k,p} \quad \forall \;\; v \in \mathrm{W}_0^{k,p}(\Omega).$$

Why are semi-norms important? Recall the Poisson equation with pure Neumann boundary conditions: here, the choice of norm is crucial. If we take the L^2-norm or the full H^1-norm, constants still remain an issue since these norms do not annihilate them, and hence the bilinear form is not coercive. In this case, the H^1 semi-norm makes sense as it vanishes on constants. Moreover, the H^1 semi-norm quantifies the gradient, which is useful for measuring the irregularity of solutions.

3.2.1 Sobolev Embedding Theorem

We know that certain L^p spaces are contained within others. This also holds for Sobolev spaces. Some inclusions are straightforward:

$$W^{m,p}(\Omega) \subset W^{k,p}(\Omega) \quad \text{for } m \geq k,$$

with k, m non-negative integers. Another one is

$$W^{k,q}(\Omega) \subset W^{k,p}(\Omega) \quad \text{for } 1 \leq p \leq q \leq \infty,$$

since $L^q \subset L^p$ when $q \geq p$.

But there are more subtle inclusions. Before stating them we define what a *continuous embedding* is.

Definition 3.63. (**Continuous Embedding**) Let X and Y be two normed spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ respectively. We say X is *continuously embedded* into Y if $X \subset Y$ and there exists a constant $C > 0$ such that

$$\|x\|_Y \leq C\|x\|_X \quad \forall\, x \in X. \tag{3.1}$$

We denote the embedding by $X \hookrightarrow Y$.

Theorem 3.64. (**Sobolev Embedding Theorem**) [16, Theorem 1.5] *Let $\Omega \subset \mathbb{R}^d$ be a domain with Lipschitz boundary. Then, for $k \geq 0$ and $1 \leq p \leq \infty$, the following continuous embeddings hold:*

$$\begin{aligned}
W^{k,p}(\Omega) &\hookrightarrow L^{p^*}(\Omega) \quad \text{with } \frac{1}{p^*} = \frac{1}{p} - \frac{k}{d}, \quad \text{if } k < \frac{d}{p},\\
W^{k,p}(\Omega) &\hookrightarrow L^{q}(\Omega) \quad \forall\, q < \infty, \quad \text{if } k = \frac{d}{p},\\
W^{k,p}(\Omega) &\hookrightarrow \mathcal{C}^{0,k-d/p}(\overline{\Omega}) \quad \text{if } \frac{d}{p} < k < \frac{d}{p} + 1,\\
W^{k,p}(\Omega) &\hookrightarrow \mathcal{C}^{0,\alpha}(\overline{\Omega}) \quad \forall\, \alpha \in (0,1), \quad \text{if } k = \frac{d}{p} + 1,\\
W^{k,p}(\Omega) &\hookrightarrow \mathcal{C}^{0,1}(\overline{\Omega}) \quad \text{if } k > \frac{d}{p} + 1.
\end{aligned}$$

The last three embeddings are into spaces of Hölder continuous functions (see Definition 1.4). Note that L^p is a much larger space than $\mathcal{C}(\overline{\Omega})$, so in the last three cases one modifies

a function $v \in \mathrm{W}^{k,p}(\Omega)$ on a set of measure zero to obtain a continuous representative $\tilde{v}$. In practice, we do not distinguish between v and $\tilde{v}$.

Why are these embeddings important? They tell us about the additional regularity that functions in Sobolev spaces possess. For example, let $d = 1$, $p = 2$, and $k = 1$. Then $d/p = 1/2$, so $k = 1 > d/p$ and hence

$$\mathrm{H}^1(\Omega) \hookrightarrow \mathcal{C}^{0,1/2}(\overline{\Omega}).$$

Thus one-dimensional H^1 functions are continuous. If $d = 2$, then $d/p = 1$ and $k = 1 = d/p$, so

$$\mathrm{H}^1(\Omega) \hookrightarrow \mathrm{L}^q(\Omega), \quad \forall\, q < \infty,$$

but not into continuous functions. This shows the critical nature of Sobolev embeddings.

Example 3.65. Consider

$$f(r) = \log\big(-\log r\big)$$

on the domain

$$\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : \; x_1^2 + x_2^2 < 1/2\},$$

where $r = \sqrt{x_1^2 + x_2^2}$. Clearly $f(r)$ is *not continuous* at $r = 0$ since $\log(-\log r) \to +\infty$ as $r \to 0$.
However, using polar coordinates, we compute the Sobolev seminorm:

$$|f|_{1,2}^2 = \int_\Omega |\nabla f|^2 \, d\mathbf{x} = \int_0^{2\pi} \int_0^{1/2} \left|\frac{\partial f}{\partial r}\right|^2 r \, dr \, d\varphi = 2\pi \int_0^{1/2} \frac{dr}{r(\log r)^2} < \infty.$$

Hence $f \in H^1(\Omega)$ despite the singularity at $r = 0$.

Remark 3.66. The critical nature of Sobolev embeddings in three dimensions is also reflected in the Millennium problem for the Navier–Stokes equations. In 3D, the embedding $\mathrm{H}^1(\mathbb{R}^3) \hookrightarrow \mathrm{L}^6(\mathbb{R}^3)$ is just sufficient to control certain nonlinear terms, but not strong enough to guarantee global existence and uniqueness of solutions. This borderline embedding is one of the reasons the global regularity problem remains open.

3.2.2 Trace

The reason we are studying Sobolev spaces is to solve Poisson boundary value problems. Now the boundary of $\Omega \subset \mathbb{R}^d$ is in $\mathbb{R}^{d-1}$ and hence its d-measure is zero (for example, the area of a line is zero). But one still needs a way to represent the boundary values by a function defined on Ω.

Let us take an example. Say $\Omega = (0,1)^2$ and $u(x_1, x_2) = x_1^{-\alpha/2}$ with $0 < \alpha < 1$. Now,

$$\begin{aligned} \|u\|_2^2 &= \int_\Omega (x_1)^{-\alpha} \, d\mathbf{x} \\ &= \int_0^1 \int_0^1 x_1^{-\alpha} \, dx_2 dx_1 \\ &= \frac{1}{1-\alpha} < \infty. \end{aligned}$$

Hence $u \in \mathrm{L}^2(\Omega)$ but $u|_{x_1=0} = \infty$. Therefore we need a way of defining the boundary values.

Theorem 3.67. (**Trace Theorem**)[13, Theorem B.52],[1, Theorem 5.36] *Let* $1 \leq p < \infty$ *and* $\Omega \subset \mathbb{R}^d$ *be a bounded domain.*

- *If* Ω *has* Lipschitz boundary Γ*, then the trace operator*

$$\mathrm{tr} : \mathrm{W}^{1,p}(\Omega) \to \mathrm{W}^{1-1/p,p}(\Gamma)$$

is bounded and surjective. Moreover,

$$\mathrm{W}_0^{1,p}(\Omega) = \{v \in \mathrm{W}^{1,p}(\Omega) : \mathrm{tr}(v) = 0\}.$$

- *More generally, if* Ω *has* $\mathcal{C}^{m,\alpha}$ *boundary, then for* $m \geq 1$*,*

$$\mathrm{tr} : \mathrm{W}^{m,p}(\Omega) \to \mathrm{W}^{m-1/p,p}(\Gamma)$$

is bounded and surjective.

Now for $p = 2$ the trace operator is

$$\mathrm{tr} : \mathrm{H}^1(\Omega) \to \mathrm{H}^{1/2}(\Gamma),$$

and as it is surjective every function in $\mathrm{H}^{1/2}(\Gamma)$ is the trace of a function in $\mathrm{H}^1(\Omega)$. This is useful in the case we seek solutions to PDEs that satisfy given values on Γ.

Corollary 3.68. [16, Corollary 1.1] *Let* $1 \leq p < \infty$ *and* $\Omega \subset \mathbb{R}^d$ *be a bounded domain with Lipschitz continuous boundary* Γ*. Then there exists a positive constant* C *such that for all* $g \in \mathrm{W}^{1-1/p,p}(\Gamma)$ *there exists* $u_g \in \mathrm{W}^{1,p}(\Omega)$ *satisfying*

$$\mathrm{tr}(u_g) = g \quad \text{and} \quad \|u_g\|_{1,p,\Omega} \leq C\|g\|_{1-1/p,p,\Gamma}.$$

The function u_g *is called the* lifting operator *of* g *in* $\mathrm{W}^{1,p}(\Omega)$*.*

Remark 3.69. The target space $\mathrm{W}^{m-1/p,p}(\Gamma)$ we have used is called the Sobolev–Slobodeckij space. If we drop the smoothness condition of $\mathcal{C}^{m,\alpha}$ boundary to just Lipschitz boundary then the natural target becomes a larger space called a *Besov space*.

Now, in the last result we have introduced half spaces like $\mathrm{H}^{1/2}(\Gamma)$. But what is half differentiability? Fractional spaces are quite an advanced topic and are clearly out of scope of this work. For completeness we mention them, but the only important thing to remember is that the solution loses $1/p$ regularity on the boundary.

For $1 \leq p < \infty$ and $0 < \sigma < 1$ we define

$$\mathrm{W}^{\sigma,p}(\Omega) = \left\{u \in \mathrm{L}^p(\Omega) : \ [u]_{\mathrm{W}^{\sigma,p}(\Omega)} < \infty\right\},$$

where

$$[u]^p_{\mathrm{W}^{\sigma,p}(\Omega)} = \int_\Omega \int_\Omega \frac{|u(\mathbf{x}) - u(\mathbf{y})|^p}{|\mathbf{x} - \mathbf{y}|^{d+\sigma p}}\, d\mathbf{x}d\mathbf{y},$$

and we define the norm on this space by

$$\|u\|^p_{\mathrm{W}^{\sigma,p}(\Omega)} = \|u\|^p_{\mathrm{L}^p(\Omega)} + [u]^p_{\mathrm{W}^{\sigma,p}(\Omega)}.$$

In the case of a general exponent $s \in \mathbb{R}^+$, we write $s = m + \sigma$ with integer m and $0 < \sigma < 1$, and define the space

$$\mathrm{W}^{s,p}(\Omega) = \{u \in \mathrm{W}^{m,p}(\Omega) : \quad D^{\boldsymbol{\alpha}} u \in \mathrm{W}^{\sigma,p}(\Omega) \ \ \forall\, |\boldsymbol{\alpha}| = m\},$$

with norm

$$\|u\|^p_{\mathrm{W}^{s,p}(\Omega)} = \|u\|^p_{\mathrm{W}^{m,p}(\Omega)} + \sum_{|\boldsymbol{\alpha}|=m} \|D^{\boldsymbol{\alpha}} u\|^p_{\mathrm{W}^{\sigma,p}(\Omega)}.$$

For an in-depth study we suggest the paper [11].

Remark 3.70. The last thing we want to mention about Sobolev spaces is their dual. Let $1 \leq p < \infty$ and q be the conjugate exponent of p. Let $\Omega \subset \mathbb{R}^d$. Then the dual of $\mathrm{W}^{k,p}_0(\Omega)$ is denoted by $\mathrm{W}^{-k,q}(\Omega)$. In particular, if $p = 2$ then $\mathrm{H}^{-k}(\Omega)$ is the dual of $\mathrm{H}^k_0(\Omega)$. Negative Sobolev spaces are often defined using Fourier transforms and in practice measure how non-smooth the functions are. We will revisit this topic later. The important thing to note is the hierarchy of spaces

$$\mathrm{H}^1_0(\Omega) \hookrightarrow \mathrm{L}^2(\Omega) \hookrightarrow \mathrm{H}^{-1}(\Omega).$$

3.2.3 Gauss and Green Formulas

One key aspect that we will use further is the generalization of the integration by parts formula. Here the regularity of the domain and the solution becomes very important.

Theorem 3.71. (**Gauss Theorem**) *Let $\Omega \subset \mathbb{R}^d$ for $d \geq 2$ be a bounded domain with Lipschitz boundary Γ. Then for $u \in \mathrm{W}^{1,1}(\Omega)$,*

$$\int_\Omega \partial_i u(\mathbf{x})\, d\mathbf{x} = \int_\Gamma u(s)\, \mathbf{n}_i(s)\, ds,$$

where $\mathbf{n} = (\mathbf{n}_1, \ldots, \mathbf{n}_d)$ is the unit outward normal vector to Γ.

Corollary 3.72. *Let the conditions of Theorem 3.71 on Ω be satisfied. Consider $u \in \mathrm{W}^{1,p}(\Omega)$ and $v \in \mathrm{W}^{1,q}(\Omega)$ with $p \in (1, \infty)$ and $1/p + 1/q = 1$. Then*

$$\int_\Omega \partial_i u(\mathbf{x})\, v(\mathbf{x})\, d\mathbf{x} = \int_\Gamma u(s)\, v(s)\, \mathbf{n}_i(s)\, ds - \int_\Omega u(\mathbf{x})\, \partial_i v(\mathbf{x})\, d\mathbf{x}.$$

Proof. Since $u \in \mathrm{W}^{1,p}(\Omega)$ and $v \in \mathrm{W}^{1,q}(\Omega)$, we have $uv \in \mathrm{W}^{1,1}(\Omega)$ (by Hölder's inequality). Applying the product rule and Theorem 3.71 to uv yields the result. □

From the above corollary we obtain Green's formula, first given by the British mathematical physicist George Green in 1828.

Corollary 3.73. (**Green's Formula**)[13, Corollary B.59] *Let the conditions of Theorem 3.71 on Ω be satisfied. Then*

$$\int_\Omega \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x})\, d\mathbf{x} = \int_\Gamma \frac{\partial u}{\partial \mathbf{n}}(s)\, v(s)\, ds - \int_\Omega \Delta u(\mathbf{x})\, v(\mathbf{x})\, d\mathbf{x},$$

for all $u \in \mathrm{H}^2(\Omega)$ and $v \in \mathrm{H}^1(\Omega)$.

Proof. The proof follows by summing over $i = 1, \ldots, d$ in the previous corollary. □

Figure 3.13: Carl Friedrich Gauss (30 April 1777 – 23 February 1855, left) and George Green (14 July 1793 – 31 May 1841, right).

3.2.4 Domains

We have seen certain types of domains that play an important role in defining results for PDEs. Now, we want to go a little deeper and give a mathematically rigorous definition of them.

Definition 3.74. (Lipschitz Domain) Let Ω be a bounded domain in $\mathbb{R}^d$. Then Ω is called a *Lipschitz domain* if for every $\mathbf{x} \in \Gamma$ there exists a neighbourhood U of $\mathbf{x}$ in $\mathbb{R}^d$ and new orthogonal coordinates $(y_1, \ldots, y_d)$ such that

1. U is a hypercube in the new coordinates, i.e.,

$$U = \{(y_1, \ldots, y_d) : \ -a_i < y_i < a_i, \quad i = 1, \ldots, d\}\,.$$

2. There exists a Lipschitz continuous function ϕ defined on

$$U' = \{(y_1, \ldots, y_{d-1}) : \ -a_i < y_i < a_i, \quad i = 1, \ldots, d-1\}\,,$$

such that

$$\begin{aligned} &|\phi(\mathbf{y}')| \le a_d \quad \text{for every } \mathbf{y}' = (y_1, \ldots, y_{d-1}) \in U', \\ &\Omega \cap U = \{\mathbf{y} = (\mathbf{y}', y_d) \in U : \ y_d < \phi(\mathbf{y}')\}\,, \\ &\Gamma \cap U = \{\mathbf{y} = (\mathbf{y}', y_d) \in U : \ y_d = \phi(\mathbf{y}')\}\,. \end{aligned}$$

Graphically, this means that for every $\mathbf{x} \in \Gamma$ there exists a local coordinate system and an open hypercube U (an interval in 1D, an open square in 2D, an open cube in 3D, etc.) such that the boundary can be represented as the graph of a Lipschitz continuous function ϕ over U', and the domain Ω locally lies on one side of this graph. In other words, $\Gamma \cap U$ corresponds

to the points on the graph $y_d = \phi(\mathbf{y}')$, while $\Omega \cap U$ corresponds to the points lying below it (i.e., $y_d < \phi(\mathbf{y}')$). This is illustrated in Fig. 3.14.

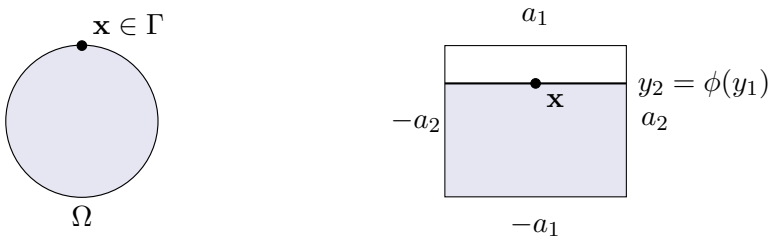


Figure 3.14: Localization of a boundary point. Left: global domain Ω and a boundary point $\mathbf{x} \in \Gamma$. Right: the zoomed neighbourhood U where the boundary is represented locally as the graph $y_2 = \phi(y_1)$ and $\Omega \cap U$ lies below the graph.

Example 3.75. 1. Domains such as balls and polygons in 2D are Lipschitz.

2. Consider the domain

$$\Omega = \left\{(x, y) : \quad x^2 + y^2 < 1\right\} \setminus \{(x, y) : \quad x \geq 0, \quad y = 0\},$$

i.e., a domain with slit. Then this domain is not Lipschitz as for any hypercube along the slit will have domain on both the sides.

3. In three dimensions, a polyhedral domain need not be Lipschitz. A simple example is shown in Fig. 3.15, where two rectangular blocks meet at a right angle. Near a point on the common edge, the boundary is made up of two perpendicular surface patches. This means that for the same base point $y' \in \mathbb{R}^2$, there are two different possible boundary heights y_3. Thus the boundary cannot be represented locally as the graph of a single Lipschitz function $y_3 = \varphi(y')$. For this reason, the Lipschitz condition fails precisely along such non-smooth edges.

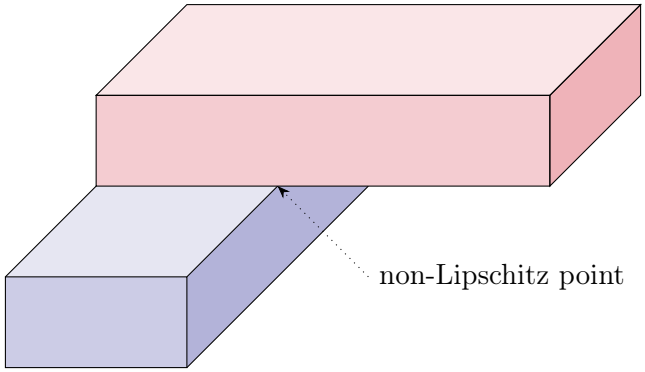


Figure 3.15: A polyhedral domain in 3D that is not Lipschitz: two blocks meet at a right angle, producing a non-Lipschitz point.

If you recall from Sec. 1.3, we defined Hölder continuous functions. The hierarchy of spaces is

$$\mathcal{C}^1(\Omega) \subset \mathcal{C}^{0,1}(\Omega) \subset \mathcal{C}^{0,\alpha}(\Omega) \subset \mathcal{C}^0(\Omega),$$

where $0 < \alpha < 1$. Theorem 1.7 requires a $\mathcal{C}^{2,\alpha}$ boundary, which means the function ϕ in Definition 3.74 must be $\mathcal{C}^{2,\alpha}$ as a function of $\mathbb{R}^{d-1}$. This excludes many important domains —

the square being the most obvious example. Therefore, in applications we usually restrict our analysis to Lipschitz domains.

Remark 3.76. In the above inclusion chain, the space $\mathcal{C}^{0,\alpha}(\Omega)$ may appear "larger" than $\mathcal{C}^{0,1}(\Omega)$, but *larger is not always better*. A typical example is $|x|^\alpha$ for $0 < \alpha < 1$. Although this function is $\mathcal{C}^{0,\alpha}$, it is not Lipschitz: near $x = 0$ its slope blows up, producing a cusp. At such points normals and tangents are not well-defined, so many geometric results (e.g. the Gauss theorem or trace theorems) fail.

3.3 Fixed Points

The final idea that we want to mention is the idea of *fixed points*, which plays a critical role in showing the existence and uniqueness of solutions.

Let us consider the following problem in a Banach space X with an operator $P : X \to X$: Find $x \in X$ such that

$$x = Px. \tag{3.2}$$

The solution $x \in X$ of the above problem is called a *fixed point* of the operator P.

Definition 3.77. (**Contraction**) The mapping $P : X \to X$ is called a *contraction* if there exists a constant $\rho < 1$ such that

$$\|Px_1 - Px_2\|_X \leq \rho \|x_1 - x_2\|_X \quad \forall\, x_1, x_2 \in X.$$

Now, we state one of the most important results regarding the existence and uniqueness of fixed points.

Theorem 3.78. (**Banach Fixed Point Theorem**) *Let X be a Banach space and $P : X \to X$ be a contraction. Then:*

1. *There exists a unique fixed point $x^* \in X$ solving Eq. (3.2).*
2. *The sequence $x_n = Px_{n-1}$, $n \geq 1$, converges to the fixed point $x^* \in X$ for any initial guess $x_0 \in X$.*
3. *The following error estimate holds:*

$$\|x_n - x^*\|_X \leq \frac{\rho^n}{1-\rho} \|Px_0 - x_0\|_X \quad \forall\, n \in \mathbb{N}.$$

Proof. We start by showing that the sequence $\{x_n\}$ is Cauchy. Observe that

$$\begin{aligned} \|x_i - x_{i-1}\|_X &= \|Px_{i-1} - Px_{i-2}\|_X \\ &\leq \rho \|x_{i-1} - x_{i-2}\|_X \\ &\leq \rho^2 \|x_{i-2} - x_{i-3}\|_X \\ &\vdots \\ &\leq \rho^{i-1} \|Px_0 - x_0\|_X \quad \forall\, i \in \mathbb{N}. \end{aligned}$$

Hence, for $m > n \geq 1$,

$$\begin{aligned}
\|x_m - x_n\|_X &= \Big\| \sum_{i=n+1}^{m} (x_i - x_{i-1}) \Big\|_X \\
&\leq \sum_{i=n+1}^{m} \|x_i - x_{i-1}\|_X \\
&\leq \sum_{i=n+1}^{m} \rho^{i-1} \|Px_0 - x_0\|_X \\
&= \|Px_0 - x_0\|_X \sum_{i=n+1}^{m} \rho^{i-1} \\
&\leq \rho^n \frac{1-\rho^{m-n}}{1-\rho} \|Px_0 - x_0\|_X \\
&\leq \frac{\rho^n}{1-\rho} \|Px_0 - x_0\|_X \longrightarrow 0,
\end{aligned}$$

as $n \to \infty$. Hence $\{x_n\}$ is a Cauchy sequence in X, and since X is complete, it converges to some $x^* \in X$.

We claim that this x^* is a fixed point of P. Indeed,

$$\begin{aligned}
\|x^* - Px^*\|_X &\leq \|x^* - x_n\|_X + \|x_n - Px^*\|_X \\
&= \|x_n - x^*\|_X + \|Px_{n-1} - Px^*\|_X \\
&\leq \|x_n - x^*\|_X + \rho \|x_{n-1} - x^*\|_X \longrightarrow 0,
\end{aligned}$$

as $n \to \infty$. Hence $x^* = Px^*$.

Next, we show uniqueness. Suppose x_1^*, x_2^* are fixed points. Then

$$\|x_1^* - x_2^*\|_X = \|Px_1^* - Px_2^*\|_X \leq \rho \|x_1^* - x_2^*\|_X.$$

Since $\rho < 1$, this implies $\|x_1^* - x_2^*\|_X = 0$, i.e., $x_1^* = x_2^*$.

Finally, for the error estimate, let $m > n$:

$$\begin{aligned}
\|x_n - x^*\|_X &\leq \|x_n - x_m\|_X + \|x_m - x^*\|_X \\
&\leq \frac{\rho^n}{1-\rho} \|Px_0 - x_0\|_X + \|x_m - x^*\|_X.
\end{aligned}$$

Letting $m \to \infty$ gives the desired bound. □

This concludes the discussion of the functional analytic concepts that are required for the numerical analysis of PDEs. Of course, the theory of partial differential equations is a vast subject, and this chapter by no means provides a complete coverage. For a more in-depth treatment, we refer the reader to the books [14, 7].

Chapter 4

Weak Solution Theory

In the last chapter, we introduced the notion of weak derivatives and Sobolev spaces. The reason for this is to generalize the notion of a solution of a differential equation to that of a weak solution. Classical solutions also suffer from strong regularity assumptions. Moreover, there are PDEs which do not admit a classical solution and only possess a weak solution, for example, Tsirelson's stochastic differential equation, named after the Russian-Israeli mathematician Boris Semyonovich Tsirelson.

Figure 4.1: Boris Semyonovich Tsirelson: 4 May 1950 – 21 January 2020.

Hence, there is a need to study weak solutions. This also motivates the development of numerical methods for weak solutions. In this chapter, we will introduce the setting for computing weak solutions, establish their existence and uniqueness, and finally provide a theoretical framework for the finite element method.

4.1 Variational Formulation

We revisit the Poisson equation in one dimension defined over $[0, 1]$. Let us consider the two-point boundary value problem

$$\begin{aligned} -u''(x) &= f(x), \quad 0 < x < 1, \\ u(0) &= u(1) = 0. \end{aligned} \tag{D}$$

where $f(x)$ is a given continuous function. We call this a boundary value problem and denote it by (D). Since $f(x)$ is continuous, we can compute the solution easily by integrating the equation twice.

Let us now consider another problem, which is a minimization problem: find $u \in V$ such that $F(u) \leq F(v)$ for all $v \in V$, where

$$F(u) = \frac{1}{2}(u', u') - (f, u),$$

is an energy functional from V to $\mathbb{R}$, and V is a vector space defined by

$$V = \{v : v \in \mathcal{C}[0,1], \quad v'(x) \text{ is piecewise continuous and bounded on } [0,1], \quad \text{and} \quad v(0) = v(1) = 0\}.$$

We denote this problem by (M).

Now, let us look at a third problem. Find $u \in V$ such that

$$(u', v') = (f, v) \quad \forall\ v \in V,$$

where $(\cdot, \cdot)$ is the L^2 inner product. We denote this problem by (V), and the reason for choosing this formulation will be made clear later.

In physics, $F(v)$ represents the total potential energy associated with the displacement $v \in V$. The minimization problem (M) corresponds to the fundamental principle of minimum potential energy in mechanics, and the problem (V) corresponds to the principle of virtual work.

We claim that the solution u of (D) is also a solution to all three problems. We first start by showing that if u solves (D), then u solves (V), which we denote by $(D) \Rightarrow (V)$. Suppose u solves (D). To show that u solves (V), we multiply Eq. (D) by a function $v \in V$ (called a *test function*) and integrate over $(0,1)$. Then,

$$-\int_0^1 u''(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx.$$

Using integration by parts and the fact that $v(0) = v(1) = 0$, we obtain

$$-\int_0^1 u''(x)v(x)\,dx = -u(x)v(x)\big|_0^1 + \int_0^1 u'(x)v'(x)\,dx = (u', v').$$

Hence u solves the problem (V).

Next, we show $(V) \Leftrightarrow (M)$. Let $u \in V$ be a solution of (V) and let $v \in V$. Set $w = v - u$, so that $v = u + w$. Now,

$$\begin{aligned} F(v) &= F(u + w) \\ &= \frac{1}{2}(u' + w', u' + w') - (f, u + w) \\ &= \frac{1}{2}(u', u') + (u', w') + \frac{1}{2}(w', w') - (f, u) - (f, w). \end{aligned}$$

Since $(u', w') = (f, w)$, we get

$$\begin{aligned} F(v) &= \frac{1}{2}(u', u') + \frac{1}{2}(w', w') - (f, u) \\ &= F(u) + \frac{1}{2}\|w'\|_{\mathrm{L}^2}^2 \geq F(u). \end{aligned}$$

As $v \in V$ was arbitrary, we conclude that $u \in V$ is the solution of (M).

Conversely, let $u \in V$ be the solution of (M). Then for any $v \in V$ and $\varepsilon \in \mathbb{R}$,

$$F(u) \leq F(u + \varepsilon v),$$

since $u + \varepsilon v \in V$. Thus the function

$$g(\varepsilon) = F(u + \varepsilon v) = \frac{1}{2}(u' + \varepsilon v', u' + \varepsilon v') - (f, u + \varepsilon v)$$

has a minimum at $\varepsilon = 0$ and hence $g'(0) = 0$. Now

$$g'(\varepsilon) = (u', v') + \varepsilon (v', v') - (f, v),$$

and $g'(0) = 0$ implies

$$(u', v') = (f, v).$$

Hence, u is a solution of (V). In fact, u is the unique solution, which can be verified as an exercise.

So far, we have shown the relation $(D) \Rightarrow (V) \Leftrightarrow (M)$. Finally, we must show that if u is the solution of (V), then u is also the solution of (D). Let $u \in V$ satisfy

$$\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx \quad \forall\ v \in V.$$

Assume $u''(x)$ exists and is continuous. Then, using integration by parts on the first term and the fact that $v(0) = v(1) = 0$, we obtain

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

As $u''(x)$ is continuous, we get

$$-\int_0^1 (u'' + f)(x)v(x)\,dx = 0, \quad \forall\ v \in V.$$

Therefore, $(u'' + f)(x) = 0$, and hence $u(x)$ is the solution.

Thus, we have three equivalent formulations of the same problem, but the problems (V) and (M) have some nice properties: they only depend on $u'(x)$, and therefore we can work with less regular spaces. The problem (V) is referred to as the *variational problem*, as we are varying over the space V. This gives the motivation to study problem (V) or (M) instead of (D), and forms the starting point of the theory of weak solutions.

4.1.1 Symmetric Problems

Before moving forward, let us introduce certain definitions and notations.

Definition 4.1. Bilinear Form A bilinear form $a(\cdot, \cdot)$ on a vector space V is a mapping $V \times V \to \mathbb{R}$ such that each of the maps $v \mapsto a(v, w)$ and $w \mapsto a(v, w)$ is linear on V. We say it is symmetric if

$$a(v, w) = a(w, v) \quad \forall\ v, w \in V.$$

Example 4.2.
1. Let $V = \mathbb{R}^2$. Then $a(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ is a symmetric bilinear form on $\mathbb{R}^2$.
2. Let $V = \mathbb{R}^2$. Then $a(\mathbf{x}, \mathbf{y}) = x_1 y_2$ is a non-symmetric bilinear form. For example, with $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, we have $a(\mathbf{x}, \mathbf{y}) = 1$ while $a(\mathbf{y}, \mathbf{x}) = 0$.

It is easy to see that an inner product is a symmetric bilinear form. As we know that an inner product defines a norm on the space, a symmetric bilinear form also induces a norm on the space, denoted by $\|\cdot\|_a$, i.e.,

$$\|v\|_a = \sqrt{a(v, v)}.$$

Let V be a Hilbert space and $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ be a *symmetric* bilinear form which is an inner product on V. Let $f \in V^*$, i.e., the dual space of V. Then we seek $u \in V$ such that

$$a(u, v) = \langle f, v \rangle \quad \forall\ v \in V.$$

Theorem 4.3. (**Existence and Uniqueness of Solution**) *Let $f \in V^*$. Then there exists a uniquely determined $u \in V$ with*

$$a(u, v) = \langle f, v \rangle \quad \forall\ v \in V,$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing.

Proof. The proof follows directly from the Riesz Representation Theorem 3.44. □

4.1.2 Non-Symmetric Problems

Now, in the case $a(\cdot, \cdot)$ is not symmetric, we still have existence and uniqueness of the solution, provided the bilinear form satisfies certain properties.

Definition 4.4. (**Bounded and Coercive**) A bilinear form $a(\cdot, \cdot)$ on a normed space V is said to be *bounded* (or continuous) if there exists $M < \infty$ such that

$$|a(u, v)| \leq M \|u\|_V \|v\|_V \quad \forall\ u, v \in V.$$

The bilinear form is said to be V-elliptic (or coercive) on V if there exists $m > 0$ such that

$$a(u, u) \geq m \|u\|_V^2 \quad \forall\ u \in V,$$

where M is independent of u, v and m is independent of u.

Theorem 4.5. (**Lax–Milgram Theorem**) *Given a Hilbert space $(V, (\cdot,\cdot))$, a continuous coercive bilinear form $a(\cdot,\cdot)$, and a continuous linear functional $f \in V^*$, there exists a unique $u \in V$ such that*

$$a(u, v) = \langle f, v\rangle \quad \forall\ v \in V. \tag{4.1}$$

Proof. The proof uses the Banach fixed point theorem 3.78. For any $u \in V$, define a functional A_u by

$$\langle A_u, v\rangle = a(u, v) \quad \forall\ v \in V.$$

Then A_u is linear, since $a(\cdot,\cdot)$ is bilinear. Moreover, A_u is continuous, since

$$|\langle A_u, v\rangle| = |a(u, v)| \le M\|u\|_V\|v\|_V.$$

Therefore,

$$\|A_u\|_{V^*} = \sup_{v\neq 0} \frac{|\langle A_u, v\rangle|}{\|v\|_V} \le M\|u\|_V < \infty,$$

hence $A_u \in V^*$.

Equation (4.1) can then be written as: find $u \in V$ such that

$$\begin{aligned} \langle A_u, v\rangle &= \langle f, v\rangle \quad \forall\ v \in V, \\ \langle A_u - f, v\rangle &= 0 \quad \forall\ v \in V. \end{aligned}$$

Hence,

$$A_u = f. \tag{4.2}$$

By the Riesz Representation Theorem, for all $\varphi \in V^*$ there exists a unique $\tau_\varphi \in V$ such that

$$\langle \varphi, v\rangle = (\tau_\varphi, v) \quad \forall\ v \in V. \tag{4.3}$$

Thus instead of finding the functional A_u in Eq. (4.2), we compute its Riesz representer, i.e.,

$$\tau_{A_u} = \tau_f,$$

since $\tau : V^* \to V$ is one-to-one. Hence the problem reduces to: find $u \in V$ such that $\tau_{A_u} = \tau_f$.

Define a mapping $T : V \to V$ by

$$T(v) = v - \rho\left(\tau_{A_v} - \tau_f\right),$$

where $\rho \neq 0$ is a constant. If T is a contraction, then by the Banach fixed point theorem there exists a unique $u \in V$ such that

$$\begin{aligned} Tu &= u, \\ u - \rho\left(\tau_{A_u} - \tau_f\right) &= u, \\ \tau_{A_u} &= \tau_f. \end{aligned}$$

It remains to show that T is a contraction. Let $v_1, v_2 \in V$ and set $v = v_1 - v_2$. Then, using linearity of τ and A, Eq. (4.3), and the boundedness and coercivity of $a(\cdot,\cdot)$, we obtain

$$\begin{aligned} \|Tv_1 - Tv_2\|_V^2 &= \|v - \rho\,(\tau_{A_v})\,\|_V^2 \\ &= \|v\|_V^2 + \rho^2\|\tau_{A_v}\|_V^2 - 2\rho(\tau_{A_v}, v) \\ &= \|v\|_V^2 - 2\rho\, a(v,v) + \rho^2\, a(v, \tau_{A_v}) \\ &\leq \|v\|_V^2 - 2\rho m\|v\|_V^2 + \rho^2 M\|v\|_V\|\tau_{A_v}\|_V \\ &\leq \left(1 - 2\rho m + \rho^2 M^2\right)\|v\|_V^2. \end{aligned}$$

Thus, if $1 - 2\rho m + \rho^2 M^2 < 1$, i.e., if there exists ρ such that $\rho M^2 - 2m < 0$, then T is a contraction. Such a ρ always exists, and hence the proof is complete. □

Hence we have existence and uniqueness of the solution for both symmetric and non-symmetric problems.

Remark 4.6. It follows from the coercivity of the bilinear form that

$$\|u\|_V \leq \frac{1}{m}\|f\|_{V^*},$$

which shows that the weak form is well-posed in the sense that it has a unique solution, which depends continuously on the data f.

4.2 Weak Solution

We move back to the original problem we stated in Chapter 1, i.e., the Poisson equation. Let us recall it. We try to solve

$$\begin{aligned} -\Delta u &= f \quad \text{in} \quad \Omega, \\ u &= 0 \quad \text{on} \quad \Gamma, \end{aligned} \tag{4.4}$$

where $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain and $f \in \mathrm{L}^2(\Omega)$. Here we are assuming homogeneous Dirichlet boundary conditions, but we will also tackle the general problem. The existence and uniqueness of a classical solution depend on Ω and f (see Theorem 1.7). In reality, f may not be differentiable or continuous. For example,

$$f(x) = 1 + \mathrm{sgn}(1 - |x|),$$

which is a piecewise continuous function (see Fig. 4.2).

Let $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$ be a classical solution of Eq. (4.4) and assume $f \in \mathcal{C}(\Omega)$. Multiplying Eq. (4.4) with $v \in \mathcal{C}_0^\infty(\Omega)$ and integrating over Ω we get

$$-\int_\Omega \Delta u\, v\, dx = \int_\Omega f v\, dx.$$

Using Corollary 3.73 and the fact that $v = 0$ on Γ, we obtain

$$\int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega f v\, dx \quad \forall \;\; v \in \mathcal{C}_0^\infty(\Omega). \tag{4.5}$$

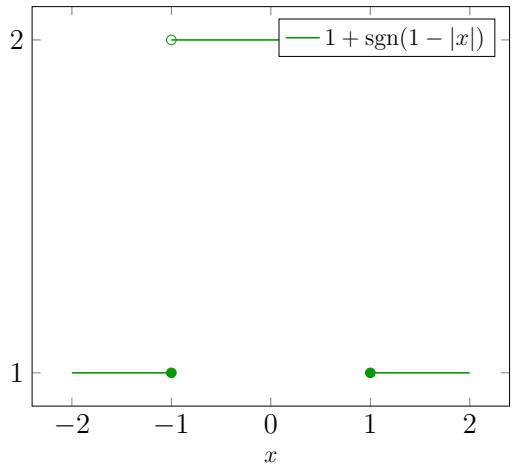


Figure 4.2: Piecewise continuous function $\notin \mathrm{L}^2(\Omega)$.

Since $\mathcal{C}_0^\infty(\Omega)$ is dense in $\mathrm{H}_0^1(\Omega)$, the solution u also satisfies Eq. (4.5) for all $v \in \mathrm{H}_0^1(\Omega)$. Hence the classical solution u need not be in $\mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$ to make sense of the integrals in Eq. (4.5); everything works for $u \in \mathrm{H}_0^1(\Omega)$ as well.

Definition 4.7. (**Weak Solution**) A function $u \in \mathrm{H}_0^1(\Omega)$ satisfying Eq. (4.5) is called a *weak solution* or a *generalized solution* of Eq. (4.4), where the partial derivatives are to be understood as weak derivatives. Eq. 4.5 is referred to as the *weak formulation* or *variational formulation*.

Hence the weak formulation of Eq. (4.4) reads as: Find $u \in V = \mathrm{H}_0^1(\Omega)$ such that

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx := F(v) \quad \forall \, v \in V. \tag{4.6}$$

Now it is easy to see that $a(\cdot, \cdot)$ is a bilinear form from $V \times V \to \mathbb{R}$ and $F(\cdot)$ is a linear form from $V \to \mathbb{R}$. Hence, for the existence and uniqueness of the solution we use the Lax-Milgram theorem.

Theorem 4.8. (**Existence and Uniqueness Theorem**) *There exists a unique weak solution $u \in V$ for Eq. (4.4).*

Proof. We need to show that the bilinear form $a(\cdot, \cdot)$ and $F(\cdot)$ satisfy the conditions of the Lax-Milgram theorem, i.e., $a(\cdot, \cdot)$ needs to be continuous and coercive, and $F(\cdot)$ needs to be continuous on $V = \mathrm{H}_0^1(\Omega)$. We recall that the norm on V is $\|\cdot\|_1 = \|\cdot\|_{\mathrm{H}_0^1(\Omega)}$, which we denote by $\|\cdot\|_V$, and is given by

$$\|u\|_V^2 = \|u\|_0^2 + \|\nabla u\|_0^2 = \|u\|_{\mathrm{L}^2(\Omega)}^2 + \|\nabla u\|_{\mathrm{L}^2(\Omega)}^2.$$

Now, using the Cauchy–Schwarz inequality for integrals and sums we get

$$\begin{aligned}
|a(u,v)| = \left| \int_\Omega \nabla u \cdot \nabla v \, dx \right| &= \left| \int_\Omega \sum_{i=1}^{d} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx \right| \\
&\leq \sum_{i=1}^{d} \int_\Omega \left| \frac{\partial u}{\partial x_i} \right| \left| \frac{\partial v}{\partial x_i} \right| dx \\
&\leq \sum_{i=1}^{d} \left(\int_\Omega \left| \frac{\partial u}{\partial x_i} \right|^2 dx \right)^{1/2} \left(\int_\Omega \left| \frac{\partial v}{\partial x_i} \right|^2 dx \right)^{1/2} \\
&\leq \left(\sum_{i=1}^{d} \int_\Omega \left| \frac{\partial u}{\partial x_i} \right|^2 dx \right)^{1/2} \left(\sum_{i=1}^{d} \int_\Omega \left| \frac{\partial v}{\partial x_i} \right|^2 dx \right)^{1/2} \\
&= \|\nabla u\|_0 \, \|\nabla v\|_0 \leq \|u\|_V \|v\|_V.
\end{aligned}$$

Hence $a(\cdot,\cdot)$ is continuous with $M = 1$. For coercivity we notice that

$$a(u,u) = \|\nabla u\|_0^2 = \int_\Omega \sum_{i=1}^{d} \left| \frac{\partial u}{\partial x_i} \right|^2 dx.$$

Now from the Poincaré inequality 3.62 we have

$$\begin{aligned}
\|u\|_0^2 &\leq C_{\mathrm{P}}^2 \|\nabla u\|_0^2, \\
\|u\|_0^2 + \|\nabla u\|_0^2 &\leq \left(C_{\mathrm{P}}^2 + 1\right) \|\nabla u\|_0^2, \\
\|\nabla u\|_0^2 &\geq \frac{1}{1 + C_{\mathrm{P}}^2} \|u\|_V^2. \qquad (4.7)
\end{aligned}$$

Hence

$$a(u,u) \geq \frac{\|u\|_V^2}{1 + C_{\mathrm{P}}^2}.$$

Thus $a(\cdot,\cdot)$ is coercive. Finally, we show that $F(\cdot)$ is bounded, which follows from the Cauchy–Schwarz inequality and the fact that $f \in \mathrm{L}^2(\Omega)$. Indeed,

$$|F(v)| = \left| \int_\Omega f(x) v(x) \, dx \right| \leq \|f\|_0 \|v\|_0 \leq \|f\|_0 \|v\|_V.$$

Hence, the existence and uniqueness of the solution follow from the Lax-Milgram theorem. □

Remark 4.9. We assumed $f \in \mathrm{L}^2(\Omega)$ in the previous theorem. In reality, we can reduce this assumption to $f \in \mathrm{H}^{-1}(\Omega)$, where $\mathrm{H}^{-1}(\Omega)$ is the dual of $\mathrm{H}_0^1(\Omega)$. In this case we write $F(v) = \langle f, v \rangle$, and

$$\langle f, v \rangle \leq \|f\|_{V^*} \|v\|_V,$$

by Lemma 3.37, and the boundedness of $F(\cdot)$ follows.

4.2.1 General Elliptic Operator

Let us consider a more general elliptic operator and see how the continuity and coercivity of $a(\cdot,\cdot)$ is computed. Consider the generalized elliptic equation of second order with mixed boundary conditions defined on the Lipschitz domain Ω

$$\begin{aligned} -\Delta u + \mathbf{b}\cdot\nabla u + cu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_{\mathrm{D}}, \\ \frac{\partial u}{\partial \mathbf{n}} &= g \quad \text{on } \Gamma_{\mathrm{N}}, \end{aligned} \tag{4.8}$$

where $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain, $\Gamma = \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}}$, and $\Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \emptyset$. The coefficients $\mathbf{b}$ and c are assumed to be sufficiently smooth, $f \in \mathrm{L}^2(\Omega)$ and $g \in \mathrm{L}^2(\Gamma_{\mathrm{N}})$.

Now, we define a new space

$$V = \mathrm{H}^1_{\mathrm{D}}(\Omega) = \left\{ v \in \mathrm{H}^1(\Omega) : \ v|_{\Gamma_{\mathrm{D}}} = 0 \right\}.$$

Obviously, $V \subset \mathrm{H}^1(\Omega)$ and the seminorm $|\cdot|_1$ is equivalent to $\|\cdot\|_1$ (a consequence of the Poincaré inequality). We take the norm on V as $\|\cdot\|_1$, i.e., the H^1 norm.

First, we need to compute the variational formulation. Similar to the Poisson equation, we multiply Eq. (4.8) with a $v \in V$ and integrate over Ω. By integration by parts,

$$\begin{aligned} -\int_\Omega \Delta u\, v \, dx + \int_\Omega \mathbf{b}\cdot\nabla u\, v \, dx + \int_\Omega c(x) uv \, dx &= \int_\Omega fv \, dx, \\ \int_\Omega \nabla u \cdot \nabla v \, dx - \int_\Gamma \frac{\partial u}{\partial \mathbf{n}} v \, ds + \int_\Omega \mathbf{b}\cdot\nabla u\, v \, dx + \int_\Omega c(x) uv \, dx &= \int_\Omega fv \, dx, \\ \int_\Omega (\nabla u \cdot \nabla v + \mathbf{b}\cdot\nabla u\, v + cuv)\, dx &= \int_\Omega fv \, dx + \int_{\Gamma_{\mathrm{N}}} \frac{\partial u}{\partial \mathbf{n}} v \, ds. \end{aligned}$$

Keeping the unknown function u on the left, we get the weak formulation: Find $u \in V$ such that

$$a(u,v) = F(v) \quad \forall\ v \in V,$$

where

$$\begin{aligned} a(u,v) &= \int_\Omega (\nabla u \cdot \nabla v + \mathbf{b}\cdot\nabla u\, v + cuv)\, dx, \\ F(v) &= \int_\Omega fv \, dx + \int_{\Gamma_{\mathrm{N}}} gv \, ds. \end{aligned}$$

Remark 4.10. Similar to Remark 4.9, g can be in $\mathrm{H}^{-1}(\Omega)$ instead of $\mathrm{L}^2(\Omega)$, in which case the boundary integral is replaced by a duality pairing $\langle\cdot,\cdot\rangle$.

After the weak formulation, for existence and uniqueness of the solution, we need to show that $a(\cdot,\cdot)$ is continuous and coercive. The verification of bilinearity and linearity is

straightforward. Now,

$$\begin{aligned}|a(u,v)| &\le \int_\Omega \Big(|\nabla u\cdot\nabla v| + |\mathbf{b}\cdot\nabla u\, v| + |cuv|\Big)\,dx \\ &\le \sum_{i=1}^d \int_\Omega \left|\frac{\partial u}{\partial x_i}\right|\left|\frac{\partial v}{\partial x_i}\right| dx + \sum_{i=1}^d \sup_{x\in\Omega}|b_i(x)| \int_\Omega \left|\frac{\partial u}{\partial x_i}\right| |v|\,dx + \sup_{x\in\Omega}|c(x)| \int_\Omega |uv|\,dx.\end{aligned}$$

Applying the Cauchy–Schwarz inequality as in the Poisson problem, we get

$$\begin{aligned}|a(u,v)| &\le |u|_1|v|_1 + \sum_{i=1}^d \sup_{x\in\Omega}|b_i(x)| \left(\int_\Omega \left|\frac{\partial u}{\partial x_i}\right|^2 dx\right)^{1/2} \|v\|_0 + \|c\|_\infty \|u\|_0\|v\|_0 \\ &\le \|u\|_1\|v\|_1 + \left(\sum_{i=1}^d \|b_i\|_\infty^2\right)^{1/2} \|u\|_1\|v\|_0 + \|c\|_\infty\|u\|_1\|v\|_1 \\ &\le M\|u\|_V\|v\|_V,\end{aligned}$$

where $M = 1 + \left(\sum_{i=1}^d \|b_i\|_\infty^2\right)^{1/2} + \|c\|_\infty$. Hence $a(\cdot,\cdot)$ is a continuous bilinear form.

Now for coercivity, let us look at

$$\begin{aligned}a(u,u) &= \int_\Omega |\nabla u|^2 dx + \int_\Omega (\mathbf{b}\cdot\nabla u + cu)\, u\; dx \\ &= \sum_{i=1}^d \int_\Omega \left|\frac{\partial u}{\partial x_i}\right|^2 dx + \sum_{i=1}^d \int_\Omega b_i(x)\frac{\partial u}{\partial x_i} u\,dx + \int_\Omega cu^2\; dx.\end{aligned}$$

The first and third terms are easy to bound, so let us focus on the second term. Using $\partial_i(u^2) = 2u\partial_i u$ and integration by parts, we get

$$\begin{aligned}\sum_{i=1}^d \int_\Omega b_i(x)\frac{\partial u}{\partial x_i} u\,dx &= \frac{1}{2}\sum_{i=1}^d \int_\Omega b_i(x)\frac{\partial(u^2)}{\partial x_i}\,dx \\ &= \frac{1}{2}\sum_{i=1}^d \left(\int_\Gamma b_i(x)\mathbf{n}_i(x)u^2\,ds - \int_\Omega \partial_{x_i} b_i(x)\, u^2\; dx\right) \\ &= \frac{1}{2}\int_{\Gamma_\mathrm{N}} \mathbf{b}\cdot\mathbf{n}\, u^2\,ds - \frac{1}{2}\int_\Omega \nabla\cdot\mathbf{b}\, u^2\,dx,\end{aligned}$$

which leads to

$$a(u,u) = |u|_1^2 + \int_\Omega \left(c(x) - \tfrac{1}{2}\nabla\cdot\mathbf{b}\right) u^2 dx + \tfrac{1}{2}\int_{\Gamma_\mathrm{N}} (\mathbf{b}\cdot\mathbf{n})u^2 ds.$$

Now if $c(x) - \dfrac{1}{2}\nabla\cdot\mathbf{b} \ge 0$ for all $x\in\Omega$ and $\mathbf{b}\cdot\mathbf{n}\ge 0$ for all $x\in\Gamma_\mathrm{N}$, then $a(u,u)\ge |u|_1^2$.

Using the Poincaré inequality as in Eq. (4.7), we get

$$a(u,u) \ge \frac{1}{1+C_\mathrm{P}^2}\|u\|_V^2 \quad \forall\ u\in V.$$

Hence $a(\cdot,\cdot)$ is a continuous and coercive bilinear form.

Now, for proving the continuity of $F(\cdot)$ we use the Cauchy–Schwarz inequality and the continuity of the trace operator, i.e.,

$$\|u\|_{\mathrm{L}^2(\Gamma_\mathrm{N})} \leq C_\mathrm{T}\|u\|_{\mathrm{H}^1(\Omega)},$$

where C_T is a constant, to get

$$\begin{aligned}|F(v)| &= \left|\int_\Omega fv\ dx + \int_{\Gamma_\mathrm{N}} gv\ ds\right| \\ &\leq \|f\|_0\|v\|_0 + \|g\|_{0,\Gamma_\mathrm{N}}\|v\|_{0,\Gamma_\mathrm{N}} \\ &\leq \left(\|f\|_0 + C_\mathrm{T}\|g\|_{0,\Gamma_\mathrm{N}}\right)\|v\|_V \\ &\leq C\|v\|_V.\end{aligned}$$

Hence $F(\cdot)$ is continuous and therefore we get existence and uniqueness of the solution using the Lax–Milgram theorem.

Remark 4.11. We notice that the Dirichlet boundary condition needs to be imposed on the function spaces, whereas the Neumann boundary condition appears naturally in the weak formulation. This is the reason Dirichlet boundary conditions are referred to as *essential boundary conditions* whereas Neumann boundary conditions are called *natural boundary conditions* in Sec. 1.1.1.

Remark 4.12. Until now we restricted to the case $u = 0$ on Γ_D, but if $u = g$ on Γ_D, then what happens? Using Corollary 3.68 we know that there exists a lifting operator $u_g \in \mathrm{H}^1(\Omega)$ such that

$$\mathrm{tr}(u_g) = g.$$

Now, consider $w = u - u_g$. This is zero on Γ_D and hence the weak formulation for Eq. (4.4) reads: Find $w \in \mathrm{H}^1_0(\Omega)$ such that

$$\int_\Omega \nabla w \cdot \nabla v\ dx = \int_\Omega fv\ dx - \int_\Omega \nabla u_g \cdot \nabla v\ dx.$$

Hence we have a different linear functional $F(\cdot)$ given by

$$F(v) = \int_\Omega fv\ dx - \int_\Omega \nabla u_g \cdot \nabla v\ dx,$$

while the bilinear form remains the same. Therefore, we need to check the continuity of $F(\cdot)$, and we notice that

$$\begin{aligned}|F(v)| &\leq \int_\Omega |fv|dx + \int_\Omega |\nabla u_g \cdot \nabla v|dx \\ &\leq \|f\|_0\|v\|_0 + \|\nabla u_g\|_0\|\nabla v\|_0 \\ &\leq \left(\|f\|_0 + \|\nabla u_g\|_0\right)\|v\|_V.\end{aligned}$$

Hence we still have the existence and uniqueness of the solution. This same idea extends to a general elliptic operator as well.

4.3 Galerkin Methods

Until now we have looked at the existence and uniqueness of the solution for the variational problem: Find $u \in V$ such that

$$a(u, v) = \langle f, v \rangle \quad \forall\, v \in V, \tag{4.9}$$

where V is a Hilbert space and $a(\cdot, \cdot)$ is either a non-symmetric bilinear form (provided it is continuous and coercive) or a symmetric bilinear form. But in the quest for existence and uniqueness we often forget about the solution itself, and hence in numerics we care about computing it. We cannot compute a solution in an infinite space (there is no concept of infinite memory), so we need to approximate it in a finite-dimensional subspace of V.

Let V_h be a finite-dimensional subspace of V, with the discretisation parameter h used to indicate that the discrete solution obtained in V_h converges to the continuous solution as $h \to 0$. The standard Galerkin method for computing the solution consists of restricting the variational problem to V_h, i.e., we solve: Find $u_h \in V_h$ such that

$$a(u_h, v_h) = \langle f, v_h \rangle \quad \forall\, v_h \in V_h. \tag{4.10}$$

We call V_h the *ansatz space* (trial space) and also the *test space* in this standard Galerkin formulation, since the trial and test spaces coincide. The existence and uniqueness of the solution for the discrete problem follows from the fact that $V_h \subset V$ and the inner product and norm are inherited: the bilinear form $a(\cdot, \cdot)$ and the linear functional $F(\cdot) = \langle f, \cdot \rangle$ remain continuous on V_h. Likewise coercivity on V implies coercivity on V_h. Hence, by the Lax–Milgram theorem there exists a unique $u_h \in V_h$ solving the discrete problem.

The Galerkin method is named after the Soviet engineer Boris Galerkin who developed this method in 1915. I. G. Bubnov also developed a similar approach for variational problems, while Galerkin independently proposed the method in the context of structural mechanics.

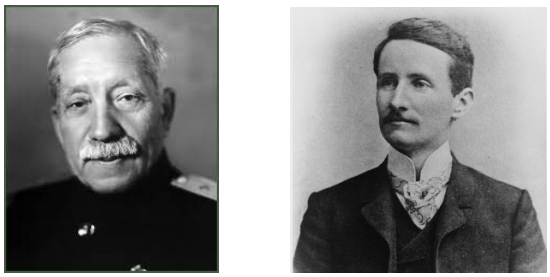


Figure 4.3: Boris Galerkin (4 March 1871 – 12 July 1945, left) and Walther Ritz (22 February 1878 – 7 July 1909, right).

We now show that the discrete variational problem is equivalent to solving a linear algebraic system. As V_h is finite-dimensional, let $\dim(V_h) = N$, i.e., there exist basis functions $\{\varphi_i\}_{i=1}^N$ that span V_h. Choosing the test functions $v_h = \varphi_i$ in the discrete variational equation yields

$$a(u_h, \varphi_i) = \langle f, \varphi_i \rangle \quad \text{for } i = 1, \ldots, N.$$

Since $u_h \in V_h$, it can be written as a linear combination of the basis:

$$u_h(x) = \sum_{j=1}^{N} u_j \varphi_j(x),$$

where the coefficients $\{u_j\}_{j=1}^N$ are unknown. Substituting this expansion and using linearity of $a(\cdot,\cdot)$ gives

$$\sum_{j=1}^{N} a(\varphi_j, \varphi_i)\, u_j = \langle f, \varphi_i \rangle \quad \text{for } i = 1, \dots, N.$$

Setting

$$a_{ij} = a(\varphi_j, \varphi_i) \quad \text{and} \quad f_i = \langle f, \varphi_i \rangle$$

we obtain the linear system

$$\mathbf{A}\mathbf{u} = \mathbf{b},$$

where $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N\times N}$ is the stiffness matrix and $\mathbf{b} = (f_i) \in \mathbb{R}^N$ is the load vector. Once the vector $\mathbf{u} = (u_j)_{j=1}^N$ is computed we recover the discrete solution u_h.

Next we show that the algebraic system is equivalent to the discrete variational statement in the sense that solving one solves the other. Let $v_h \in V_h$ be arbitrary; then

$$v_h(x) = \sum_{i=1}^{N} v_i \varphi_i(x).$$

By linearity of $a(\cdot,\cdot)$ and of the duality pairing,

$$a(u_h, v_h) = a\Big(u_h, \sum_{i=1}^{N} v_i \varphi_i\Big) = \sum_{i=1}^{N} v_i\, a(u_h, \varphi_i) = \sum_{i=1}^{N} v_i \,\langle f, \varphi_i \rangle = \langle f, v_h \rangle,$$

so the two formulations are equivalent.

This approximation method is referred to as the Galerkin method.

Remark 4.13. 1. In the case $a(\cdot,\cdot)$ is symmetric, one may equivalently obtain the discrete solution by solving the finite-dimensional minimization problem: find $u_h \in V_h$ that minimizes the energy functional

$$F(u) = \frac{1}{2} a(u,u) - \langle f, u \rangle.$$

Numerically solving this minimization problem is referred to as the Ritz method (see [6, §2.5.1]). It is also easy to observe that the stiffness matrix $\mathbf{A}$ is symmetric when the bilinear form is symmetric.

2. If $a(\cdot,\cdot)$ is coercive with coercivity constant $m > 0$, then the matrix $\mathbf{A}$ is positive definite. Indeed, for any vector $\xi \in \mathbb{R}^N$ define $v = \sum_{j=1}^N \xi_j \varphi_j \in V_h$. Then

$$\xi^\top \mathbf{A}\, \xi = \sum_{i=1}^{N}\sum_{j=1}^{N} \xi_j a_{ij} \xi_i = a\Big(\sum_{j=1}^{N} \xi_j \varphi_j, \sum_{i=1}^{N} \xi_i \varphi_i\Big) = a(v,v) \geq m \|v\|_V^2.$$

In particular, if $v \neq 0$ then $\xi^\top \mathbf{A}\, \xi > 0$.

Remark 4.14. (**Petrov–Galerkin Method**) In the standard Galerkin method the trial (ansatz) and test spaces coincide. In the Petrov–Galerkin method they may be different: find $u_h \in W_h$ such that

$$a(u_h, w_h) = \langle f, w_h \rangle \quad \forall\ w_h \in V_h,$$

where W_h and V_h are finite-dimensional subspaces of possibly different Hilbert spaces W and V. The conditions on $a(\cdot,\cdot)$ and $F(\cdot)$ must be adapted accordingly (continuity, inf–sup conditions, or appropriate stability assumptions). Petrov–Galerkin methods (introduced by Georgy Petrov) are important in applications such as stabilized methods for advection-dominated problems (e.g. SUPG) [8].

4.3.1 Abstract Error Estimates

Now the next thing we want to do is estimate the error that arises when approximating Eq. 4.9 with Eq. (4.10).

Lemma 4.15. (**Céa's Lemma[BS07]**) *Let V be a Hilbert space. Suppose that the bilinear form $a(\cdot,\cdot)$ is continuous and coercive, and the linear form $F(\cdot)$ is continuous. Then, for the unique solutions u and u_h of Eq.* (4.9) *and Eq.* (4.10)*, respectively, we have*

$$\|u-u_h\|_V \leq \frac{M}{m}\inf_{v_h\in V_h}\|u-v_h\|_V,$$

where M and m are the continuity and coercivity constants, respectively.

Proof. Since $a(u,v)=\langle f,v\rangle$ for all $v\in V$ and $a(u_h,v_h)=\langle f,v_h\rangle$ for all $v_h\in V_h$, and as $V_h\subset V$, we get

$$a(u-u_h,v_h)=0 \quad \forall v_h\in V_h. \tag{4.11}$$

Due to coercivity, we have

$$\begin{aligned} m\|u-u_h\|_V^2 &\leq a(u-u_h,u-u_h)\\ &= a(u-u_h,u-v_h)+a(u-u_h,v_h-u_h) \quad \forall v_h\in V_h. \end{aligned}$$

Since $u_h,v_h\in V_h$, we have $u_h-v_h\in V_h$, and hence the second term vanishes due to Eq. (4.11). Using the continuity of $a(\cdot,\cdot)$, we get

$$\begin{aligned} m\|u-u_h\|_V^2 &\leq a(u-u_h,u-u_h)\\ &\leq M\|u-u_h\|_V\,\|u-v_h\|_V,\\ \Rightarrow \|u-u_h\|_V &\leq \frac{M}{m}\|u-v_h\|_V \quad \forall v_h\in V_h. \end{aligned}$$

Therefore,

$$\|u-u_h\|_V \leq \frac{M}{m}\inf_{v_h\in V_h}\|u-v_h\|_V.$$

□

Céa's lemma was proved by Jean Céa a French mathematician in his PhD thesis.

Remark 4.16. In the case where $a(\cdot,\cdot)$ is a symmetric bilinear form, the norm on V is induced by $\sqrt{a(\cdot,\cdot)}$, i.e.,

$$\|v\|_V^2 = a(v,v),$$

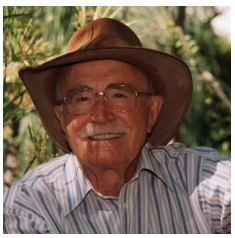


Figure 4.4: Jean Céa: 8 February 1932 - 9 January 2024.

then in the proof of Lemma 4.15 we obtain

$$\|u - u_h\|_V \leq \inf_{v_h \in V_h} \|u - v_h\|_V,$$

since continuity of $a(\cdot, \cdot)$ is replaced by the Cauchy–Schwarz inequality, and coercivity follows from the definition of the norm. Hence, the solution is the best approximation.

For a non-symmetric bilinear form, we get that the error is bounded by the best approximation error, i.e., it is quasi-optimal.

Lemma 4.17. (**Galerkin Orthogonality**) *Let u and u_h be the solutions of Eq.* (4.9) *and Eq.* (4.10)*, respectively. Then*

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h.$$

Proof. As shown in the proof of Lemma 4.15. □

For a symmetric bilinear form, the Galerkin orthogonality states that $u - u_h \perp v_h$ for all $v_h \in V_h$, and hence u_h is the best approximation (see Fig. 4.5).

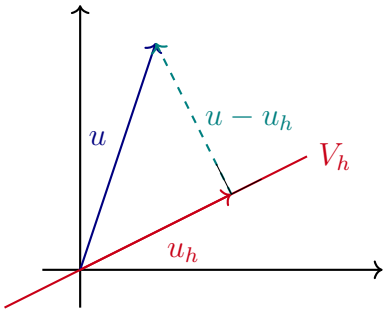


Figure 4.5: Galerkin orthogonality: the error $u - u_h$ is orthogonal to the subspace V_h.

Now, the idea in the development of numerical methods is to find a finite-dimensional subspace V_h such that the matrix $\mathbf{A}$ is easy to solve. What we want is that the entries can be computed efficiently and that the matrix is sparse. This is the origin of the *finite element method*.

Chapter 5

Finite Element Method

In the previous chapter, we introduced the idea of Galerkin methods, where the goal is to find an approximate solution in a finite-dimensional subspace $V_h \subset V$. In this chapter, we explore this concept in detail.

The finite element method (FEM) was originally developed by civil engineers to analyze the stress and strain in structures such as buildings and bridges. However, the lack of mathematical rigor in early implementations sometimes led to catastrophic failures. One well-known example is the Tacoma Narrows Bridge, built in the 1940s, which collapsed just four months after its inauguration. This incident highlights the importance of studying the mathematical foundations of the finite element method.

Figure 5.1: Collapse of the Tacoma Narrows Bridge.

In its simplest form, the finite element method provides a systematic way to construct finite-dimensional subspaces V_h, called *finite element spaces*. The construction of such spaces is characterized by three basic concepts:

FEM 1 Triangulation of $\overline{\Omega}$: The domain $\overline{\Omega}$ is divided into a finite number of subsets K, called *finite elements*, that together cover $\overline{\Omega}$:

$$\overline{\Omega} = \bigcup_K K.$$

FEM 2 Finite Element Space: Once a triangulation is defined, we construct a finite element

space V_h consisting of functions that are either polynomial or "close" to polynomial on each element.

FEM 3 Basis Functions: The third fundamental concept is the existence of at least one *canonical basis* of V_h, whose elements have small (ideally local) supports.

With these foundational ideas in mind, we now examine each in detail, beginning with the notion of triangulation.

5.1 Triangulation

The first step in the finite element method is the decomposition of Ω into geometrical shapes such as triangles or quadrilaterals. This decomposition is called a *triangulation*, denoted by $\mathcal{T}_h$. In general, one can also use other polygonal or polyhedral elements such as pentagons or hexagons; however, the most commonly used shapes are triangles (or tetrahedra in three dimensions) and quadrilaterals (or hexahedra).

Each polyhedron (or cell) in $\mathcal{T}_h$ is called a *mesh cell* and is denoted by K.

Definition 5.1 (**Admissible Triangulation**)**.** A triangulation $\mathcal{T}_h$ is called *admissible* if the following conditions hold:

1. $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$.
2. Each element K is closed, and its interior $\mathring{K}$ is non-empty.
3. For each $K \in \mathcal{T}_h$, the boundary ∂K is Lipschitz-continuous.
4. The intersection of any two mesh cells is either empty or a common m-face, where $m \in \{0, 1, \ldots, d-1\}$.

In some references, the elements K are taken to be open sets. In that case,

$$\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} \overline{K}.$$

Figure 5.2 shows examples of admissible (right) and non-admissible (left) triangulations. The left mesh is non-admissible because $K_1 \cap K_2$ is an *incomplete face*. In some literature, admissible triangulations are also referred to as *conforming triangulations*.

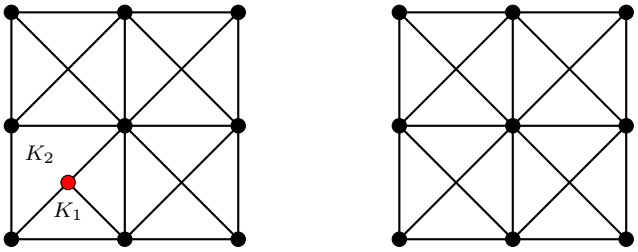


Figure 5.2: Non-admissible triangulation (left) and admissible triangulation (right).

5.1.1 Simplex

The three-dimensional counterpart of a triangle is a tetrahedron. More generally, both triangles and tetrahedra can be viewed as specific cases of a higher-dimensional object called a *simplex*. Before defining a simplex, let us first recall the notion of a convex hull.

Definition 5.2 (**Convex Hull**). The *convex hull* of a set of points is the smallest convex set that contains those points.

Definition 5.3 (**Simplex**). A *d-simplex* in $\mathbb{R}^d$ is the convex hull of $d+1$ points $\{\mathbf{a}_j\}_{j=0}^{d} \subset \mathbb{R}^d$ such that the matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_0 & \mathbf{a}_1 & \cdots & \mathbf{a}_d \\ 1 & 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$$

is non-singular; that is, $\det(\mathbf{A}) \neq 0$.

At first glance, this definition might appear abstract. To gain intuition, consider the case $d = 2$. Then, the simplex is a triangle determined by three non-collinear points, each having two coordinates. In this case, the determinant in the above definition is twice the area of the triangle—equivalently, twice the result obtained using the *shoelace formula* for polygonal areas.

Now, we can write $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_0 & \mathbf{a}_1 - \mathbf{a}_0 & \cdots & \mathbf{a}_d - \mathbf{a}_0 \\ 1 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

Therefore

$$\det(\mathbf{A}) = \det\left[(\mathbf{a}_1 - \mathbf{a}_0)\cdots(\mathbf{a}_d - \mathbf{a}_0)\right],$$

which is the area of the parallelopiped spanned by the vectors $(\mathbf{a}_1 - \mathbf{a}_0)$, $(\mathbf{a}_2 - \mathbf{a}_0), \cdots, (\mathbf{a}_d - \mathbf{a}_0)$ and

$$\det(\mathbf{A}) = d!|K|$$

where $|K|$ is the Lebesgue measure of the d-simplex K is $\mathbb{R}^d$. It's clear to see in 2d that $|K| = 0.5\det(\mathbf{A})$, in 3d as well we get the volume of tetrahedron is $\det(\mathbf{A})/6$. You can decompose a parallelopiped into 6 tetrahedron (see Fig. 5.3).

5.1.2 Barycentric Coordinates

A point in space is usually represented using Cartesian coordinates. However, there exist other coordinate systems that are often more natural in specific contexts. One such system is the *barycentric coordinate system*, introduced by August Möbius in 1827.

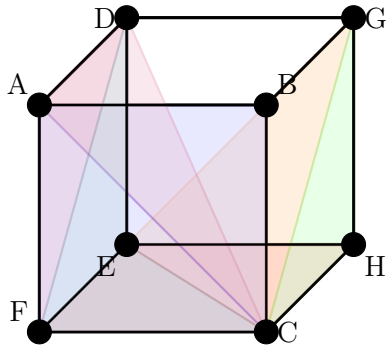


Figure 5.3: Decomposition of parallelopiped into six tetrahedron. The six tetrahedron FBEC, BECH, BEGH, BDGE, DABE, and FBEA.

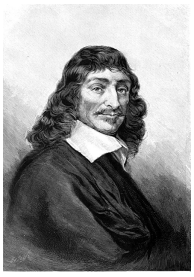
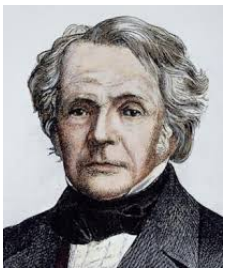



Figure 5.4: René Descartes (31 March 1596 – 11 February 1650, left) and August Möbius (17 November 1790 – 26 September 1868, right).

Definition 5.4 (Barycentric Coordinates). Let $\{\mathbf{a}_j\}_{j=0}^d$ be the $(d+1)$ vertices of a d-simplex in $\mathbb{R}^d$. The *barycentric coordinates* of any point $\mathbf{x} \in \mathbb{R}^d$ with respect to these vertices are the functions $\{\lambda_j(\mathbf{x})\}_{j=0}^d$ satisfying

$$\sum_{j=0}^{d} \lambda_j(\mathbf{x})\, \mathbf{a}_j = \mathbf{x}, \qquad \sum_{j=0}^{d} \lambda_j(\mathbf{x}) = 1.$$

Equivalently, component-wise,

$$\sum_{j=0}^{d} \lambda_j(\mathbf{x})\, a_{ji} = x_i, \qquad \text{for } 1 \le i \le d.$$

In this definition, we have $(d+1)$ unknowns $\{\lambda_j(\mathbf{x})\}_{j=0}^d$ and $(d+1)$ equations: d from the first condition and one from the second.

Hence, the barycentric coordinates are the solution of the linear system

$$\begin{bmatrix} a_{01} & a_{11} & \cdots & a_{d1} \\ a_{02} & a_{12} & \cdots & a_{d2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{0d} & a_{1d} & \cdots & a_{dd} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_d \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \\ 1 \end{bmatrix}.$$

Compactly, we may write

$$\begin{bmatrix} \mathbf{a}_0 & \mathbf{a}_1 & \cdots & \mathbf{a}_d \\ 1 & 1 & \cdots & 1 \end{bmatrix} \lambda = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}.$$

By Cramer's rule,[1] the barycentric coordinates admit a geometric interpretation. Let $\mathbf{x}$ be a point in the d-simplex K, and let $K_i(\mathbf{x})$ denote the simplex formed by replacing vertex $\mathbf{a}_i$ of K by $\mathbf{x}$ (see Fig. 5.5). Then

$$\lambda_i(\mathbf{x}) = \frac{|K_i(\mathbf{x})|}{|K|}.$$

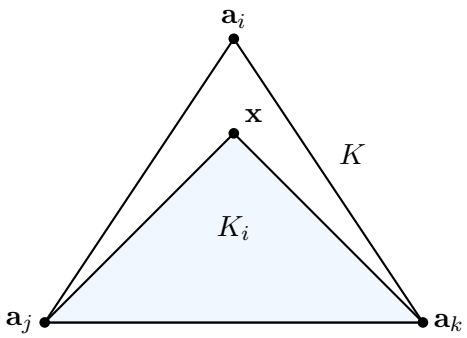


Figure 5.5: Geometric interpretation of the barycentric coordinates.

Hence, $\lambda_i(\mathbf{x})$ is a function of $\mathbf{x}$ that vanishes on the face of the d-simplex K opposite to $\mathbf{a}_i$ and equals 1 at $\mathbf{a}_i$.

Example 5.5. Consider the simplex with vertices $\mathbf{a}_0 = (0,0)$, $\mathbf{a}_1 = (0,1)$, and $\mathbf{a}_2 = (1,0)$. We compute $\{\lambda_i(\mathbf{x})\}_{i=0}^2$. In two dimensions, the general affine function has the form $a + bx + cy$. For $\lambda_0(\mathbf{x})$ corresponding to $\mathbf{a}_0$, we require

$$\lambda_0(0,0) = 1, \qquad \lambda_0(0,1) = 0, \qquad \lambda_0(1,0) = 0.$$

Thus,

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

which gives $a = 1$, $b = -1$, and $c = -1$. Hence using the same ideas we get Barycentric coordinates as,

$$\lambda_0(\mathbf{x}) = 1 - x - y, \qquad \lambda_1(\mathbf{x}) = y, \qquad \lambda_2(\mathbf{x}) = x.$$

[1] For a linear system $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{n\times n}$,

$$x_i = \frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})},$$

where $\mathbf{A}_i$ is obtained from $\mathbf{A}$ by replacing its i^{th} column with $\mathbf{b}$.

5.1.3 Affine Mapping

Once the triangulation is constructed, we have many elements K. Defining basis functions individually on each K would be inefficient. Instead, we define them on a *reference element* and map them to each physical element through an affine map.

Definition 5.6 (Affine Map). Let $U \subset \mathbb{R}^m$ and $V \subset \mathbb{R}^n$. A map $f : U \to V$ is said to be *affine* if, for all $\{\mathbf{x}_i\}_{i=1}^k \subset U$ and coefficients $\{\beta_i\}_{i=1}^k$ satisfying $\sum_{i=1}^k \beta_i = 1$, we have

$$f\left(\sum_{i=1}^{k} \beta_i \mathbf{x}_i\right) = \sum_{i=1}^{k} \beta_i \, f(\mathbf{x}_i).$$

For $k = 2$, affine transformations preserve lines; for $k = 3$, they preserve planes. More generally, affine maps preserve parallelism but not necessarily distances or angles.

Example 5.7. 1. Every linear transformation is affine.
2. Let $U = V = \mathbb{R}^2$. Then

$$f(\mathbf{x}) = (-2x + y + 5,\ 3x + 8y - 2)$$

is affine but not linear.

Reference Element

Let $\hat{K}$ denote the *reference simplex* spanned by the vectors $\{\mathbf{e}_i\}_{i=0}^d \subset \mathbb{R}^d$, where

$$\mathbf{e}_0 = (0, 0, \ldots, 0)^\top, \quad \mathbf{e}_1 = (1, 0, \ldots, 0)^\top, \quad \ldots, \quad \mathbf{e}_d = (0, 0, \ldots, 1)^\top.$$

Let $\mathbf{a}_i = F_K(\mathbf{e}_i)$ for $i = 0, 1, \ldots, d$. The affine map $F_K : \hat{K} \to K$ is defined by (see Fig. 5.6)

$$F_K(\hat{\mathbf{x}}) = \mathbf{B}_K \hat{\mathbf{x}} + b_K, \qquad \mathbf{B}_K = \begin{bmatrix} \mathbf{a}_1 - \mathbf{a}_0 & \cdots & \mathbf{a}_d - \mathbf{a}_0 \end{bmatrix}, \qquad b_K = \mathbf{a}_0.$$

It is straightforward to verify that F_K is bijective.

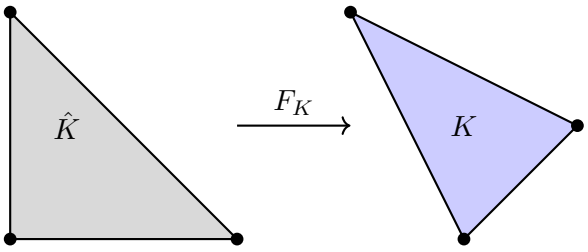


Figure 5.6: Reference map from $\hat{K}$ to K.

Remark 5.8. Whenever the hat notation (e.g., $\hat{K}$) is used, it refers to the reference element and quantities defined on it.

Lemma 5.9. *Let $\{\lambda_i\}_{i=0}^d$ be the barycentric coordinates of the d-simplex $K \subset \mathbb{R}^d$, and let $F_K : \hat{K} \to K$ be the affine mapping from the reference simplex $\hat{K}$ to K. Then the functions*

$$\hat{\lambda}_i = \lambda_i \circ F_K, \qquad i = 0, 1, \ldots, d,$$

are the barycentric coordinates of the reference simplex $\hat{K}$.

Proof. Let $\hat{\mathbf{x}} \in \hat{K}$ and define $\hat{\lambda}_i(\hat{\mathbf{x}}) = \lambda_i(F_K(\hat{\mathbf{x}}))$ for $i = 0, 1, \ldots, d$. We must show that $\{\hat{\lambda}_i(\hat{\mathbf{x}})\}_{i=0}^d$ satisfy

$$\sum_{i=0}^{d} \hat{\lambda}_i(\hat{\mathbf{x}})\mathbf{e}_i = \hat{\mathbf{x}}, \qquad \sum_{i=0}^{d} \hat{\lambda}_i(\hat{\mathbf{x}}) = 1.$$

Let $\mathbf{x} = F_K(\hat{\mathbf{x}}) = \mathbf{B}_K\hat{\mathbf{x}} + b_K$. Since $\{\lambda_i(\mathbf{x})\}_{i=0}^d$ are the barycentric coordinates of K, we have

$$\sum_{i=0}^{d} \lambda_i(\mathbf{x})\, \mathbf{a}_i = \mathbf{x}, \qquad \sum_{i=0}^{d} \lambda_i(\mathbf{x}) = 1.$$

Applying F_K and using $\hat{\lambda}_i = \lambda_i \circ F_K$, we immediately obtain

$$\sum_{i=0}^{d} \hat{\lambda}_i(\hat{\mathbf{x}}) = 1.$$

For the geometric condition, using $\mathbf{a}_i = \mathbf{B}_K\mathbf{e}_i + b_K$, we get

$$\sum_{i=0}^{d} \lambda_i(\mathbf{x})\, \mathbf{a}_i = \sum_{i=0}^{d} \hat{\lambda}_i(\hat{\mathbf{x}})(\mathbf{B}_K\mathbf{e}_i + b_K) = \mathbf{B}_K \left(\sum_{i=0}^{d} \hat{\lambda}_i(\hat{\mathbf{x}})\mathbf{e}_i \right) + b_K.$$

Since the left-hand side equals $\mathbf{x} = \mathbf{B}_K\hat{\mathbf{x}} + b_K$, it follows that

$$\mathbf{B}_K \left(\sum_{i=0}^{d} \hat{\lambda}_i(\hat{\mathbf{x}})\mathbf{e}_i \right) = \mathbf{B}_K\hat{\mathbf{x}}.$$

Because $\mathbf{B}_K$ is invertible, we conclude

$$\sum_{i=0}^{d} \hat{\lambda}_i(\hat{\mathbf{x}})\mathbf{e}_i = \hat{\mathbf{x}}.$$

Hence, the functions $\hat{\lambda}_i$ are indeed the barycentric coordinates of $\hat{K}$. □

This mapping of barycentric coordinates plays a crucial role later, as it allows us to define basis functions on the reference element $\hat{K}$ and map them efficiently to each element K in the triangulation.

5.1.4 Shape Regularity

The last property we mention for a triangulation is its *shape-regularity*.

Definition 5.10 (Shape regular). Let $K \in \mathcal{T}_h$ be a triangle (or, more generally, a d-simplex). Denote by h_K the diameter of K, by ρ_K the diameter of the largest ball inscribed in K (so that $\rho_K = 2r_K$ when r_K is the inradius), and set $h = \max_{K\in\mathcal{T}_h} h_K$. The family $\mathcal{T}_h$ is called *shape-regular* if there exists a constant $C > 0$, independent of h_K and ρ_K, such that for all $K \in \mathcal{T}_h$

$$\frac{h_K}{\rho_K} \leq C.$$

This condition forbids arbitrarily narrow ("skinny") elements in the mesh. In particular, in a shape-regular mesh the ratio h_K/ρ_K is uniformly bounded; see Fig. 5.7 (right) for a non-shape-regular example where $\rho_K \ll h_K$.

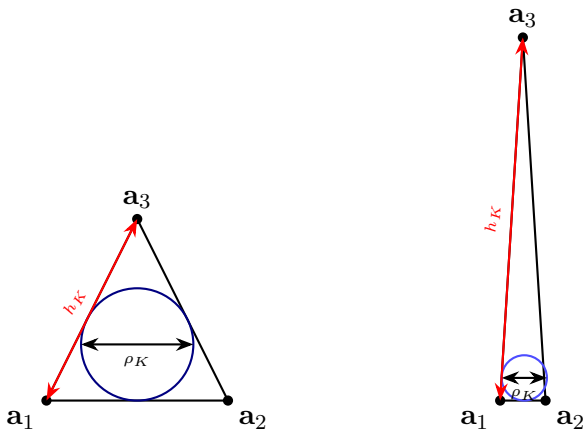


Figure 5.7: Left: a shape-regular triangle with incircle (diameter ρ_K) and diameter h_K. Right: a non-shape-regular (skinny) triangle; its incircle diameter ρ_K is much smaller compared with h_K.

A triangulation is called *quasi-uniform* if it is shape-regular and, in addition, there exists a constant $C > 0$ such that

$$Ch \leq h_K \qquad \forall\, K \in \mathcal{T}_h,$$

i.e., the diameters of all elements are of comparable size.

We next record some useful relations between geometric quantities of K and those of the reference element $\hat{K}$.

Lemma 5.11. *Let $F_K : \hat{K} \to K$ be the affine map*

$$F_K(\hat{\mathbf{x}}) = \mathbf{B}_K \hat{\mathbf{x}} + b_K,$$

with $\mathbf{B}_K \in \mathbb{R}^{d\times d}$ invertible. Then

$$\det(\mathbf{B}_K) = \frac{|K|}{|\hat{K}|}, \qquad \|\mathbf{B}_K\| \le \frac{h_K}{\rho_{\hat{K}}}, \qquad \|\mathbf{B}_K^{-1}\| \le \frac{\hat{h}}{\rho_K},$$

where $\|\cdot\|$ denotes the operator norm induced by the Euclidean vector norm, $\hat{h} = \operatorname{diam}(\hat{K})$, and $\rho_{\hat{K}}$ is the diameter of the largest inscribed ball in $\hat{K}$.

Proof. The Jacobian J_{F_K} is constant and equal to $\mathbf{B}_K$, hence the change of variables gives

$$|K| = \int_K 1\, d\mathbf{x} = \int_{\hat{K}} \left|\det(J_{F_K})\right| d\hat{\mathbf{x}} = |\det(\mathbf{B}_K)|\,|\hat{K}|,$$

which yields the first identity

$$\det(\mathbf{B}_K) = \frac{|K|}{|\hat{K}|}.$$

For the second inequality recall the operator norm definition

$$\|\mathbf{B}_K\| = \sup_{\xi\neq 0} \frac{\|\mathbf{B}_K\xi\|}{\|\xi\|} = \frac{1}{\rho_{\hat{K}}} \sup_{\|\xi\|=\rho_{\hat{K}}} \|\mathbf{B}_K\xi\|.$$

By definition of $\rho_{\hat{K}}$ (the diameter of the largest inscribed ball) every vector $\xi \in \mathbb{R}^d$ with $\|\xi\| = \rho_{\hat{K}}$ can be written as a difference $\xi = \hat{y}_1 - \hat{y}_2$ of two points $\hat{y}_1, \hat{y}_2 \in \hat{K}$ (indeed take the center $\hat{x}_0$ of the inscribed ball and set $\hat{y}_{1,2} = \hat{x}_0 \pm \xi/2$). Therefore

$$\|\mathbf{B}_K\xi\| = \|\mathbf{B}_K\hat{y}_1 - \mathbf{B}_K\hat{y}_2\| = \|F_K(\hat{y}_1) - F_K(\hat{y}_2)\| \le \operatorname{diam}(K) = h_K.$$

Taking the supremum over $\|\xi\| = \rho_{\hat{K}}$ and dividing by $\rho_{\hat{K}}$ gives

$$\|\mathbf{B}_K\| \le \frac{h_K}{\rho_{\hat{K}}}.$$

The third bound follows by exchanging the roles of K and $\hat{K}$. Indeed apply the preceding estimate to the inverse affine map $F_K^{-1}(\mathbf{x}) = \mathbf{B}_K^{-1}\mathbf{x} - \mathbf{B}_K^{-1}b_K$, whose linear part is $\mathbf{B}_K^{-1}$, and note that $\operatorname{diam}(\hat{K}) = \hat{h}$ and the inscribed-ball diameter of K is ρ_K. This yields

$$\|\mathbf{B}_K^{-1}\| \le \frac{\hat{h}}{\rho_K},$$

as claimed. □

5.2 Finite Element Space

5.2.1 Polynomial Spaces

Once the domain is triangulated, the next step is to define a suitable function space on it. Since the method we study is the *finite element method*, we focus on finite-dimensional spaces, denoted by $\mathcal{P}_K$, which are typically composed of polynomials. We denote the dimension of $\mathcal{P}_K$ by N_K.

Our primary focus will be on *triangular elements*. For such elements, we define the space of polynomials of degree less than or equal to k, denoted by $\mathbb{P}_k(K)$:

$$\mathbb{P}_k(K) = \left\{ p : K \to \mathbb{R} \;\middle|\; p(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}| \le k} a_{\boldsymbol{\alpha}} \, \mathbf{x}^{\boldsymbol{\alpha}}, \quad \mathbf{x} \in K \right\},$$

where the coefficients $a_{\boldsymbol{\alpha}} \in \mathbb{R}$ are constants, and $K \in \mathcal{T}_h$.

Example 5.12. If $k = 1$, we obtain the space of linear polynomials:

$$\mathbb{P}_1(K) = \left\{ p : K \to \mathbb{R} \;\middle|\; p(\mathbf{x}) = a_0 + \sum_{i=1}^{d} a_i x_i \right\}.$$

There are $(d+1)$ coefficients, hence $\dim\big(\mathbb{P}_1(K)\big) = d + 1$.
For instance:

- In one dimension ($d = 1$), the basis is $\{1, x\}$.
- In two dimensions ($d = 2$), the basis is $\{1, x, y\}$.

In general, the dimension of $\mathbb{P}_k$ is

$$\dim(\mathbb{P}_k) = \binom{d+k}{k}.$$

For $k = 1$, this gives $\binom{d+1}{1} = d + 1$, as expected.

Although we will not explore rectangular elements in depth, we briefly mention their polynomial spaces for completeness. For a rectangular (or tensor-product) element K, we define

$$\mathbb{Q}_k(K) = \left\{ p : K \to \mathbb{R} \;\middle|\; p(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|_\infty \le k} a_{\boldsymbol{\alpha}} \, \mathbf{x}^{\boldsymbol{\alpha}}, \quad \mathbf{x} \in K \right\},$$

where $|\boldsymbol{\alpha}|_\infty = \max_{i \in \{1,\dots,d\}} \alpha_i$.

Example 5.13. If $d = 2$ and $k = 1$, then

$$\mathbb{Q}_1(K) = \left\{ p : K \to \mathbb{R} \;\middle|\; p(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|_\infty \leq 1} a_{\boldsymbol{\alpha}}\, \mathbf{x}^{\boldsymbol{\alpha}}, \quad \mathbf{x} \in K \right\}.$$

The possible multi-indices $\boldsymbol{\alpha}$ are $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$. Hence, a general polynomial in $\mathbb{Q}_1$ can be written as

$$p(x, y) = c_0 + c_1 x + c_2 y + c_3 xy.$$

Such elements are called *bilinear elements*, as the polynomial is linear in each argument separately.
In general,

$$\dim(\mathbb{Q}_k) = (k+1)^d.$$

For example, for $d = 2$ and $k = 1$, we have $\dim(\mathbb{Q}_1) = 4$, consistent with the basis $\{1, x, y, xy\}$.

5.2.2 Nodal Functionals

After defining the polynomial space $\mathcal{P}_K$, the next step is to describe how these polynomials are evaluated. For instance, in 2D, one can define a linear polynomial by specifying its values at the vertices of a triangle, or alternatively by its values at the midpoints of the edges. Such point evaluations can be viewed as *functionals* acting on $\mathcal{P}_K$.

Hence, we must define a suitable collection of functionals on $\mathcal{P}_K$. In practice, these functionals are defined on a slightly larger space — typically a subspace of $\mathcal{C}^s(K)$ — to ensure sufficient smoothness. Since $\mathcal{P}_K \subset \mathcal{C}^s(K)$ for some $s \in \mathbb{N} \cup \{0\}$, we take our functionals to be mappings

$$\Phi_{K,i} : \mathcal{C}^s(K) \longrightarrow \mathbb{R}.$$

The number of functionals equals the dimension of $\mathcal{P}_K$, as their role is to uniquely determine the coefficients of a polynomial in $\mathcal{P}_K$.

Let

$$\Sigma_K = \{\Phi_{K,i}\}_{i=1}^{N_K}, \qquad \Phi_{K,i} : \mathcal{C}^s(K) \to \mathbb{R},$$

be a set of N_K linearly independent functionals. We choose exactly N_K functionals so that they capture all degrees of freedom associated with $\mathcal{P}_K$.

Typical examples of functionals used in FEM are:

1. **Point evaluation:** $\Phi(v) = v(\mathbf{x})$ for some $\mathbf{x} \in K$.
2. **Derivative evaluation:** For $K \subset \mathbb{R}$, $\Phi(v) = \dfrac{dv}{dx}(x)$ for $x \in K$.
3. **Integral mean value:** For $K \subset \mathbb{R}^d$ with $d \geq 2$,

$$\Phi(v) = \frac{1}{|E|} \int_E v(s)\, ds,$$

where E is an edge (or face) of K.

The smoothness parameter s is chosen so that all $\Phi_{K,i}$ are continuous on $\mathcal{C}^s(K)$. For example, the derivative functional requires $s = 1$, while point and integral evaluations require $s = 0$.

Definition 5.14 (**Unisolvence**)**.** The polynomial space $\mathcal{P}_K$ is said to be *unisolvent* with respect to the set of functionals $\{\Phi_{K,i}\}_{i=1}^{N_K}$ if, for every $\mathbf{a} = (a_1, \ldots, a_{N_K})^\top \in \mathbb{R}^{N_K}$, there exists a unique $p \in \mathcal{P}_K$ such that

$$\Phi_{K,i}(p) = a_i, \qquad 1 \leq i \leq N_K.$$

Intuitively, the functionals $\Phi_{K,i}$ act as measurement operators that extract the coefficients (or degrees of freedom) of a polynomial in $\mathcal{P}_K$. Thus, for every $\mathbf{a} \in \mathbb{R}^{N_K}$, there exists exactly one polynomial $p \in \mathcal{P}_K$ whose functionals evaluate to the prescribed values a_i.

To construct a corresponding *basis* of $\mathcal{P}_K$, let $\{\mathbf{e}_i\}_{i=1}^{N_K}$ denote the standard basis of $\mathbb{R}^{N_K}$, where $\mathbf{e}_i$ has 1 in the i^{th} position and zeros elsewhere. For each $\mathbf{e}_j$, there exists a unique polynomial $\varphi_{K,j} \in \mathcal{P}_K$ such that

$$\Phi_{K,i}(\varphi_{K,j}) = \delta_{ij}, \qquad 1 \leq i, j \leq N_K.$$

The set of polynomials $\{\varphi_{K,j}\}_{j=1}^{N_K}$ forms a *local basis* of $\mathcal{P}_K$ associated with the functionals Σ_K.

Remark 5.15. The unisolvence of $\mathcal{P}_K$ with respect to Σ_K means that $\{\Phi_{K,i}\}_{i=1}^{N_K}$ forms a basis of the dual space $\mathcal{P}_K^*$. Hence, the sets $\{\Phi_{K,i}\}_{i=1}^{N_K}$ and $\{\varphi_{K,i}\}_{i=1}^{N_K}$ can be viewed as *dual bases* in the algebraic sense.

5.2.3 Finite Element

Now, we are in a position to formally define what a finite element is.

Definition 5.16 (**Finite Element**)**.** A triple $(K, \mathcal{P}_K, \Sigma_K)$ is called a *finite element*, where

- K is an open cell (or element) in the triangulation $\mathcal{T}_h$,
- $\mathcal{P}_K$ is a local finite-dimensional function space defined on K, and
- Σ_K is a set of nodal functionals acting on $\mathcal{P}_K$.

The set of functionals Σ_K is assumed to be *unisolvent* for $\mathcal{P}_K$, that is, for every $\mathbf{a} = (a_1, \ldots, a_{N_K})^\top \in \mathbb{R}^{N_K}$ there exists a unique $p \in \mathcal{P}_K$ such that

$$\Phi_{K,i}(p) = a_i, \qquad i = 1, \ldots, N_K.$$

This definition of a finite element first appeared in the classical monograph by Ciarlet and is therefore often referred to as the *finite element in the sense of Ciarlet* [10].

Now, we look at certain finite elements.

Lagrange Finite Elements

The most widely used finite element is the *Lagrange finite element*, which is a generalization of the classical Lagrange interpolant. Let us briefly recall the Lagrange interpolation formula. Given points $\{\mathbf{x}_i\}_{i=0}^m$, the Lagrange interpolating polynomial of order m is defined such that the j^{th} basis function satisfies

$$\mathbb{L}_j^m(\mathbf{x}_i) = \delta_{ij}, \qquad 0 \le i, j \le m,$$

i.e., it takes the value 1 at the node $\mathbf{x}_j$ and 0 at all other nodes. This idea naturally motivates the definition of the Lagrange finite element.

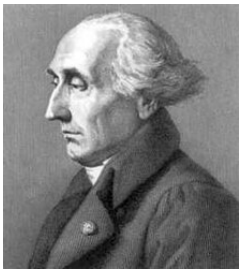


Figure 5.8: Joseph-Louis Lagrange (25 January 1736 – 10 April 1813).

Let $\mathcal{T}_h$ be a triangulation of Ω, and let $\mathcal{P}_K$ be the space of polynomials of degree $\le k$, i.e. $\mathcal{P}_K = \mathbb{P}_k(K)$. The nodal functionals Σ_K are defined as point evaluations on K, that is,

$$\Phi_{K,i}(p) = p(\mathbf{a}_i), \qquad p \in \mathcal{P}_K,\ \mathbf{a}_i \in K.$$

The points $\{\mathbf{a}_i\}$ are referred to as *nodes*. Depending on the polynomial degree k, the number and placement of nodes are chosen accordingly.

Linear Elements ($k = 1$). For $k = 1$, we have $\mathbb{P}_1$ elements, whose dimension is $(d + 1)$. Hence, we select $(d + 1)$ nodes. The simplest choice is to take the vertices of the triangle (or tetrahedron in 3D) as the nodes. Figure 5.9 shows the nodal positions for $\mathbb{P}_1$ elements in two and three dimensions.

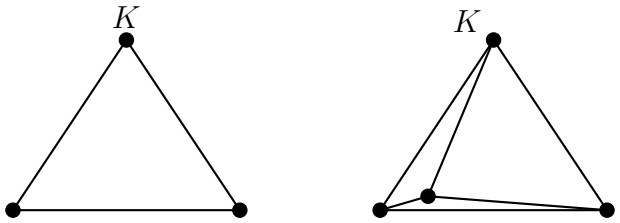


Figure 5.9: Nodes for $\mathbb{P}_1$ linear elements in 2D and 3D.

Once the nodes are identified, the next step is to determine the corresponding local basis functions. For a $\mathbb{P}_1$ element, the local basis function associated with node i must satisfy

$$\Phi_{K,j}(\varphi_i) = \delta_{ij}, \qquad 1 \le i, j \le N_K,$$

which in the case of point evaluations means

$$\varphi_i(\mathbf{a}_j) = \delta_{ij}.$$

We have already encountered such functions — the *barycentric coordinates* $\{\lambda_i\}_{i=1}^{d+1}$. For instance, in 2D, if the vertices of a triangle are $\{\mathbf{a}_i\}_{i=1}^{3}$, then $\varphi_i = \lambda_i$ satisfies the above property. Thus, the barycentric coordinates form the local basis of $\mathbb{P}_1$. See Fig. 5.10 for a schematic representation of the basis functions in 2D.

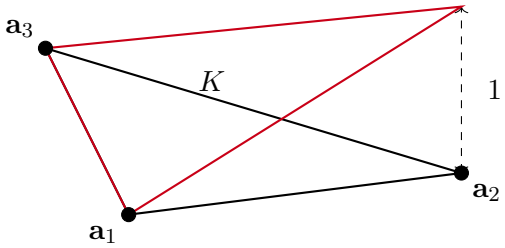


Figure 5.10: Local basis functions for $\mathbb{P}_1$ elements in 2D.

Note that the local basis of $\mathbb{P}_1$ is not the same as the *standard algebraic basis* of $\mathbb{P}_1$. In two dimensions, the local basis is given by $\{1-x-y,\ x,\ y\}$, whereas the standard polynomial basis is $\{1, x, y\}$.

The nodal functionals are *unisolvent* with respect to $\mathbb{P}_1$, since this space possesses a corresponding local basis associated with Σ_K.

Quadratic Elements ($\mathbb{P}_2$). For quadratic elements, we consider the space $\mathbb{P}_2$, whose dimension is

$$\dim(\mathbb{P}_2) = \frac{(d+1)(d+2)}{2}.$$

Compared to $\mathbb{P}_1$ elements, $\mathbb{P}_2$ elements have additional nodes — specifically, there are

$$\frac{d(d+1)}{2}$$

extra nodes corresponding to the midpoints of the edges. Hence, the nodal values are taken both at the vertices and at the midpoints of the edges joining these vertices.

Let $\mathbf{a}_{ij}$ denote the midpoint of the edge connecting the vertices $\mathbf{a}_i$ and $\mathbf{a}_j$. Then, the complete set of nodes for a $\mathbb{P}_2$ element is

$$\{\mathbf{a}_i\}_{i=1}^{d+1} \cup \{\mathbf{a}_{ij}\}_{i<j}.$$

Figure 5.11 shows the nodal configuration for $\mathbb{P}_2$ elements in two and three dimensions.

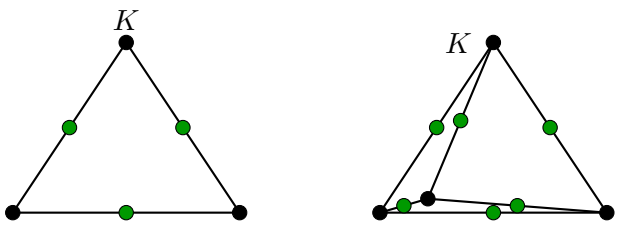


Figure 5.11: Nodes for $\mathbb{P}_2$ quadratic elements in 2D and 3D.

The local basis functions for $\mathbb{P}_2$ elements can be expressed in terms of the barycentric coordinates $\{\lambda_i\}_{i=1}^{d+1}$ as

$$\varphi_i = \lambda_i(2\lambda_i - 1), \qquad i = 1, \ldots, d+1,$$

for vertex-associated basis functions, and

$$\varphi_{ij} = 4\lambda_i\lambda_j, \qquad i, j = 1, \ldots, d+1,\ i < j,$$

for edge-associated basis functions.

Since the barycentric coordinates λ_i satisfy $\lambda_i(\mathbf{a}_j) = \delta_{ij}$, we observe that

$$\varphi_i(\mathbf{a}_j) = \delta_{ij}, \qquad \varphi_{ij}(\mathbf{a}_{kl}) = \begin{cases} 1, & \text{if } \{i,j\} = \{k,l\}, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, since $\lambda_i(\mathbf{a}_{ij}) = \lambda_j(\mathbf{a}_{ij}) = \frac{1}{2}$, the midpoints satisfy

$$\lambda_i(\mathbf{a}_{ij}) = \lambda_j(\mathbf{a}_{ij}) = \tfrac{1}{2}, \qquad \lambda_k(\mathbf{a}_{ij}) = 0 \text{ for } k \neq i, j.$$

These properties show that the basis functions $\{\varphi_i, \varphi_{ij}\}$ are unisolvent with respect to the set of nodal functionals corresponding to the vertices and edge midpoints.

Figure 5.12 illustrates the local basis functions for a $\mathbb{P}_2$ element in 2D. The functions φ_{ij} in two dimensions are often referred to as *bubble functions*, as they attain their maximum value inside the element and vanish at all vertices.

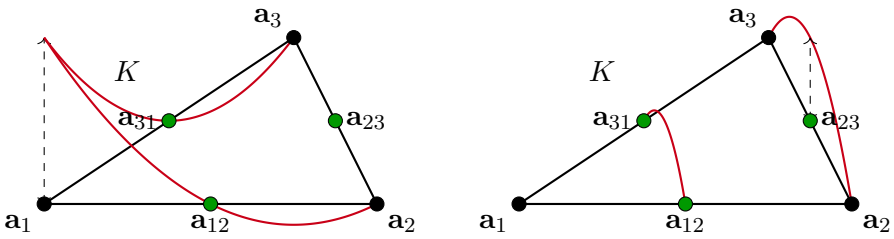


Figure 5.12: Local basis functions for $\mathbb{P}_2$ elements in 2D.

Cubic Elements ($\mathbb{P}_3$). For cubic elements, the local polynomial space is $\mathbb{P}_3$, whose dimension is

$$\dim(\mathbb{P}_3) = \frac{(d+3)(d+2)(d+1)}{6}.$$

The number of nodes increases accordingly. Since this is a cubic polynomial space, we require four points along each edge of the d-simplex, together with additional points inside the higher-dimensional faces.

Let us illustrate this in 2D and 3D.

In 2D: A $\mathbb{P}_3$ element has 10 nodes:

- 3 nodes at the vertices,
- 2 nodes on each of the 3 edges (giving 6 edge nodes),

- 1 node in the interior of the triangle.

The edge nodes are placed at equal distances along each edge, and the interior node is located at the centroid of the triangle. See Fig. 5.13 for the nodal layout.

In 3D: A $\mathbb{P}_3$ tetrahedral element has 20 nodes:

- 4 at the vertices,
- 2 on each of the 6 edges (giving 12 edge nodes),
- 4 on the faces of the tetrahedron (one on each face, typically near the centroid).

The placement follows the same principle as in 2D — uniformly distributed along edges, and symmetrically located within faces.

Figure 5.13: Nodes for $\mathbb{P}_3$ cubic elements in 2D and 3D.

The local basis functions for $\mathbb{P}_3$ elements can be conveniently expressed using barycentric coordinates $\{\lambda_i\}_{i=1}^{d+1}$ as follows:

$$\begin{aligned}
\varphi_i &= \tfrac{1}{2}\lambda_i(3\lambda_i - 1)(3\lambda_i - 2), && i = 1, 2, \ldots, d+1, \\
\varphi_{ij} &= \tfrac{9}{2}\lambda_i\lambda_j(3\lambda_i - 1), && i, j = 1, 2, \ldots, d+1,\ i \neq j, \\
\varphi_{ijk} &= 27\,\lambda_i\lambda_j\lambda_k, && i, j, k = 1, 2, \ldots, d+1,\ i < j < k.
\end{aligned}$$

In 2D, the functions φ_{ijk} are referred to as *bubble functions*, since they attain their maximum value inside the element and vanish at all vertices and edges, resembling a "bubble" on the surface.

Crouzeix Reaviart Finite Element

All the finite elements discussed so far have been of the *Lagrange type*, where the degrees of freedom are defined through point evaluations at the vertices (or additional nodal points) of the elements. However, other types of functionals can be used as well. For instance, instead of using vertex evaluations, we can define the functionals using *midpoint values* or *edge averages*.

Consider $\mathcal{P}_K = \mathbb{P}_1(K)$ and define the following functionals:

$$\begin{aligned}
\Phi_i^{\mathrm{CR}}(v) &= v(\mathbf{a}_{i-1,i+1}), && \text{for } d = 2, \\
\Phi_i^{\mathrm{CR}}(v) &= v(\mathbf{a}_{i-2,i-1,i+1}), && \text{for } d = 3,
\end{aligned}$$

where $\mathbf{a}_{i-1,i+1}$ (or $\mathbf{a}_{i-2,i-1,i+1}$ in 3D) denotes the midpoint of the edge (or face) opposite to vertex i.

It can be verified that, for linear polynomials,

$$\frac{1}{|E|}\int_E v(s)\,ds = v(\mathbf{m}), \tag{5.1}$$

where $\mathbf{m}$ is the midpoint of the edge (or face) E. Hence, the cell average of a linear polynomial over an edge (or face) is equivalent to the point evaluation at its midpoint.

This property is particularly useful in problems where the continuity of the solution is required along edges or faces, rather than at the vertices. The finite element defined by the triplet

$$(K,\, \mathbb{P}_1,\, \Sigma_K^{\mathrm{CR}})$$

is called the *Crouzeix–Raviart (CR) finite element*, named after the French mathematicians Michel Crouzeix and Pierre-Arnaud Raviart. Based on Eq. (5.1), Fig. 5.14 shows the placement of nodes for the CR element in 2D and 3D.

Figure 5.14: Nodes for linear Crouzeix–Raviart elements in 2D and 3D.

The local basis functions for the CR element are given in terms of the barycentric coordinates $\{\lambda_i\}_{i=1}^{d+1}$ as

$$\varphi_i = 1 - d\,\lambda_i, \qquad i = 1, 2, \ldots, d+1.$$

These basis functions satisfy

$$\varphi_i(\mathbf{m}_j) = \delta_{ij},$$

where $\mathbf{m}_j$ are the midpoints of the edges (or faces) (see Fig. 5.15 for basis function in 2D).

The existence of such local basis functions implies that Σ_K^{CR} is unisolvent with respect to $\mathbb{P}_1$.

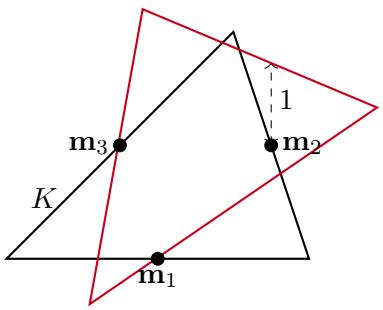


Figure 5.15: Basis for linear Crouzeix–Raviart element in 2D.

Exotic Elements

Besides Lagrange and Crouzeix–Raviart elements, we can construct other finite elements that satisfy specific continuity or derivative conditions of interest. For example, suppose we require not only the function values but also the *directional derivatives* to match at the vertices of the elements. Such elements are known as *Hermite finite elements*.

In one dimension, this idea corresponds to the classical *Hermite interpolant*. Given two points $\{x_i, x_j\}$, and the data values $f(x_i)$, $f'(x_i)$, $f(x_j)$, and $f'(x_j)$, we can construct a unique cubic polynomial $p \in \mathbb{P}_3$ that satisfies

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i), \quad p(x_j) = f(x_j), \quad p'(x_j) = f'(x_j).$$

Thus, the Hermite interpolant in 1D matches both the function and its first derivative at the endpoints.

This idea can be extended to two or three dimensions. In 2D, a cubic Hermite element is based on $\mathbb{P}_3$, whose dimension is 10. To uniquely determine a polynomial in this space, we must specify 10 independent degrees of freedom.

A typical choice is as follows:

- function values at the three vertices,
- the two directional derivatives (along the local coordinate directions) at each vertex, providing $3 \times 3 = 9$ values,
- one additional value at the centroid (or interior point) of the triangle.

These together provide 10 degrees of freedom, sufficient to define the cubic Hermite interpolant in two dimensions.

Figure 5.16 illustrates the placement of nodes and associated derivative degrees of freedom for the 2D Hermite element.

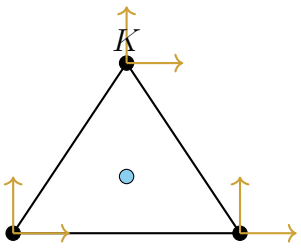


Figure 5.16: Nodes and derivative degrees of freedom for the cubic Hermite element in 2D.

We will not delve into the details of such higher-order or mixed-type elements here. However, for a comprehensive overview of various classical and exotic finite elements, readers are encouraged to consult the *Periodic Table of Finite Elements*5.17 an excellent resource summarizing the properties and relationships among different element families.

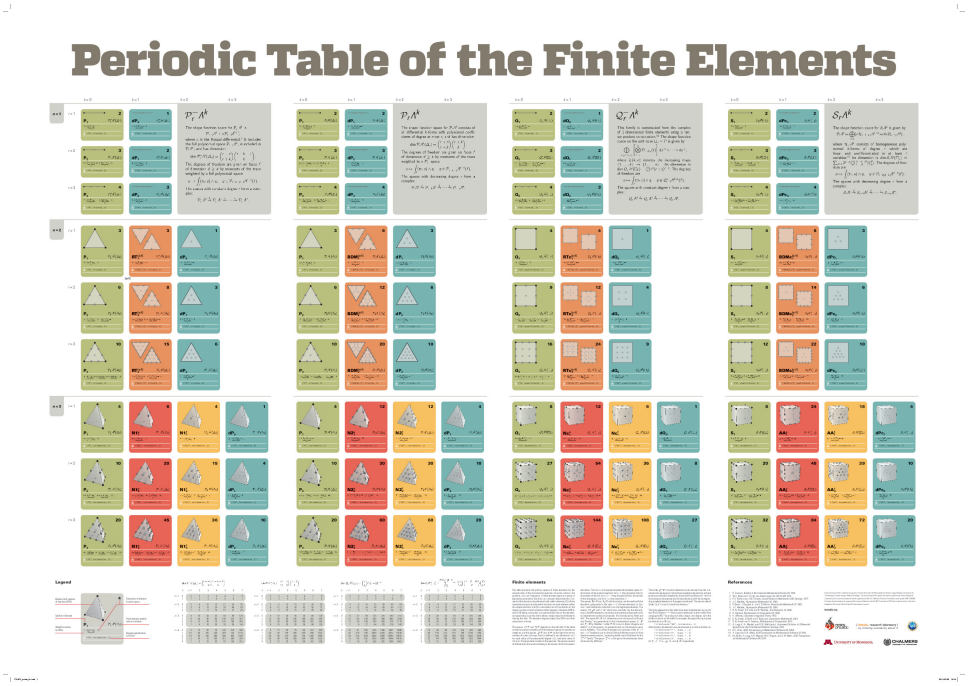


Figure 5.17: Periodic Table of Finite Elements

5.2.4 Finite Element Space

Once the local functionals are defined on each element, they can be combined to form *global functionals* over the entire triangulation.

Global Functional. Let N denote the total number of global nodes of the triangulation $\mathcal{T}_h$. Based on the local functionals $\{\Phi_{K,i}\}_{i=1}^{N_K}$, we define the corresponding global functionals

$$\{\Phi_i\}_{i=1}^{N} : \left\{v \in L^\infty(\Omega) : v|_K \in \mathcal{P}_K \ \forall K \in \mathcal{T}_h\right\} \longrightarrow \mathbb{R}.$$

Here, the restriction $v|_K \in \mathcal{P}_K$ is understood in the sense that the polynomial on K is extended continuously up to the boundary ∂K.

The restriction of a global functional Φ_i to an element K defines a local functional in Σ_K, that is,

$$\Phi_i|_{\mathcal{C}^s(K)} = \Phi_{K,i}, \qquad i = 1, \ldots, N_K,$$

where $\{\Phi_{K,i}\}_{i=1}^{N_K}$ are assumed to be unisolvent on $\mathcal{P}_K$.

For each global functional Φ_i, we define the associated *patch* ω_i as the union of all mesh cells K for which there exists a $p \in \mathcal{P}_K$ satisfying $\Phi_i(p) \neq 0$, i.e.,

$$\omega_i = \bigcup_{\substack{K \in \mathcal{T}_h \\ \Phi_i(p) \neq 0}} K.$$

Example 5.17. Let $d = 2$, and suppose Φ_i is defined as the nodal value of a function v at a point $\mathbf{x} \in K$. Then:

- If $\mathbf{x} \in \mathring{K}$ (the interior of K), then $\omega_i = K$.
- If $\mathbf{x}$ lies on an edge of K (but not at a vertex), then ω_i is the union of all triangles sharing that edge.
- If $\mathbf{x}$ is a vertex of K, then ω_i is the union of all triangles sharing that vertex.

See Fig. 5.18 for a schematic illustration of the different cases.

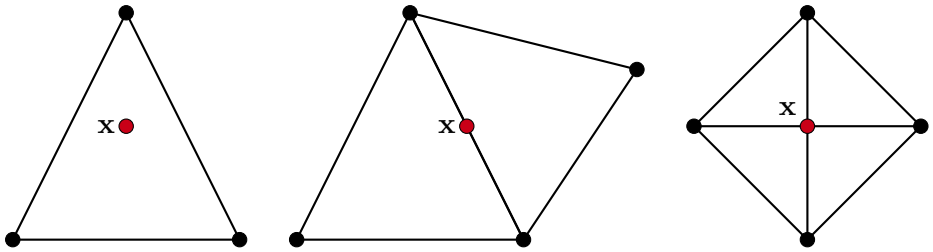


Figure 5.18: Examples of patches ω_i for different node locations in 2D.

Finite Element Space. Now we are in a position to define what a *finite element space* is.

Definition 5.18 (Finite Element Space). A function $v : \Omega \to \mathbb{R}$ with $v|_K \in \mathcal{P}_K$ for all $K \in \mathcal{T}_h$ is said to be *continuous with respect to the global functionals* Φ_i if

$$\Phi_i(v|_{K_1}) = \Phi_i(v|_{K_2}) \qquad \forall\, K_1, K_2 \subset \omega_i.$$

The corresponding finite element space is defined as

$$V_h = \Big\{ v \in L^\infty(\Omega) : \; v|_K \in \mathcal{P}_K \text{ for all } K \in \mathcal{T}_h, \; v \text{ is continuous w.r.t. } \{\Phi_i\}_{i=1}^N \Big\}.$$

The global basis functions $\{\varphi_i\}_{i=1}^N$ of V_h are defined by the duality condition

$$\Phi_i(\varphi_j) = \delta_{ij}, \qquad 1 \le i, j \le N.$$

For Lagrange linear elements, these global basis functions take the familiar "hat" shape, as shown in Fig. 5.19.

After defining the finite element space V_h, we need to ensure that its functions possess the same regularity as the space in which our original PDE was formulated — typically $\mathrm{H}_0^1(\Omega)$ or $\mathrm{H}^1(\Omega)$.

Theorem 5.19. *If, for every $K \in \mathcal{T}_h$, we have $\mathcal{P}_K \subset \mathrm{H}^1(K)$ and $V_h \subset \mathcal{C}(\overline{\Omega})$, then $V_h \subset \mathrm{H}^1(\Omega)$. Moreover, if in addition $v = 0$ on $\partial\Omega$ for all $v \in V_h$, then $V_h \subset \mathrm{H}_0^1(\Omega)$.*

Proof. Let $v \in V_h$. We need to show that $v \in \mathrm{H}^1(\Omega)$, i.e., $D^{\boldsymbol{\alpha}} v \in \mathrm{L}^2(\Omega)$ for all $|\boldsymbol{\alpha}| \le 1$.

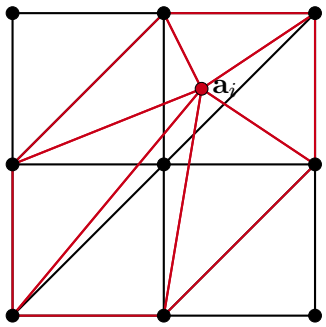


Figure 5.19: Global basis function φ_i for a Lagrange linear finite element in 2D (the "hat function").

Since $V_h \subset \mathcal{C}(\overline{\Omega})$, we immediately have $v \in \mathrm{L}^2(\Omega)$. It remains to show that $D^{\boldsymbol{\alpha}} v \in \mathrm{L}^2(\Omega)$ for all $|\boldsymbol{\alpha}| = 1$, i.e., the weak derivative exists.

For each element $K \in \mathcal{T}_h$, set

$$w := D^{\boldsymbol{\alpha}}(v|_K) \in \mathrm{L}^2(K),$$

since $\mathcal{P}_K \subset \mathrm{H}^1(K)$ implies that derivatives of $v|_K$ are square-integrable.

Let $\varphi \in \mathcal{C}_0^\infty(\Omega)$. Using integration by parts on each K, we have

$$\begin{aligned}\int_\Omega (w\varphi + v D^{\boldsymbol{\alpha}}\varphi)\,dx &= \sum_{K\in\mathcal{T}_h} \int_K (D^{\boldsymbol{\alpha}} v\,\varphi + v D^{\boldsymbol{\alpha}}\varphi)\,dx \\ &= \sum_{K\in\mathcal{T}_h} \int_{\partial K} v|_K\,\varphi\, n_K^{\boldsymbol{\alpha}}\,ds \\ &= \sum_{E\in\mathcal{E}_h} \int_E v|_E\,\varphi\, n_E^{\boldsymbol{\alpha}}\,ds,\end{aligned}$$

where $\mathcal{E}_h$ is the set of all faces (edges in 2D), and $n_E^{\boldsymbol{\alpha}}$ denotes the $\boldsymbol{\alpha}$-component of the unit outer normal on E.

Now, the boundary $\mathcal{E}_h$ can be divided into two parts: faces on $\partial\Omega$ and interior faces. For $E \subset \partial\Omega$, $\varphi = 0$ since $\varphi \in \mathcal{C}_0^\infty(\Omega)$. For an interior face E shared by elements K and K', the continuity $v \in \mathcal{C}(\overline{\Omega})$ ensures

$$v|_{E\subset\partial K} = v|_{E\subset\partial K'} \quad \text{and} \quad n_E^{\boldsymbol{\alpha}}|_K = -\,n_E^{\boldsymbol{\alpha}}|_{K'}.$$

Hence, the contributions from both sides cancel, and the right-hand side vanishes. Therefore, w is the weak derivative of v, proving that $v \in \mathrm{H}^1(\Omega)$. □

According to Theorem 5.19, continuity of v across the faces of neighboring elements is sufficient to guarantee that $V_h \subset \mathrm{H}^1(\Omega)$. To ensure this continuity, it is necessary that the local polynomial spaces have enough degrees of freedom along each face.

For instance, in the case of the Lagrange linear finite element, for any face $E \subset \partial K$ (a $(d-1)$-simplex), d points uniquely determine a function in $\mathcal{P}_1(E)$. In 2D, two points determine

a line. Therefore, the finite element space of piecewise linear functions can be defined as

$$V_h := \{v : \Omega \to \mathbb{R} \ : \ v|_K \in \mathbb{P}_1(K) \ \forall K \in \mathcal{T}_h, \ v \text{ continuous at interior vertices}\}.$$

If we require the homogeneous Dirichlet boundary condition, we define

$$\begin{aligned} V_{h0} \ := \ \Big\{ & v : \Omega \to \mathbb{R} \ : \ v|_K \in \mathbb{P}_1(K) \ \forall K \in \mathcal{T}_h, \\ & v \text{ continuous at interior vertices}, \ v = 0 \text{ on boundary vertices} \Big\}. \end{aligned}$$

Similarly, for higher-order Lagrange elements, $V_h \subset \mathrm{H}^1(\Omega)$ since each face E contains sufficient nodal points to ensure continuity of the polynomial across element boundaries.

Nonconforming Elements. For the Crouzeix–Raviart (CR) element discussed earlier, the continuity condition fails because along an edge or face, only one point determines the linear polynomial, which does not ensure matching traces on both sides. Hence, $V_h \not\subset \mathrm{H}^1(\Omega)$ but rather $V_h \subset V$, where V denotes a larger, less regular space. In this case, the finite element is called a *nonconforming finite element*, and its analysis and implementation differ from those of conforming finite elements.

5.3 Implementation of FEM

One key aspect of the finite element method is its implementation. After introducing the concept of finite elements, let us now look at their application to the Poisson equation in 2D. The ideas can easily be extended to general elliptic operators.

We assume V_h to be the space of linear Lagrange finite elements defined on triangles. Let $\mathcal{T}_h$ be the triangulation of Ω, whose elements are denoted by K. Let $\hat{K}$ be the reference element with vertices $(0,0)$, $(1,0)$, and $(0,1)$. Let F_K denote the affine map from $\hat{K}$ to K, defined by

$$F_K(\hat{\mathbf{x}}) = B_K \hat{\mathbf{x}} + b_K,$$

where $B_K \in \mathbb{R}^{2\times 2}$ and $b_K \in \mathbb{R}^2$. Let F_K^{-1} be the inverse map from K to $\hat{K}$.

Let $(\hat{K}, \mathcal{P}_{\hat{K}}, \Sigma_{\hat{K}})$ be the finite element on $\hat{K}$ and $(K, \mathcal{P}_K, \Sigma_K)$ be the finite element on K. As seen in Section 5.1.2, the barycentric coordinates on K can be expressed using those on $\hat{K}$ as $\lambda = \hat{\lambda} \circ F_K^{-1}$. Similarly, any polynomial (and hence any basis function) defined on K can be written in terms of a polynomial defined on $\hat{K}$. If $\hat{p} \in \mathcal{P}_{\hat{K}}$ is a polynomial on $\hat{K}$, then

$$p = \hat{p} \circ F_K^{-1} \in \mathcal{P}_K.$$

Since we are considering Lagrange finite elements, if $\hat{\mathbf{a}}_i$ is a node on $\hat{K}$, then its image under F_K,

$$\mathbf{a}_i = F_K(\hat{\mathbf{a}}_i),$$

is a node on K.

Gradient Transformation. The bilinear form for the Poisson equation involves gradients. Hence, before we apply the method, let us see how the gradient of a polynomial on $\mathcal{P}_K$ relates to that of $\hat{p} \in \mathcal{P}_{\hat{K}}$. We have

$$\begin{aligned}\nabla p(\mathbf{x}) &= \nabla p(F_K(\hat{\mathbf{x}})) \\ &= \nabla \left(p \circ F_K\right)(\hat{\mathbf{x}}) \\ &= \left(\frac{\partial \hat{p}}{\partial x_1}(\hat{\mathbf{x}}), \frac{\partial \hat{p}}{\partial x_2}(\hat{\mathbf{x}})\right).\end{aligned}$$

Using the chain rule, we get

$$\frac{\partial \hat{p}}{\partial x_1}(\hat{\mathbf{x}}) = \frac{\partial \hat{p}}{\partial \hat{x}_1}\frac{\partial \hat{x}_1}{\partial x_1} + \frac{\partial \hat{p}}{\partial \hat{x}_2}\frac{\partial \hat{x}_2}{\partial x_1},$$

and similarly,

$$\frac{\partial \hat{p}}{\partial x_2}(\hat{\mathbf{x}}) = \frac{\partial \hat{p}}{\partial \hat{x}_1}\frac{\partial \hat{x}_1}{\partial x_2} + \frac{\partial \hat{p}}{\partial \hat{x}_2}\frac{\partial \hat{x}_2}{\partial x_2}.$$

Hence,

$$\nabla \hat{p}(\hat{\mathbf{x}}) = \begin{bmatrix} \frac{\partial \hat{x}_1}{\partial x_1} & \frac{\partial \hat{x}_2}{\partial x_1} \\ \frac{\partial \hat{x}_1}{\partial x_2} & \frac{\partial \hat{x}_2}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial \hat{p}}{\partial \hat{x}_1} \\ \frac{\partial \hat{p}}{\partial \hat{x}_2} \end{bmatrix}.$$

Therefore,

$$\nabla p(\mathbf{x}) = (B_K^{-1})^\top \hat{\nabla} \hat{p}(\hat{\mathbf{x}}). \tag{5.2}$$

5.3.1 Meshing

Solving a PDE using the finite element method is broadly divided into three major steps. The first is the *triangulation* (or meshing) of the domain. There are various ways to triangulate a domain, and in fact, the *topology of triangulation* itself is an active area of research within a branch of mathematics known as *Discrete Geometry*.

The triangulation can be either structured (see Fig. 5.20, left) or unstructured (see Fig. 5.20, right). In either case, the important property is that the triangulation must be *shape-regular*. Certain PDEs also require specific triangulations — for example, Delaunay triangulations. For most practical applications, it is preferable to use external triangulation software, especially for complex geometries. A popular choice is TetGen [24].

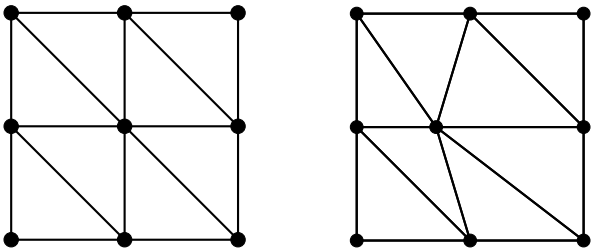


Figure 5.20: Structured grid (left) and unstructured grid (right).

For simple domains, such as the unit square shown in Fig. 5.20 (left), it is easy to generate a triangulation programmatically. For a triangulation, we need to keep track of two quantities:

1. the coordinates of the vertices, and
2. the connectivity (cells or elements), i.e., which vertices form each triangle.

It is convenient to use a pre-existing meshing package in Python; one such lightweight package is `meshzoo`. To install it, use:

```
pip install meshzoo
```

Suppose the grid has $n_x + 1$ points in each direction. Then `meshzoo` returns two arrays:

- `points`: of size $(n_x + 1)^2 \times 2$, containing the coordinates of all vertices,
- `cells`: of size (#triangles)$\times 3$, containing the indices of the vertices forming each triangle.

A simple pseudocode for generating the mesh of a unit square is shown below.

```
def msh_unit_square(nx):
    xs = np.linspace(0, 1, nx + 1)  # Create an array of size nx+1
    ys = np.linspace(0, 1, nx + 1)
    points, cells = meshzoo.rectangle_tri(xs, ys, variant="up")
    return points, cells
```

The keyword `variant` in the above code can take the values `"up"`, `"down"`, or `"zigzag"`. Figure 5.21 shows the resulting grids for each variant.

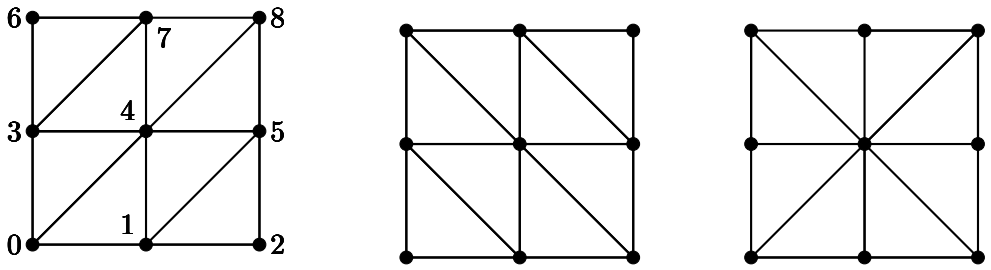


Figure 5.21: Meshes generated by `meshzoo` using `variant="up"` (left), `"down"` (middle), and `"zigzag"` (right).

Let us now look at what this code produces. For the grid in Fig. 5.21 (left), we obtain:

- `points` as an array of size 9×2, and
- `cells` as an array of size 8×3.

We number the grid points from 0 to 8, starting from the bottom left corner and moving from left to right, then bottom to top. Thus, `cells[0]` returns `[0, 1, 4]`, which corresponds to the triangle formed by vertices 0, 1, and 4. If we call `points[cells[0]]`, we obtain the coordinates of these vertices, e.g., `[0, 0]` for vertex 0.

5.3.2 Assembly

Once we have the triangulation, the next step is to assemble the system of equations. Let $V_h \subset \mathrm{H}_0^1(\Omega)$ be spanned by $\{\varphi_i\}_{i=0}^{N_K-1}$, i.e., $\dim(V_h) = N_K$. Using the standard Galerkin finite

element method, we obtain a system of equations of the form

$$\sum_{j=0}^{N_K-1} a_{ij} u_j = b_i, \quad \text{for } i = 0, 1, \ldots, N_K - 1,$$

where $\{u_j\}_{j=0}^{N_K-1}$ are the unknown coefficients. Further, the non-zero entries of the stiffness matrix $\mathbf{A}$ and load vector $\mathbf{b}$ are given by

$$a_{ij} = \sum_{k=0}^{\text{n\_cells}-1} \int_{K_k} \nabla\varphi_j \cdot \nabla\varphi_i \, d\mathbf{x}, \qquad b_i = \sum_{k=0}^{\text{n\_cells}-1} \int_{K_k} f\, \varphi_i \, d\mathbf{x}.$$

Hence, we need to compute the local contributions on each cell and then combine them to obtain the global matrix and vector. Let us consider the above integrals on a single cell K.

Transformation to the Reference Element. We can transform the above integrals to the reference element $\hat{K}$. We write $\varphi_i = \hat{\varphi}_i \circ F_K^{-1}$, and from Eq. (5.2), we have

$$\nabla\varphi_i = (B_K^{-1})^\top \hat{\nabla}\hat{\varphi}_i(\hat{\mathbf{x}}).$$

Hence,

$$\begin{aligned}
\int_K \nabla\varphi_i \cdot \nabla\varphi_j \, d\mathbf{x} &= \int_{\hat{K}} \left((B_K^{-1})^\top \hat{\nabla}\hat{\varphi}_i\right) \cdot \left((B_K^{-1})^\top \hat{\nabla}\hat{\varphi}_j\right) \det(B_K)\, d\hat{\mathbf{x}} \\
&= \int_{\hat{K}} \left(\hat{\nabla}\hat{\varphi}_i\right)^\top B_K^{-1}(B_K^\top)^{-1} \hat{\nabla}\hat{\varphi}_j \, \det(B_K)\, d\hat{\mathbf{x}} \\
&= \int_{\hat{K}} \left(\hat{\nabla}\hat{\varphi}_i\right)^\top \left(B_K^\top B_K\right)^{-1} \hat{\nabla}\hat{\varphi}_j \, \det(B_K)\, d\hat{\mathbf{x}}.
\end{aligned}$$

Similarly, we can transform the right-hand side to

$$\int_K f(\mathbf{x})\varphi_i(\mathbf{x})\, d\mathbf{x} = \int_{\hat{K}} \hat{f}(\hat{\mathbf{x}})\, \hat{\varphi}_i(\hat{\mathbf{x}}) \, \det(B_K)\, d\hat{\mathbf{x}}.$$

Numerical Quadrature. To approximate these integrals over each mesh cell, we use the Gaussian quadrature formula. Since we are dealing with polynomial basis functions, Gaussian quadrature is particularly useful because an n-point Gaussian rule is exact for polynomials of degree up to $2n-1$. The integral over $\hat{K}$ can be approximated as

$$\int_{\hat{K}} f(\hat{\mathbf{x}})\, d\hat{\mathbf{x}} \approx \sum_{q=0}^{N_q-1} w_q\, f(\hat{\mathbf{x}}_q),$$

where N_q is the number of quadrature points, w_q are the quadrature weights, and $\hat{\mathbf{x}}_q$ are the quadrature points. Here again we see the advantage of using a reference element: we only need to compute $\{\hat{\mathbf{x}}_q, w_q\}$ once for $\hat{K}$, and then use them for all mesh cells.

Remark 5.20. We may encounter cases where $f(\cdot)$ is not a polynomial, e.g., $f(x) = \sin(x)$, or when the matrix $\mathbf{A}$ contains non-constant coefficients (e.g., a diffusion coefficient $a(\mathbf{x})$). In such cases, a numerical error arises because Gaussian quadrature is exact only for polynomials. To mitigate this, one can approximate $f(x)$ using the finite element basis functions, i.e.,

$$f(x) \approx \sum_{i=0}^{N_K-1} f_i \, \varphi_i(\mathbf{x}),$$

and then apply a sufficiently high-order quadrature rule. However, numerical errors can still occur if the approximation is not accurate enough. Such errors are sometimes referred to as *variational crimes*.

Once the local matrices are assembled, we combine them to form the global matrix. Each local matrix contributes values at p nodes, where p depends on the polynomial degree of the finite element. For example, for linear elements in 2D, $p = 3$. When we assemble all local matrices, the global matrix becomes larger, and the entries corresponding to shared nodes are summed due to the continuity across elements.

To illustrate the process, consider the grid shown in Fig. 5.22, which has 8 cells and 9 nodes. For simplicity, assume the domain is $[0, 2]^2$. Cell I is formed by the vertices $\{0, 8, 1\}$. The only thing we must ensure is that all triangles have the same orientation, i.e., counterclockwise. We use $\mathbb{P}_1$ Lagrange finite elements for this problem. Since we are using linear elements, the nodes are located at the vertices of each triangle.

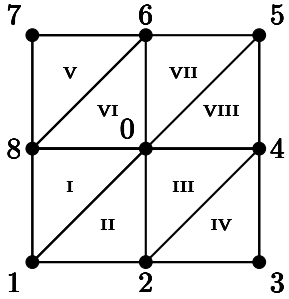


Figure 5.22: Example of a grid used to solve the Poisson equation using FEM.

Now, the triangulation presented in Fig. 5.22 has 9 nodes and hence 9 unknowns in the system of equations. We write our numerical solution

$$u_h(\mathbf{x}) = \sum_{i=0}^{8} u_i \, \Phi_i(\mathbf{x}),$$

where $\{\Phi_i(\mathbf{x})\}_{i=0}^{8}$ are the global basis functions and $\{u_i\}_{i=0}^{8}$ are the nodal values. As seen previously, the assembly of the global matrix follows from that of the local matrices. Hence, we first assemble the matrix locally and then, using continuity along the edges, assemble it globally.

We can notice that all the triangles are of equal area, i.e., $|K| = 0.5$. One important point to consider here is that the orientation of each cell is important; otherwise, the area becomes negative.

Now let us consider Cell I with vertices $\{0, 8, 1\}$. Let F_K be the reference map from $\hat{K}$ to K_{I}, where $\mathbf{e}_0 := (0,0)$ is mapped to $\mathbf{a}_8$, $\mathbf{e}_1 := (1,0)$ is mapped to $\mathbf{a}_1$, and $\mathbf{e}_2 := (0,1)$ is mapped to $\mathbf{a}_0$. Hence, in the affine map $F_{K_{\mathrm{I}}}(\hat{\mathbf{x}}) = B_{K_{\mathrm{I}}}\hat{\mathbf{x}} + \mathbf{b}_{K_{\mathrm{I}}}$ we have

$$\mathbf{b}_{K_{\mathrm{I}}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad B_{K_{\mathrm{I}}} = \begin{bmatrix} \mathbf{a}_1 - \mathbf{a}_8 & \mathbf{a}_0 - \mathbf{a}_8 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

and therefore

$$F_{K_{\mathrm{I}}}(\hat{\mathbf{x}}) = \begin{bmatrix} \hat{x}_2 \\ -\hat{x}_1 + 1 \end{bmatrix}.$$

Also,

$$B_{K_{\mathrm{I}}}^\top B_{K_{\mathrm{I}}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

To compute the stiffness matrix entries, using the change of variables we have

$$\int_{K_{\mathrm{I}}} \nabla\varphi_j \cdot \nabla\varphi_i \, d\mathbf{x} = \int_{\hat{K}} \hat{\nabla}\hat{\varphi}_j^\top \left(B_{K_{\mathrm{I}}}^\top B_{K_{\mathrm{I}}}\right)^{-1} \hat{\nabla}\hat{\varphi}_i \, \det(B_{K_{\mathrm{I}}}) \, d\hat{\mathbf{x}} = \int_{\hat{K}} \hat{\nabla}\hat{\varphi}_j^\top \hat{\nabla}\hat{\varphi}_i \, \det(B_{K_{\mathrm{I}}}) \, d\hat{\mathbf{x}}.$$

On the reference triangle, the three basis functions are

$$\hat{\varphi}_0 = 1 - \hat{x}_1 - \hat{x}_2, \qquad \hat{\varphi}_1 = \hat{x}_1, \qquad \hat{\varphi}_2 = \hat{x}_2.$$

Hence, their gradients are

$$\hat{\nabla}\hat{\varphi}_0 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \hat{\nabla}\hat{\varphi}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \hat{\nabla}\hat{\varphi}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Since the bilinear form is symmetric, we only need to compute the upper-triangular entries of the local matrix, i.e., $a_{00}, a_{11}, a_{88}, a_{01}, a_{08}$, and a_{18}.

Now,

$$\begin{aligned} a_{00} &= \int_{K_{\mathrm{I}}} \nabla\varphi_0 \cdot \nabla\varphi_0 \, d\mathbf{x} = \int_{\hat{K}} \hat{\nabla}\hat{\varphi}_2^\top \left(B_{K_{\mathrm{I}}}^\top B_{K_{\mathrm{I}}}\right)^{-1} \hat{\nabla}\hat{\varphi}_2 \, \det(B_{K_{\mathrm{I}}}) \, d\hat{\mathbf{x}} \\ &= \int_{\hat{K}} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} d\hat{\mathbf{x}} = 0.5. \end{aligned}$$

Similarly,

$$a_{11} = 0.5, \qquad a_{88} = 1.$$

Now, let us move to the off-diagonal elements. Here we notice that

$$\begin{aligned} a_{01} &= \int_{K_{\mathrm{I}}} \nabla\varphi_1 \cdot \nabla\varphi_0 \, d\mathbf{x} = \int_{\hat{K}} \hat{\nabla}\hat{\varphi}_1^\top \left(B_{K_{\mathrm{I}}}^\top B_{K_{\mathrm{I}}}\right)^{-1} \hat{\nabla}\hat{\varphi}_2 \, \det(B_{K_{\mathrm{I}}}) \, d\hat{\mathbf{x}} \\ &= \int_{\hat{K}} \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} d\hat{\mathbf{x}} = 0. \end{aligned}$$

Similarly,

$$a_{08} = -0.5, \qquad a_{18} = -0.5.$$

After the local assembly we obtain the local matrix $\mathbf{A}_{\mathrm{I}}$ as

$$\mathbf{A}_{\mathrm{I}} = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Now, let us compute the right-hand side. For simplicity, we assume f to be a constant c. Then

$$\begin{aligned} b_i^{\mathrm{I}} &= \int_{\hat{K}} c\,\hat{\varphi}_i(\hat{\mathbf{x}})\,\det(B_{K_{\mathrm{I}}})\,d\hat{\mathbf{x}} \\ &= c\int_{\hat{K}} \hat{\varphi}_i(\hat{\mathbf{x}})\,d\hat{\mathbf{x}} \\ &= \frac{c}{6}, \end{aligned}$$

where we have used that $\int_{\hat{K}} \hat{\varphi}_i(\hat{\mathbf{x}})\,d\hat{\mathbf{x}}$ is the same for all i and equals $1/6$. In case f is not constant, we follow the ideas in Remark 5.20. Hence, we get a local system of equations of the form

$$\mathbf{A}_{\mathrm{I}}\mathbf{u}^{\mathrm{I}} = \mathbf{b}^{\mathrm{I}},$$

where $\mathbf{u}^{\mathrm{I}} = \begin{bmatrix} u_0 & u_8 & u_1 \end{bmatrix}^\top$.

Now, one can perform such local assembly over each element and obtain local matrices $\mathbf{A}^k$ and local right-hand sides $\mathbf{b}^k$, for $k = \mathrm{I}, \mathrm{II}, \ldots, \mathrm{VIII}$. Let $\mathbf{A}$ be the 9×9 global stiffness matrix corresponding to the finite element space and $\mathbf{b}$ be the global right-hand side. Then, using the local assemblies, we can compute them.

Let us try to understand this by combining four elements from their local matrices into the global matrix. For this, we consider elements I, II, III, and V. Let us denote their local matrix entries by a_{ij}^k, where $k = \mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{V}$. Then the global matrix has the structure shown below:

$$\begin{bmatrix} a_{00}^{\mathrm{I}} + a_{00}^{\mathrm{II}} + a_{00}^{\mathrm{III}} & a_{01}^{\mathrm{I}} + a_{01}^{\mathrm{II}} & a_{02}^{\mathrm{II}} + a_{02}^{\mathrm{III}} & a_{04}^{\mathrm{III}} & & & a_{08}^{\mathrm{I}} \\ & a_{11}^{\mathrm{I}} + a_{11}^{\mathrm{II}} & a_{12}^{\mathrm{II}} & & & & a_{18}^{\mathrm{I}} \\ & & a_{22}^{\mathrm{II}} + a_{22}^{\mathrm{III}} & a_{24}^{\mathrm{III}} & & & \\ & & & & & & \\ & & & a_{44}^{\mathrm{III}} & & & \\ & & & & & & \\ & & & & a_{66}^{\mathrm{V}} & a_{67}^{\mathrm{V}} & a_{68}^{\mathrm{V}} \\ & & & & & a_{77}^{\mathrm{V}} & a_{78}^{\mathrm{V}} \\ & & & & & & a_{88}^{\mathrm{I}} + a_{88}^{\mathrm{V}} \end{bmatrix}$$

For compactness, we have shown only the upper-triangular part of the matrix.

We can observe that the entry corresponding to a_{00} has three local contributions, as node 0 is shared by three elements. Similarly, other matrix entries accumulate contributions from all elements sharing the corresponding nodes.

In general, when all local matrices are assembled, the global entry a_{00} will have α local contributions, where α is the number of cells containing node 0. In our case, $\alpha = 6$.

It is also important to mention that not all local entries a_{ij}^k are the same. For example, in cell I, $a_{00}^{\mathrm{I}} = 0.5$, whereas in cell III, $a_{00}^{\mathrm{II}} = 1$. The same pattern of assembly applies to the right-hand side as well.

Remark 5.21. While using the affine map from $\hat{K}$ to K, any node of $\hat{K}$ can be mapped to any node of K. That is, if $\{\hat{0}, \hat{1}, \hat{2}\}$ are the nodes of $\hat{K}$ and $\{0, 8, 1\}$ are the nodes of K, then $\hat{0}$ can be mapped to 0, 8, or 1. The only thing to remember is that the new nodes should still preserve the counter-clockwise orientation. For example, $\hat{1}$ should be mapped to 8 and not 1, because then $\{0, 1, 8\}$ would be oriented clockwise.

Algorithm 2 presents the local assembly of the element and Algorithm 3 presents the global assembly.

Algorithm 2 Local Element Matrix Assembly for Linear Triangular Element

Given: Triangle vertex coordinates $X \in \mathbb{R}^{3\times 2}$ with rows $X_0 = (x_0, y_0)$, $X_1 = (x_1, y_1)$, $X_2 = (x_2, y_2)$
Find: Element Area $|K|$, Local stiffness matrix $\mathbf{A}^e \in \mathbb{R}^{3\times 3}$

Step 1: Compute Jacobian and Element Area
Form the Jacobian matrix:

$$\mathbf{B}_K = \begin{bmatrix} x_1 - x_0 & x_2 - x_0 \\ y_1 - y_0 & y_2 - y_0 \end{bmatrix}$$

Compute determinant: $\det(\mathbf{B}_K)$
Compute element area: $|K| = \frac{1}{2}\,|\det(\mathbf{B}_K)|$

Step 2: Gradients on the Reference Element
Define gradients of barycentric basis functions on the reference triangle:

$$\mathbf{G}_{\text{ref}} = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Step 3: Map Gradients to the Physical Element
Compute inverse-transpose of the Jacobian: $\mathbf{J}^{-\top}$
Compute physical gradients:

$$\nabla\boldsymbol{\varphi} = \mathbf{G}_{\text{ref}}\,\mathbf{B}_K^{-\top} \quad \text{(each row is } \nabla\varphi_i\text{)}$$

Step 4: Form Local Stiffness Matrix
Initialize $\mathbf{A}^e = \mathbf{0}_{3\times 3}$
for $i = 0$ **to** 2 **do**
  for $j = 0$ **to** 2 **do**
    $\mathbf{A}^e[i, j] = |K|\,(\nabla\varphi_i \cdot \nabla\varphi_j)$
  end for
end for

---

return $|K|$, $\mathbf{A}^e$

Algorithm 3 Global Assembly of Linear Triangular Finite Element Matrices

Given: Set of node coordinates `points` $= \{(x_i, y_i)\}_{i=1}^{N}$, set of triangular cells `cells` $= \{(i_1, i_2, i_3)\}$, constant right-hand side function $f_c(x, y)$
Find: Global stiffness matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, load vector $\mathbf{b} \in \mathbb{R}^N$

Step 1: Initialization
Let $N =$ number of nodes
Initialize global matrices: $\mathbf{A} = \mathbf{0}_{N \times N}$, $\mathbf{b} = \mathbf{0}_N$

Step 2: Loop over all elements
for each element K in `cells` **do**
Extract vertex coordinates $X =$ `points`$[K] \in \mathbb{R}^{3 \times 2}$
Compute local quantities using Algorithm 2:

$$|K|,\ \mathbf{A}^e\ = \text{LocalElementMatrices}(X)$$

Step 2.1: Assemble local stiffness matrix into global matrix
for $i = 0$ **to** 2 **do**
for $j = 0$ **to** 2 **do**
$\mathbf{A}[K[i], K[j]]$ += $\mathbf{A}^e[i, j]$
end for
end for

Step 2.2: Assemble load vector
for $i = 0$ **to** 2 **do**
$\mathbf{b}[K[i]]$ += $f_c \cdot |K|/3$
end for
end for

---

return **A**, **b**

5.3.3 Boundary Conditions

The last step in the implementation is the treatment of boundary conditions. One can notice that without applying boundary conditions, the global stiffness matrix is non-invertible, as the sum of each row is zero. Hence, we must modify the system. (You may recall a similar issue in the finite difference method for the pure Neumann problem.)

A convenient way to implement the boundary conditions is to partition the global matrix into four blocks:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{\text{in}} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{A}_{\text{BC}} \end{bmatrix},$$

where $\mathbf{A}_{\text{in}}$ corresponds to the part associated with the inner nodes and is of size $M \times M$ (with

M being the number of non-Dirichlet nodes). $\mathbf{A}_{\text{BC}}$ corresponds to the Dirichlet portion of the domain and is of size $(N-M)\times(N-M)$. The matrix $\mathbf{B}$ represents the coupling between inner and boundary nodes. These matrices need not be square and can be rectangular.

Similarly, we partition the right-hand side vector into two parts, $\mathbf{b}_{\text{in}}$ and $\mathbf{b}_{\text{BC}}$. To impose the Dirichlet boundary conditions, we replace $\mathbf{A}_{\text{BC}}$ by the identity matrix of size $(N-M)\times(N-M)$ and set $\mathbf{b}_{\text{BC}}$ to the corresponding Dirichlet values.

We then have two systems of equations:

$$\begin{aligned}\mathbf{A}_{\text{in}}\mathbf{u}_{\text{in}}+\mathbf{B}\mathbf{u}_{\text{BC}} &= \mathbf{b}_{\text{in}},\\ \mathbf{B}^{\top}\mathbf{u}_{\text{in}}+\mathbb{I}\mathbf{u}_{\text{BC}} &= \mathbf{g}.\end{aligned}$$

We do not need to solve the second system explicitly. Hence, the modified system can be written as

$$\begin{bmatrix}\mathbf{A}_{\text{in}} & \mathbf{B}\\ \mathbf{0} & \mathbb{I}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{\text{in}}\\ \mathbf{u}_{\text{BC}}\end{bmatrix}=\begin{bmatrix}\mathbf{b}_{\text{in}}\\ \mathbf{g}\end{bmatrix}\quad\text{or equivalently}\quad\begin{bmatrix}\mathbf{A}_{\text{in}} & \mathbf{0}\\ \mathbf{0} & \mathbb{I}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{\text{in}}\\ \mathbf{u}_{\text{BC}}\end{bmatrix}=\begin{bmatrix}\mathbf{b}_{\text{in}}-\mathbf{B}\mathbf{g}\\ \mathbf{g}\end{bmatrix}.$$

Therefore, we only need to invert the inner matrix $\mathbf{A}_{\text{in}}$, yielding

$$\mathbf{u}_{\text{in}}=\mathbf{A}_{\text{in}}^{-1}\left(\mathbf{b}_{\text{in}}-\mathbf{B}\mathbf{g}\right).$$

Remark 5.22. In the example presented earlier, we had only one inner node, labelled as 0. However, in general (for instance, in `meshzoo`), the nodes are not necessarily ordered with all inner nodes first, followed by boundary nodes. In such cases, we must reorder the rows and columns of the global matrix so that inner nodes appear first, after which the boundary condition modification can be applied.

In most FEM packages, such as FEniCS [23], the numbering convention already labels the inner nodes first and the boundary nodes afterward.

Remark 5.23. In case of higher-order elements, say $\mathbb{P}_2$, the size of both the local and global matrices increases. Here, the derivatives of the local basis functions are not constant, so one needs to use numerical quadrature to evaluate the local integrals accurately.

Remark 5.24. For Neumann boundary conditions, the nodes corresponding to the Neumann boundary are treated as inner nodes and are therefore not replaced with the identity matrix. In general, the node labelling follows the order: inner nodes, Neumann boundary nodes, and then Dirichlet boundary nodes.

Algorithm 4 presents the algorithm for applying the boundary conditions to a homogeneous Poisson problem defined on an unit square.

Algorithm 4 Apply Homogeneous Dirichlet Boundary Conditions and Solve

Given: Global stiffness matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, load vector $\mathbf{b} \in \mathbb{R}^N$, node coordinates $\texttt{points} = \{(x_i, y_i)\}_{i=1}^N$
Find: Solution vector $\mathbf{u} \in \mathbb{R}^N$

Step 1: Identify boundary and interior nodes
Define boundary nodes as those lying on the unit square boundary:

$$\texttt{boundary} = \{\, i \mid x_i = 0 \text{ or } x_i = 1 \text{ or } y_i = 0 \text{ or } y_i = 1 \,\}$$

Define interior nodes:

$$\texttt{interior} = \{\, i \mid i \notin \texttt{boundary} \,\}$$

Step 2: Reduce the system to interior nodes
Extract reduced stiffness matrix and load vector:

$$\mathbf{A}_{\text{in}} = \mathbf{A}(\texttt{interior}, \texttt{interior}), \quad \mathbf{b}_{\text{in}} = \mathbf{b}(\texttt{interior})$$

Step 3: Solve the reduced system
Solve:

$$\mathbf{A}_{\text{in}}\, \mathbf{u}_{\text{in}} = \mathbf{b}_{\text{in}}$$

Step 4: Construct the global solution
Initialize global vector $\mathbf{u} = \mathbf{0}_N$
Assign interior values:

$$\mathbf{u}(\texttt{interior}) = \mathbf{u}_{\text{in}}$$

Boundary nodes remain zero due to homogeneous Dirichlet conditions.

---

return $\mathbf{u}$

5.3.4 Computation of Error

The last thing we want to discuss from an implementation point of view is the computation of error. The same ideas used in the local assembly can be applied here. We will present the computation for the error in the L^2-norm, although these ideas can easily be extended to other norms such as the L^2-norm of the gradient.

Let u be the analytical (exact) solution of the PDE and u_h be the FEM solution. Then

$$\begin{aligned}\|u - u_h\|^2_{\mathrm{L}^2(\Omega)} &= \int_\Omega |u(\mathbf{x}) - u_h(\mathbf{x})|^2 \, d\mathbf{x} \\ &= \sum_{K \in \mathcal{T}_h} \int_K |u(\mathbf{x}) - u_h(\mathbf{x})|^2 \, d\mathbf{x} \\ &= \sum_{K \in \mathcal{T}_h} \|u - u_h\|^2_{\mathrm{L}^2(K)}.\end{aligned}$$

Hence, we need to compute the error on each element K. Using the mapping from the reference element, we obtain

$$\begin{aligned}\|u - u_h\|^2_{\mathrm{L}^2(K)} &= \int_K |u(\mathbf{x}) - u_h(\mathbf{x})|^2 \, d\mathbf{x} \\ &= \int_{\hat{K}} \left| \hat{u}(\hat{\mathbf{x}}) - \sum_{i=1}^{N_K - 1} u_i \, \hat{\varphi}_i(\hat{\mathbf{x}}) \right|^2 \det(B_K) \, d\hat{\mathbf{x}},\end{aligned}$$

where N_K is the number of local nodes, $\hat{u} = u \circ F_K$ is the pullback of u to the reference element, and $\hat{\varphi}_i$ are the local basis functions on $\hat{K}$.

The above integral can be evaluated numerically using a suitable Gaussian quadrature rule on the reference element. This yields an accurate estimate of the L^2 error for the FEM solution.

5.4 Error Estimates

The last aspect we want to discuss about finite elements is their order of error. From Céa's lemma (Lemma 4.15), we already have an estimate of the form

$$\|u - u_h\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V,$$

where C is a constant depending on the coercivity and continuity of the bilinear form $a(\cdot, \cdot)$.

On the discrete space V_h, we can choose $v_h = \Pi_h u$, where $\Pi_h u$ denotes the interpolation (or projection) of u in V_h. Then we can bound

$$\|u - u_h\|_V \leq C \, \|u - \Pi_h u\|_V.$$

Hence, if we are able to bound the interpolation error in terms of the mesh size h, we immediately obtain the desired estimate for the FEM error.

The first step is therefore to construct an appropriate interpolation operator Π_h and then estimate the interpolation error. The general strategy is to decompose the global error over Ω into element-wise contributions on each K, derive the error estimate on the reference element $\hat{K}$, and then transform it to the physical element K.

We already have geometric relationships between norms and semi-norms on K and $\hat{K}$ (see Lemma 5.11). Now, let us define the equivalence of semi-norms between these two elements. Recall that we have a bijective affine map $F_K : \hat{K} \to K$ defined by

$$F_K(\hat{\mathbf{x}}) = B_K\hat{\mathbf{x}} + \mathbf{b}_K.$$

Theorem 5.25 (Equivalence of Semi-Norms, [13, Lemma 1.101])**.** *Let $v \in \mathrm{W}^{m,p}(K)$ for $m \in \mathbb{Z}^+ \cup \{0\}$ and $1 \le p \le \infty$. Then $\hat{v} = v \circ F_K \in \mathrm{W}^{m,p}(\hat{K})$, and it holds that*

$$|\hat{v}|_{m,p,\hat{K}} \le C \, \|B_K\|^m \, |\det(B_K)|^{-1/p} \, |v|_{m,p,K}.$$

Conversely, if $\hat{v} \in \mathrm{W}^{m,p}(\hat{K})$, then $v = \hat{v} \circ F_K^{-1} \in \mathrm{W}^{m,p}(K)$, and

$$|v|_{m,p,K} \le C \, \|B_K^{-1}\|^m \, |\det(B_K)|^{1/p} \, |\hat{v}|_{m,p,\hat{K}}.$$

Proof. The result follows directly from the definition of Sobolev semi-norms, the chain rule for derivatives, and the change of variables formula for integrals. The key observation is that each derivative with respect to x introduces one power of B_K, and hence the factor $\|B_K\|^m$ appears for derivatives of order m. □

5.4.1 Canonical Interpolation

The first thing we want to define is an interpolation operator on the space V_h. For this, we begin by defining what it means for a family of finite elements to be *affine equivalent*.

Definition 5.26 (**Affine Equivalent**)**.** A family of finite elements $(K, \mathcal{P}_K, \Sigma_K)_{K \in \mathcal{T}_h}$ is called *affine equivalent* if there exists a fixed reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ such that for any $K \in \mathcal{T}_h$ there is a bijective affine map

$$F_K : \hat{K} \to K, \qquad F_K(\hat{\mathbf{x}}) = B_K\hat{\mathbf{x}} + \mathbf{b}_K,$$

satisfying

$$\begin{aligned} K &= F_K(\hat{K}), \\ \mathcal{P}_K &= \left\{ p : K \to \mathbb{R} : \quad p = \hat{p} \circ F_K^{-1}, \quad \hat{p} \in \hat{\mathcal{P}} \right\}, \\ \Sigma_K &= \left\{ \Phi(p) : \ \Phi(p) = \Phi(\hat{p} \circ F_K^{-1}) = \hat{\Phi}(\hat{p}), \quad \hat{\Phi} \in \hat{\Sigma} \right\}. \end{aligned}$$

Let $(K, \mathcal{P}_K, \Sigma_K)$ be a finite element with $\dim(\mathcal{P}_K) = N_K$. Since Σ_K is unisolvent with respect to $\mathcal{P}_K$, there exist local basis functions $\{\varphi_j\}_{j=1}^{N_K} \subset \mathcal{P}_K$ satisfying

$$\Phi_i(\varphi_j) = \delta_{ij}, \qquad i, j = 1, 2, \ldots, N_K.$$

Now we define the *canonical interpolation operator* Π_K of $v \in V$ as the unique function $\Pi_K v \in \mathcal{P}_K$ satisfying

$$\Phi_i(\Pi_K v) = \Phi_i(v), \qquad i = 1, 2, \ldots, N_K,$$

i.e., the interpolant shares the same nodal values as v. Then we can express the interpolant as

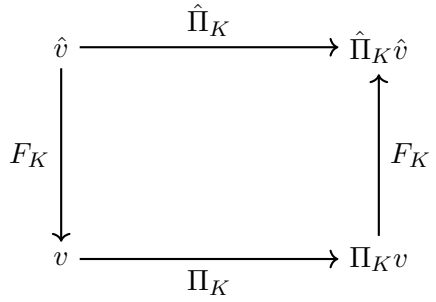
$$\Pi_K v(\mathbf{x}) = \sum_{i=1}^{N_K} \Phi_i(v)\, \varphi_i(\mathbf{x}).$$

The domain of Π_K is the intersection of the domains of all the nodal functionals.

Using the affine map F_K, we define an interpolation operator $\hat{\Pi}_K$ on the reference domain by

$$v \mapsto \hat{\Pi}_K \hat{v} := \left(\Pi_K \left(\hat{v} \circ F_K^{-1}\right)\right) \circ F_K = \widehat{\Pi_K v}, \tag{5.3}$$

which is the inverse pullback of $\Pi_K v$.



We have thus defined the interpolation operator on $\hat{K}$ as the pullback of Π_K. Next, we show that for a family of affine equivalent finite elements, this interpolation operator coincides with the canonical interpolation on $\hat{K}$.

Lemma 5.27. *Let $(K, \mathcal{P}_K, \Sigma_K)_{K \in \mathcal{T}_h}$ be an affine equivalent family of finite elements. Then $\hat{\Pi}_K$ defined by Eq. (5.3) is equal to the canonical interpolation operator $\hat{\Pi} := \hat{\Pi}_K$ on the reference element.*

Proof. For affine equivalent finite elements, it holds that

$$\Phi_i(v) = \hat{\Phi}_i(\hat{v}),$$

and in particular,

$$\Phi_i(\varphi_j) = \hat{\Phi}_i(\hat{\varphi}_j) = \delta_{ij}, \qquad i, j = 1, 2, \ldots, N_K.$$

Now, we show that $\hat{\Pi}_K$ coincides with $\hat{\Pi}$. For $\hat{v} \in \hat{\mathcal{P}}$, we have

$$\begin{aligned} \hat{\Pi}_K \hat{v} &= \widehat{\Pi_K v} \\ &= \sum_{i=1}^{N_K} \Phi_i(v)\, (\varphi_i \circ F_K) \\ &= \sum_{i=1}^{N_K} \hat{\Phi}_i(\hat{v})\, \hat{\varphi}_i. \end{aligned}$$

Hence, the interpolation operator $\hat{\Pi}_K$ satisfies

$$\hat{\Phi}_j\left(\hat{\Pi}_K \hat{v}\right) = \sum_{i=1}^{N_K} \hat{\Phi}_i(\hat{v})\, \hat{\Phi}_j(\hat{\varphi}_i) = \sum_{i=1}^{N_K} \hat{\Phi}_i(\hat{v})\, \delta_{ij} = \hat{\Phi}_j(\hat{v}) \quad \forall j = 1, \ldots, N_K.$$

This is precisely the same defining condition as the canonical interpolation $\hat{\Pi}$ on the reference cell. Since the canonical interpolation is unique,[2] we conclude that $\hat{\Pi}_K = \hat{\Pi}$. □

5.4.2 Local and Global Error Estimates

Now we present some local and global error estimates. Before that we recall an important result named after Bramble and Hilbert.

Theorem 5.28. (**Bramble–Hilbert Lemma**)[13, Lemma B.68] *Let $\Omega \subset \mathbb{R}^d$ be open with Lipschitz boundary. For some $k \in \mathbb{Z}^+ \cup \{0\}$ and $1 \leq p \leq \infty$, let L be a continuous linear operator from $\mathrm{W}^{k+1,p}(\Omega)$ into a normed space Y, with*

$$L(p) = 0 \quad \forall\ p \in \mathbb{P}_k(\Omega).$$

Then there is a constant $C(\Omega)$ such that

$$\|Lv\|_Y \leq C(\Omega)\, \|L\|\, |v|_{k+1,p,\Omega},$$

where

$$\|L\| := \sup_{\substack{v \in \mathrm{W}^{k+1,p}(\Omega) \\ \|v\|_{k+1,p,\Omega}=1}} \|Lv\|_Y.$$

Informally, the Bramble–Hilbert lemma says that if a continuous linear operator is zero on all polynomials of degree $\leq k$, then its value on any v is controlled by the $(k+1)$-seminorm of v.

Remark 5.29. The condition $L(p) = 0$ for $p \in \mathbb{P}_k(\Omega)$ is not only sufficient but also necessary for the inequality, since $|p|_{k+1,p,\Omega} = 0$ for all $p \in \mathbb{P}_k(\Omega)$.

We next state the local interpolation estimate.

[2] The uniqueness follows from the standard argument of assuming two interpolants and using unisolvence to obtain a contradiction.

Theorem 5.30. *Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ be the reference element of an affine-equivalent family of finite elements. Assume for some integers $m, k \in \mathbb{Z}^+ \cup \{0\}$ and $1 \le p, q \le \infty$ that the following hold:*

1. *The nodal functionals $\hat{\Phi}_i : \mathrm{W}^{k+1,p}(\hat{K}) \to \mathbb{R}$, $i = 1, \ldots, N_K$, are linear and continuous.*
2. *The embedding $\mathrm{W}^{k+1,p}(\hat{K}) \hookrightarrow \mathrm{W}^{m,q}(\hat{K})$ is continuous, i.e.*

$$\|v\|_{m,q,\hat{K}} \le C \, \|v\|_{k+1,p,\hat{K}}$$

   for all $v \in \mathrm{W}^{k+1,p}(\hat{K})$.
3. *$\mathbb{P}_k(\hat{K}) \subset \hat{\mathcal{P}} \subset \mathrm{W}^{m,q}(\hat{K})$.*

Then there exists a constant $C = C(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ such that for every affine-equivalent element $(K, \mathcal{P}_K, \Sigma_K)$ and all $v \in \mathrm{W}^{k+1,p}(K)$,

$$|v - \Pi_K v|_{m,q,K} \le C \, |K|^{\frac{1}{q} - \frac{1}{p}} \frac{h_K^{k+1}}{\rho_K^m} \, |v|_{k+1,p,K},$$

where $\Pi_K v$ is the canonical $\mathcal{P}_K$-interpolant of v on K.

Proof. By the embedding hypothesis we have $m \le k+1$ and the nodal functionals and local basis functions make the canonical interpolation

$$\Pi_K v = \sum_{i=1}^{N_K} \Phi_i(v) \, \varphi_i$$

well-defined as a map $\Pi_K : \mathrm{W}^{k+1,p}(K) \to \mathrm{W}^{m,q}(K)$.

Using the inverse affine transformation and Theorem 5.25 we obtain

$$|v - \Pi_K v|_{m,q,K} \le C \, |\det(B_K)|^{1/q} \, \|B_K^{-1}\|^m \, |\hat{v} - \widehat{\Pi_K v}|_{m,q,\hat{K}}.$$

Since $\widehat{\Pi_K v} = \hat{\Pi} \hat{v}$ on the reference cell, write $L(\hat{v}) = \hat{v} - \hat{\Pi}\hat{v}$. The operator $L : \mathrm{W}^{k+1,p}(\hat{K}) \to \mathrm{W}^{m,q}(\hat{K})$ is linear and continuous. We show this by using the embedding and continuity of the nodal functionals and basis norms,

$$\begin{aligned}
\|L\hat{v}\|_{m,q,\hat{K}} &= \|\hat{v} - \hat{\Pi}\hat{v}\|_{m,q,\hat{K}} \\
&= \|\hat{v} - \sum_{i=1}^{N_K} \hat{\Phi}_i(\hat{v}) \hat{\varphi}_i\|_{m,q,\hat{K}} \\
&\le \|\hat{v}\|_{m,q,\hat{K}} + \|\sum_{i=1}^{N_K} \hat{\Phi}_i(\hat{v}) \hat{\varphi}_i\|_{m,q,\hat{K}} \\
&\le C_1 \|\hat{v}\|_{k+1,p,\hat{K}} + \sum_{i=1}^{N_K} |\hat{\Phi}_i(\hat{v})| \|\hat{\varphi}_i\|_{m,q,\hat{K}} \\
&\le C_1 \|\hat{v}\|_{k+1,p,\hat{K}} + \sum_{i=1}^{N_K} C_2 \|\hat{v}\|_{k+1,p,\hat{K}} \|\hat{\varphi}_i\|_{m,q,\hat{K}} \\
&\le C \|\hat{v}\|_{k+1,p,\hat{K}}.
\end{aligned}$$

As $\mathbb{P}_k(\hat{K}) \subset \hat{\mathcal{P}}$ the interpolation preserves polynomials of degree k and hence $L(p) = 0$ for all $p \in \mathbb{P}_k(\hat{K})$. Thus Bramble–Hilbert gives

$$\|L\hat{v}\|_{m,q,\hat{K}} \leq C \, |\, \hat{v} \,|_{k+1,p,\hat{K}}.$$

Combining with the previous inequality yields

$$|v - \Pi_K v|_{m,q,K} \leq C \, |\det(B_K)|^{1/q} \, \|B_K^{-1}\|^m \, |\hat{v}|_{k+1,p,\hat{K}}.$$

Applying the seminorm equivalence (Theorem 5.25) to bound $|\hat{v}|_{k+1,p,\hat{K}}$ in terms of $|v|_{k+1,p,K}$ gives

$$|v - \Pi_K v|_{m,q,K} \leq C \, |\det(B_K)|^{1/q} \, \|B_K^{-1}\|^m \, \|B_K\|^{k+1} \, |\det(B_K)|^{-1/p} \, |v|_{k+1,p,K}.$$

Using Lemma 5.11,

$$|\det(B_K)| = \frac{|K|}{|\hat{K}|}, \qquad \|B_K\| \leq \frac{h_K}{\hat{\rho}}, \qquad \|B_K^{-1}\| \leq \frac{\hat{h}}{\rho_K},$$

and absorbing the fixed reference quantities $(\hat{h}, \hat{\rho}, |\hat{K}|)$ into the constant yields the stated estimate. □

Once the local estimate is available we define the global interpolant.

Definition 5.31. The *global interpolation* $\Pi_h : \mathrm{W}^{k+1,p}(\Omega) \to V_h$ is given elementwise by

$$(\Pi_h v)|_K = \Pi_K v \qquad \text{for each } K \in \mathcal{T}_h.$$

Theorem 5.32. [6, Theorem 4.4.4] *Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ be the reference element of a shape-regular, affine-equivalent family of finite elements with $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} \overline{K}$. Assume that for some integers m and k with $0 \leq m \leq k+1$ the following hypotheses are fulfilled:*

- *The nodal functionals $\hat{\Phi}_i : \mathrm{H}^{k+1}(\hat{K}) \to \mathbb{R}$, for $i = 1, 2, \ldots, N_K$, are linear and continuous.*
- $\mathbb{P}_k(\hat{K}) \subset \hat{\mathcal{P}} \subset \mathrm{H}^m(\hat{K})$.

Then there exists a constant $C > 0$ such that for all $v \in \mathrm{H}^{k+1}(\Omega)$,

$$\begin{aligned} \|v - \Pi_h v\|_{m,\Omega} &\leq C h^{k+1-m} |v|_{k+1,\Omega}, \quad m = 0, 1, \\ \left(\sum_{K \in \mathcal{T}_h} \|v - \Pi_h v\|_{m,K}^2 \right)^{1/2} &\leq C h^{k+1-m} |v|_{k+1,\Omega}, \quad m = 2, 3, \ldots, k+1, \end{aligned}$$

where $\Pi_h v$ is the global canonical interpolation in $V_h \subset \mathcal{C}(\overline{\Omega})$.

Proof. For $p = q = 2$ and $k + 1 \geq m$, the Sobolev embedding $\mathrm{W}^{k+1,2}(\hat{K}) = \mathrm{H}^{k+1}(\hat{K})$ continuously embeds into $\mathrm{W}^{m,2}(\hat{K}) = \mathrm{H}^m(\hat{K})$. Then, by the local interpolation estimate,

$$|v - \Pi_K v|_{m,K} \leq C \, \frac{h_K^{k+1}}{\rho_K^m} \, |v|_{k+1,K}.$$

Since $\mathcal{T}_h$ is shape regular, there exists a constant $C_2 > 0$ such that $h_K \leq C_2 \rho_K$ for all K (see Definition 5.10). As $h_K \leq h$, we have

$$|v - \Pi_K v|_{m,K} \leq C\, h^{k+1-m} |v|_{k+1,K},$$

where C is a generic constant absorbing C_1 and C_2.

Now we compute

$$\begin{aligned}
\|v - \Pi_K v\|_{m,K} &= \left(\sum_{j=0}^{m} |v - \Pi_K v|_{j,K}^2 \right)^{1/2} \\
&\leq C \left(\sum_{j=0}^{m} h^{2(k+1-j)} \right)^{1/2} |v|_{k+1,K} \\
&= C|v|_{k+1,K} \left(h^{2(k+1)} + h^{2k} + \cdots + h^{2(k+1-m)} \right)^{1/2} \\
&\leq C\, h^{k+1-m}\, |v|_{k+1,K},
\end{aligned}$$

where we have used that $\sum_{i=0}^{2m} h^i \leq C$ for sufficiently small h.

Summing over all elements gives

$$\|v - \Pi_h v\|_{m,\Omega} = \left(\sum_{K \in \mathcal{T}_h} |v - \Pi_K v|_{m,K}^2 \right)^{1/2} \leq C h^{k+1-m} |v|_{k+1,\Omega}.$$

The same argument extends to $m \geq 2$, completing the proof. □

The above theorem states that if v has regularity $\mathrm{H}^{k+1}(\Omega)$, then the interpolation error in the H^m-norm decays as $O(h^{k+1-m})$.

Now, returning to our model problem: we seek $u \in \mathrm{H}_0^1(\Omega)$ such that

$$\begin{aligned}
-\Delta u &= f \quad \text{in } \Omega, \\
u &= 0 \quad \text{on } \Gamma,
\end{aligned}$$

for $f \in \mathrm{L}^2(\Omega)$. Let $u_h \in V_h$ be the Galerkin solution using continuous piecewise polynomials of degree $k \geq 1$. If $u \in \mathrm{H}_0^1(\Omega) \cap \mathrm{H}^{k+1}(\Omega)$, then both the continuous and discrete problems are well-posed, and the following error estimate holds:

$$\begin{aligned}
\|u - u_h\|_{1,\Omega} &\leq C \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \\
&\leq C\, \|u - \Pi_h u\|_{1,\Omega} \\
&\leq C\, h^k |u|_{k+1,\Omega}.
\end{aligned}$$

Thus, to achieve order k convergence in the H^1-norm, the solution u must possess H^{k+1} regularity. For instance, if $f \in \mathrm{L}^2(\Omega)$ and Ω is convex, then $u \in \mathrm{H}^2(\Omega)$. However, for non-convex domains such as the L-shaped region, $u \notin \mathrm{H}^2(\Omega)$, and the convergence rate deteriorates (see Fig. 5.23). This loss of regularity can be mitigated by using *adaptive mesh refinement*.

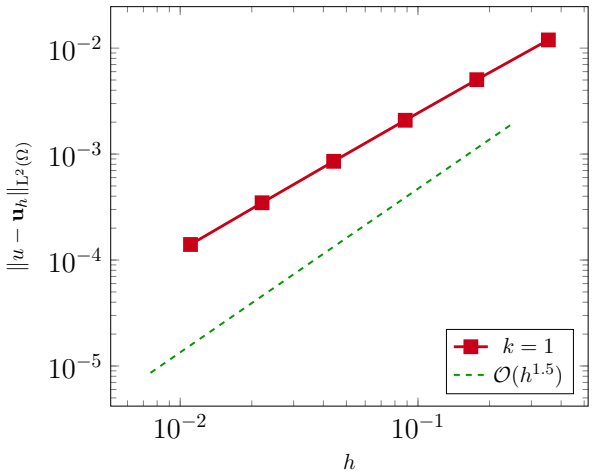


Figure 5.23: Log-log plot of error vs. mesh size h for $\mathbb{P}_1$ element on L-shaped domain.

Finally, note that

$$\|u - u_h\|^2_{\mathrm{L}^2(\Omega)} + \|\nabla(u - u_h)\|^2_{\mathrm{L}^2(\Omega)} = \|u - u_h\|^2_{\mathrm{H}^1(\Omega)},$$

so in particular,

$$\|\nabla(u - u_h)\|_{\mathrm{L}^2(\Omega)} \leq C\, h^k |u|_{k+1,\Omega}.$$

Now, if we look at the interpolation estimate then one sees that for the error in the L^2-norm ($m = 0$) we expect order $k + 1$, while for the H^1-norm we expect order k. Since $\mathrm{H}^1(\Omega) \hookrightarrow \mathrm{L}^2(\Omega)$ continuously, we always have at least order k in the L^2-norm.

Let us look at an example. Take the Poisson equation on the unit square with right-hand side and boundary data chosen so that

$$u(x, y) = \sin(\pi x) \sin(\pi y)$$

is the exact solution. We plot the L^2-error (left) and the H^1-seminorm error (right) in Fig. 5.24. The numerics show decay with orders $k + 1$ and k, respectively, matching the interpolation theory. The sharper L^2-order can be proved analytically using the Aubin–Nitsche lemma.

We now present the continuous and discrete Poisson problems used above.

Find $u \in V = \mathrm{H}^1_0(\Omega)$ such that

$$(\nabla u, \nabla v) = \langle f, v \rangle \quad \forall\ v \in V. \tag{5.4}$$

Find $u_h \in V_h \subset V$ such that

$$(\nabla u_h, \nabla v_h) = \langle f, v_h \rangle \quad \forall\ v_h \in V_h. \tag{5.5}$$

Both problems are well posed by the Lax–Milgram lemma. Subtracting (5.5) from (5.4) and letting $v = v_h = z_h \in V_h$ gives the Galerkin orthogonality:

$$(\nabla(u - u_h), \nabla z_h) = 0 \quad \forall\ z_h \in V_h.$$

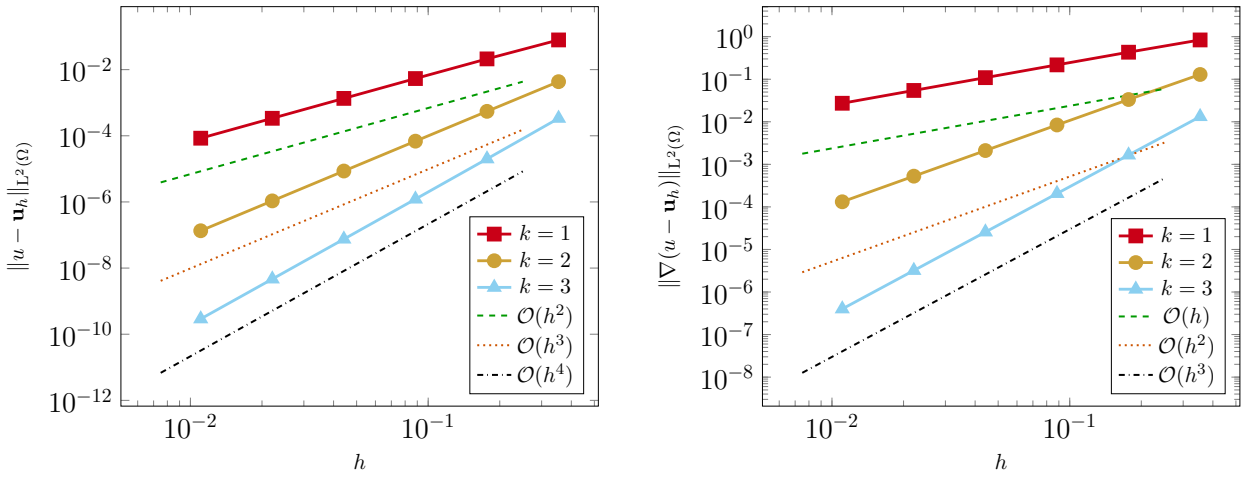


Figure 5.24: Log–log plot of error vs. mesh size h for various $\mathbb{P}_k$ elements. Left: $\|u - u_h\|_{L^2}$; Right: $\|\nabla(u - u_h)\|_{L^2}$.

In functional-analytic terms, for a Hilbert space inner product $(\cdot, \cdot)$ an operator A has an adjoint A^* defined by $(Ax, y) = (x, A^*y)$. For the Poisson operator in variational form, the adjoint problem is the same Poisson problem (see [14, Sec. 6.2.3]. For a given $g \in L^2(\Omega)$ we consider:

Find $z_g \in V$ such that

$$(\nabla v, \nabla z_g) = \langle g, v \rangle \quad \forall\, v \in V. \tag{5.6}$$

Definition 5.33 (H^2**-regularity**)**.** The family of adjoint problems is called H^2-regular if for every $g \in L^2(\Omega)$ the solution z_g satisfies $z_g \in H_0^1(\Omega) \cap H^2(\Omega)$ and

$$|z_g|_{2,\Omega} \leq C \|g\|_{0,\Omega}.$$

Theorem 5.34 (Aubin–Nitsche lemma)**.** *Let the exact solution u of* (5.4) *belong to $H_0^1(\Omega) \cap H^{k+1}(\Omega)$, let $V_h \subset H_0^1(\Omega)$ consist of piecewise polynomials of degree k, and assume the adjoint problems* (5.6) *are H^2-regular. Then*

$$\|u - u_h\|_{0,\Omega} \leq C\, h^{k+1}\, |u|_{k+1,\Omega}.$$

Proof. Take $g := u - u_h \in L^2(\Omega)$ and let $z_g \in V$ solve (5.6) with this right-hand side. Then

$$\|u - u_h\|_{0,\Omega}^2 = (g, u - u_h) = (\nabla(u - u_h), \nabla z_g).$$

Split $z_g = z_g - z_h + z_h$ with arbitrary $z_h \in V_h$. Using Galerkin orthogonality, the z_h-term vanishes:

$$(\nabla(u - u_h), \nabla z_g) = (\nabla(u - u_h), \nabla(z_g - z_h)).$$

Choose $z_h = \Pi_h^1 z_g \in V_h$ as the piecewise linear interpolant (or any suitable interpolant) of

$z_g \in \mathrm{H}^2(\Omega)$. Then, by Cauchy–Schwarz and the interpolation estimate,

$$\begin{aligned}\|u-u_h\|_{0,\Omega}^2 &= (\nabla(u-u_h), \nabla(z_g - \Pi_h^1 z_g)) \\ &\le \|\nabla(u-u_h)\|_{0,\Omega}\, \|\nabla(z_g - \Pi_h^1 z_g)\|_{0,\Omega} \\ &\le \|\nabla(u-u_h)\|_{0,\Omega}\, Ch\, |z_g|_{2,\Omega}.\end{aligned}$$

By the H^2-regularity of the adjoint problem, $|z_g|_{2,\Omega} \le C\|g\|_{0,\Omega} = C\|u-u_h\|_{0,\Omega}$. Therefore

$$\|u-u_h\|_{0,\Omega}^2 \le Ch\, \|\nabla(u-u_h)\|_{0,\Omega}\, \|u-u_h\|_{0,\Omega}.$$

Dividing both sides by $\|u-u_h\|_{0,\Omega}$ (unless the error is zero) gives

$$\|u-u_h\|_{0,\Omega} \le Ch\, \|\nabla(u-u_h)\|_{0,\Omega}.$$

Finally use the H^1-error bound $\|\nabla(u-u_h)\|_{0,\Omega} \le Ch^k|u|_{k+1,\Omega}$ (from Céa + interpolation) to obtain

$$\|u-u_h\|_{0,\Omega} \le Ch^{k+1}|u|_{k+1,\Omega},$$

which is the desired result. □

The Aubin–Nitsche lemma generalizes to abstract Hilbert spaces; see e.g. [13, Lemma 2.3.4].

Remark 5.35. The canonical nodal interpolation requires sufficient smoothness of the underlying function (pointwise values must be defined). By Sobolev embedding, $\mathrm{W}^{m,p}(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$ when $m > d/p$, so nodal interpolation is directly admissible only when this holds. For non-smooth u (or for $d > 1$ with low regularity) one uses other interpolants such as the Clément interpolant [9] or the Scott–Zhang interpolant [22], which are defined for rougher functions and (in the Scott–Zhang case) preserve homogeneous boundary conditions.

5.5 Inverse Estimate

As the name suggests, inverse estimates are inequalities that bound higher-order derivatives in terms of lower-order ones. They are called "inverse" because, unlike approximation estimates (which bound smoother functions by less smooth ones), these go in the opposite direction.

Theorem 5.36 (Inverse Estimates)**.** *Let $\mathcal{P}_K$ be a finite-dimensional space on K consisting of polynomials, let $0 \le k \le l$ be integers, and let $p, q \in [1,\infty]$. Then there exists a constant $C_{\mathrm{inv}} > 0$, depending only on k, l, p, q, such that*

$$|v_h|_{l,q,K} \le C_{\mathrm{inv}}\, h_K^{k-l-d\left(\frac{1}{p}-\frac{1}{q}\right)} |v_h|_{k,p,K} \quad \forall\, v_h \in \mathcal{P}_K.$$

Proof. As in previous results, we first work on the reference element $\hat{K}$ and then map back to K. Let $\hat{v}_h$ be the pullback of v_h to $\hat{K}$. Since $\hat{\mathcal{P}}$ is finite-dimensional, all norms on it are equivalent. In particular, for any $\hat{v}_h \in \hat{\mathcal{P}}$,

$$|\hat{v}_h|_{l,q,\hat{K}} \le \|\hat{v}_h\|_{l,q,\hat{K}} \le C\, \|\hat{v}_h\|_{k,p,\hat{K}}. \tag{5.7}$$

If $k > 0$, define the auxiliary polynomial space

$$\widetilde{\mathcal{P}} = \left\{ \partial^{\boldsymbol{\alpha}} \hat{v}_h : \ \hat{v}_h \in \hat{\mathcal{P}}, \ |\boldsymbol{\alpha}| = k \right\},$$

which is again finite-dimensional. Applying (5.7) to $\widetilde{\mathcal{P}}$ gives

$$\begin{aligned} |\hat{v}_h|_{l,q,\hat{K}} &= \left(\sum_{|\boldsymbol{\alpha}|=k} \|D^{l-k}(\partial^{\boldsymbol{\alpha}} \hat{v}_h)\|_{0,q,\hat{K}}^q \right)^{1/q} \\ &\leq C \left(\sum_{|\boldsymbol{\alpha}|=k} \|\partial^{\boldsymbol{\alpha}} \hat{v}_h\|_{0,p,\hat{K}}^p \right)^{1/p} = C\, |\hat{v}_h|_{k,p,\hat{K}}. \end{aligned}$$

Now we transform this estimate to the physical element K using Theorem 5.25, which relates norms on K and $\hat{K}$:

$$\begin{aligned} |v_h|_{l,q,K} &\leq C\, \|B_K^{-1}\|^l\, |\det(B_K)|^{1/q}\, |\hat{v}_h|_{l,q,\hat{K}} \\ &\leq C\, \|B_K^{-1}\|^l\, |\det(B_K)|^{1/q}\, |\hat{v}_h|_{k,p,\hat{K}} \\ &\leq C\, \|B_K^{-1}\|^l\, |\det(B_K)|^{1/q}\, \|B_K\|^k\, |\det(B_K)|^{-1/p}\, |v_h|_{k,p,K}. \end{aligned}$$

Using the geometric relations from Lemma 5.11,

$$\|B_K^{-1}\| \leq \frac{\hat{h}}{\rho_K}, \qquad \|B_K\| \leq \frac{h_K}{\hat{\rho}}, \qquad |\det(B_K)| = \frac{|K|}{|\hat{K}|} = C\, h_K^d,$$

we obtain

$$\begin{aligned} |v_h|_{l,q,K} &\leq C\, \frac{\hat{h}^l}{\rho_K^l} \frac{h_K^k}{\hat{\rho}^k} \left(\frac{|K|}{|\hat{K}|} \right)^{\frac{1}{q}-\frac{1}{p}} |v_h|_{k,p,K} \\ &\leq C_{\text{inv}}\, h_K^{k-l-d\left(\frac{1}{p}-\frac{1}{q}\right)} |v_h|_{k,p,K}, \end{aligned}$$

where C_{inv} absorbs all reference-element constants. □

Now, we can also obtain a global inverse estimate by combining the results of the local error estimate.

Remark 5.37. If we take $p = q = 2$, $l = 1$, and $k = 0$, the inverse estimate becomes

$$|v_h|_{1,2,K} \leq C_{\text{inv}}\, h_K^{-1}\, |v_h|_{0,2,K},$$

illustrating that higher derivatives are bounded by lower ones with a negative power of h_K.

Remark 5.38. A key point in the proof is the equivalence of norms, which only holds on finite-dimensional spaces. Therefore, inverse estimates do not extend to infinite-dimensional function spaces.

5.6 Summary

In this chapter, we discussed the finite element method and examined its key aspects. As in Chapter 2, the concepts introduced here are easily extendable to general elliptic operators.

As with any introductory course, several advanced topics naturally follow from the material presented. A few of these are listed below, along with references for further study.

1. **Finite Elements for Parabolic Problems.** Consider the heat equation: find $u \in V$ such that

$$\begin{aligned} u_t - \Delta u &= f \quad \text{in } \Omega \times (0,T], \\ u &= g \quad \text{on } \Gamma \times (0,T], \\ u(0,\cdot) &= u_0, \end{aligned}$$

   where $\Omega \subset \mathbb{R}^d$. In this case, along with spatial discretisation, we also need to discretise in time. This leads to the notion of *stability*, analogous to the concept of stability in the numerical treatment of initial value problems. For a detailed treatment, see the classic monograph by Thomée [28].

2. **Mixed Finite Elements.** The Navier–Stokes equations are fundamental in computational fluid dynamics (CFD). We seek $(\mathbf{u}, p) \in V \times Q$ satisfying

$$\begin{aligned} -\nu\Delta\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + \nabla p &= f \quad \text{in } \Omega, \\ \nabla\cdot\mathbf{u} &= 0 \quad \text{in } \Omega, \end{aligned}$$

   with appropriate boundary conditions. Here $\mathbf{u} = (u_1, u_2, u_3)^\top$ is a vector-valued velocity field and p is the pressure. This system naturally leads to the study of *mixed finite element methods*, where both $\mathbf{u}_h$ and p_h are approximated simultaneously. For an in-depth treatment, see the book by Boffi,Brezzi, and Fortin [4].

3. **Adaptive Finite Elements.** In Fig. 5.23, we observed a loss of optimal convergence order for L-shaped domains due to the lack of full regularity ($u \notin \mathrm{H}^{k+1}(\Omega)$). Instead of using a uniform mesh, we can employ an *adaptive mesh refinement* strategy, where finer elements are placed near singularities (see Fig. 5.25).

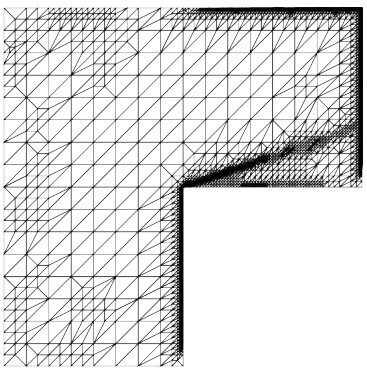


Figure 5.25: Adaptive grid on an L-shaped domain for the convection–diffusion equation.

   Such adaptivity is guided by *a posteriori* error estimators that provide local measures of the error on each cell. For further details, refer to the book by Ainsworth and Oden [2].

Finite elements also form the foundation for several modern high-order discretisation methods, such as the *hybrid high-order* (HHO) method [12] and the *virtual element method* (VEM) [3]. A rigorous mathematical treatment of finite element theory can be found in the standard references [6, 13, 10, 5].

List of Algorithms

Index

Bibliography

[1] R. A. Adams. *Sobolev spaces*. Pure and Applied Mathematics, Vol. 65. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975, pp. xviii+268.

[2] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000, pp. xx+240. ISBN: 0-471-29411-X. DOI: 10.1002/9781118032824.

[3] L. Beirão da Veiga et al. "Basic principles of virtual element methods". In: *Math. Models Methods Appl. Sci.* 23.1 (2013), pp. 199–214. ISSN: 0218-2025,1793-6314. DOI: 10.1142/S0218202512500492.

[4] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*. Vol. 44. Springer Series in Computational Mathematics. Springer, Heidelberg, 2013, pp. xiv+685. ISBN: 978-3-642-36518-8; 978-3-642-36519-5. DOI: 10.1007/978-3-642-36519-5.

[5] D. Braess. *Finite elements*. Third. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker. Cambridge University Press, Cambridge, 2007, pp. xviii+365. ISBN: 978-0-521-70518-9; 0-521-70518-5. DOI: 10.1017/CBO9780511618635.

[6] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*. Third. Vol. 15. Texts in Applied Mathematics. Springer, New York, 2008, pp. xviii+397. ISBN: 978-0-387-75933-3. DOI: 10.1007/978-0-387-75934-0.

[7] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011, pp. xiv+599. ISBN: 978-0-387-70913-0.

[8] A. N. Brooks and T. J. R. Hughes. "Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations". In: *Comput. Methods Appl. Mech. Engrg.* 32.1-3 (1982). FENOMECH '81, Part I (Stuttgart, 1981), pp. 199–259. ISSN: 0045-7825. DOI: 10.1016/0045-7825(82)90071-8.

[9] C. Carstensen. "Clément Interpolation and Its Role in Adaptive Finite Element Error Control". In: *Partial Differential Equations and Functional Analysis*. Birkhäuser Basel, 2006, pp. 27–43. ISBN: 9783764376017. DOI: 10.1007/3-7643-7601-5_2.

[10] P. G. Ciarlet. *The finite element method for elliptic problems*. Studies in Mathematics and its Applications, Vol. 4. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978, pp. xix+530. ISBN: 0-444-85028-7.

[11] E. Di Nezza, G. Palatucci, and E. Valdinoci. "Hitchhikers guide to the fractional Sobolev spaces". In: *Bulletin des Sciences Mathematiques* 136.5 (July 2012), pp. 521–573. ISSN: 0007-4497. DOI: 10.1016/j.bulsci.2011.12.004.

[12] D. A. Di Pietro and J. Droniou. *The hybrid high-order method for polytopal meshes*. Vol. 19. MS&A. Modeling, Simulation and Applications. Design, analysis, and applica-

tions. Springer, Cham, [2020] ©2020, pp. xxxi+525. ISBN: 978-3-030-37202-6; 978-3-030-37203-3. DOI: 10.1007/978-3-030-37203-3.

[13] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*. Vol. 159. Applied Mathematical Sciences. Springer-Verlag, New York, 2004, pp. xiv+524. ISBN: 0-387-20574-8. DOI: 10.1007/978-1-4757-4355-5.

[14] L. C. Evans. *Partial differential equations*. Second. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2010, pp. xxii+749. ISBN: 978-0-8218-4974-3. DOI: 10.1090/gsm/019.

[15] J. D. Faires and R. Burden. *Numerical methods*. Second. With 1 IBM-PC floppy disk (3.5 inch; HD). Brooks/Cole Publishing Co., Pacific Grove, CA, 1998, pp. xii+594. ISBN: 0-534-35187-5.

[16] S. Ganesan and L. Tobiska. *Finite Elements: Theory and Algorithms*. 1st. New York, NY, USA: Cambridge University Press, 2017. ISBN: 1108415709, 9781108415705.

[17] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Second. Vol. 224. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1983, pp. xiii+513. ISBN: 3-540-13025-X. DOI: 10.1007/978-3-642-61798-0.

[18] D. F. Griffiths, J. W. Dold, and D. J. Silvester. *Essential Partial Differential Equations: Analytical and Computational Aspects*. Springer International Publishing, 2015. ISBN: 9783319225692. DOI: 10.1007/978-3-319-22569-2.

[19] S. Kesavan. *Functional analysis*. Second. Vol. 52. Texts and Readings in Mathematics. Springer, Singapore; Hindustan Book Agency, New Delhi, [2023] ©2023, pp. xv+268. ISBN: 978-981-19-7633-9; 978-819-51-9613-5. DOI: 10.1007/978-981-19-7633-9.

[20] E. Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, New York-London-Sydney, 1978, pp. xiv+688. ISBN: 0-471-50731-8.

[21] W. Rudin. *Principles of mathematical analysis*. Third. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, 1976, pp. x+342.

[22] L. R. Scott and S. Zhang. "Finite element interpolation of nonsmooth functions satisfying boundary conditions". In: *Mathematics of Computation* 54.190 (1990), pp. 483–493. ISSN: 1088-6842. DOI: 10.1090/s0025-5718-1990-1011446-7.

[23] M. W. Scroggs et al. "Construction of arbitrary order finite element degree-of-freedom maps on polygonal and polyhedral cell meshes". In: *ACM Trans. Math. Software* 48.2 (2022), Art. 18, 23. ISSN: 0098-3500,1557-7295. DOI: 10.1145/3524456.

[24] H. Si. "TetGen, a Delaunay-based quality tetrahedral mesh generator". In: *ACM Trans. Math. Software* 41.2 (2015), Art. 11, 36. ISSN: 0098-3500,1557-7295. DOI: 10.1145/2629697.

[25] G. Strang. *Linear algebra and its applications*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1976, pp. xi+374.

[26] M. Stynes. "Steady-state convection-diffusion problems". In: *Acta Numer.* 14 (2005), pp. 445–508. ISSN: 0962-4929. DOI: 10.1017/S0962492904000261.

[27] T. Tao. *An Introduction to Measure Theory*. Graduate Studies in Mathematics. American Mathematical Society, 2021. ISBN: 9781470466404. URL: https://books.google.co.in/books?id=k0lDEAAAQBAJ.

[28] V Thomée. *Galerkin finite element methods for parabolic problems*. Second. Vol. 25. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2006, pp. xii+370. ISBN: 978-3-540-33121-6; 3-540-33121-2.