Numerical Methods for Partial Differential Equations

Steffen Börm

Compiled August 19, 2020, 16:33

Contents

1. Introduction

Differential equations have been established as one of the most important representations of the laws governing natural phenomena, e.g., the movement of bodies in a gravitational field or the growth of populations.

If all functions appearing in the equation depend only on one variable, we speak of an *ordinary differential equation*. Ordinary differential equations frequently describe the behaviour of a system over time, e.g., the movement of an object depends on its velocity, and the velocity depends on the acceleration.

Ordinary differential equations can be treated by a variety of numerical methods, most prominently by time-stepping schemes that evaluate the derivatives in suitably chosen points to approximate the solution.

If the functions in the equation depend on more than one variable and if the equation therefore depends on partial derivatives, we speak of a *partial differential equation*. Partial differential equations can be significantly more challenging than ordinary differential equations, since we may not be able to split the computation into discrete (time-)steps and have to approximate the entire solution at once.

A typical example is the *potential equation* of electrostatics. Given a domain $\Omega \subseteq \mathbb{R}^3$, we consider

$$\frac{\partial^2 u}{\partial x_1^2}(x) + \frac{\partial^2 u}{\partial x_2^2}(x) + \frac{\partial^2 u}{\partial x_3^2}(x) = f(x) \qquad \text{for all } x \in \Omega,$$

where $\frac{\partial^\nu u}{\partial x_i^\nu}$ denotes the ν-th partial derivative with respect to the i-th variable.

Explicit solutions for this equation are only known in special situations, e.g., if $\Omega = \mathbb{R}^3$ or $\Omega = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$, while the general case usually has to be handled by numerical methods.

Since computers have only a finite amount of storage at their disposal, they are generally unable to represent the solution u as an element of the infinite-dimensional space $C^2(\Omega)$ exactly. Instead we look for an *approximation* of the solution in a finite-dimensional space that can be represented by a computer. Since the approximation is usually constructed by replacing the domain Ω by a grid of discrete points, the approximation of the solution is called a *discretization*.

A fairly simple discretization technique is the method of *finite differences*: we replace the derivatives by difference quotients and replace Ω by a grid Ω_h such that the difference quotients in the grid points can be evaluated using only values in grid points. In the case of the potential equation, this leads to a system of linear equations that can be solved in order to obtain an approximation u_h of u.

We have to investigate the *discretization error*, i.e., the difference between u_h and u in the grid points. This task can be solved rather elegantly by establishing the *consistency* and the *stability* of the discretization scheme: consistency means that applying the approximated derivatives to the real solution u yields an error that can be controlled, and stability means that small perturbations of the forcing term f lead only to small perturbations of the solution u_h. Once both properties have been established, we find that the discretization scheme is *convergent*, i.e., that we can reach any given accuracy as long as we use a sufficiently fine grid.

For time-dependent problems like the heat equation and the wave equations, it is a good idea to treat the time variable separately. An attractive approach is the *method of lines* that uses a discretization in space to obtain a system of ordinary differential equations that can be treated by standard time-stepping algorithms.

Since the Lipschitz constant arising in this context is quite large, it is a good idea to consider implicit time-stepping schemes that provide better stability and do not require us to use very small time steps in order to avoid oscillations.

The wave equation conserves the total energy of the system, and we would like to have a numerical scheme that shares this property. If we replace the total energy by a suitable discretized counterpart, we find that the *Crank-Nicolson method* guarantees that the discretized total energy indeed remains constant.

In order to prove consistency of finite difference methods, we frequently have to assume that the solution u is quite smooth, e.g., a standard approach for the potential equation requires u to be four times continuously differentiable. This is an assumption that is only rarely satisfied in practice, so we have to consider alternative discretization schemes.

Variational methods are particularly attractive, since they are based on an elegant reformulation of the partial differential equation in terms of Hilbert spaces. We can prove that the variational equation has a unique generalized solution in a *Sobolev space*, and that this generalized solution coincides with the classical solution if the latter exists. Variational formulations immediately give rise to the *Galerkin discretization* scheme that leads to a system of equations we can solve to obtain an approximation of the solution.

If we use a *finite element method*, this system has a number of desirable properties, most importantly it is *sparse*, i.e., each row of the corresponding matrix contains only a small number of non-zero entries. This allows us to apply particularly efficient solvers to obtain the approximate solution.

In order to be able to approximate the solution even with fairly weak regularity assumptions, we investigate the approximation properties of averaged Taylor polynomials and obtain the *Bramble-Hilbert lemma*, a generalized error estimate for these polynomials, and the *Sobolev lemma*, an embedding result for Sobolev spaces that allows us to use standard interpolation operators to construct the finite element approximation.

Acknowledgements

I would like to thank Jens Liebenau, Alexander Dobrick, Mario Santer, Nils Krütgen, and Jonas Lorenzen for corrections and suggestions for improvements of these lecture notes.

2. Finite difference methods

This chapter provides an introduction to a first simple discretization technique for elliptic partial differential equations: the finite difference approach replaces the domain by a grid consisting of discrete points and the derivatives in the grid points by difference quotients using only adjacent grid points. The resulting system of linear equations can be solved in order to obtain approximations of the solution in the grid points.

2.1. Potential equation

A typical example for an *elliptic* partial differential equation is the *potential equation*, also known as *Poisson's equation*. As its name suggests, the potential equation can be used to find potential functions of vector fields, e.g., the electrostatic potential corresponding to a distribution of electrical charges.

In the unit square $\Omega := (0,1) \times (0,1)$ the equation takes the form

$$-\frac{\partial^2 u}{\partial x_1^2}(x) - \frac{\partial^2 u}{\partial x_2^2}(x) = f(x) \qquad \text{for all } x = (x_1, x_2) \in \Omega.$$

In order to obtain a unique solution, we have to prescribe suitable conditions on the boundary

$$\partial\Omega := \overline{\Omega} \cap \overline{\mathbb{R}^2 \setminus \Omega} = \{0,1\} \times [0,1] \cup [0,1] \times \{0,1\}$$

of the domain. Particularly convenient for our purposes are *Dirichlet boundary conditions* given by

$$u(x) = 0 \qquad \text{for all } x = (x_1, x_2) \in \partial\Omega.$$

In the context of electrostatic fields, these conditions correspond to a superconducting boundary: if charges can move freely along the boundary, no potential differences can occur.

In order to shorten the notation, we introduce the *Laplace operator*

$$\Delta u(x) = \frac{\partial^2 u}{\partial x_1^2}(x) + \frac{\partial^2 u}{\partial x_2^2}(x) \qquad \text{for all } x = (x_1, x_2) \in \Omega,$$

and summarize our task as follows:

Find $u \in C(\bar{\Omega})$ with $u|_\Omega \in C^2(\Omega)$, and

$$-\Delta u(x) = f(x) \qquad \text{for all } x \in \Omega, \tag{2.1a}$$

$$u(x) = 0 \qquad \text{for all } x \in \partial\Omega. \tag{2.1b}$$

Solving this equation "by hand" is only possible in special cases, the general case is typically handled by numerical methods.

The solution u is an element of an infinite-dimensional space of functions on the domain Ω, and we can certainly not expect a computer with only a finite amount of storage to represent it accurately. That is why we employ a *discretization*, in this case of the domain Ω: we replace it by a finite number of discrete points and focus on approximating the solution only in these points.

Using only discrete points means that we have to replace the partial derivatives in the equation by approximations that require only the values of the function in these points.

Lemma 2.1 (Central difference quotient) *Let $h \in \mathbb{R}_{>0}$ and $g \in C^4[-h,h]$. We can find $\eta \in (-h,h)$ with*

$$\frac{g(h) - 2g(0) + g(-h)}{h^2} = g''(0) + \frac{h^2}{12} g^{(4)}(\eta).$$

Proof. Using Taylor's theorem, we find $\eta_+ \in (0,h)$ and $\eta_- \in (-h,0)$ with

$$\begin{aligned} g(h) &= g(0) + hg'(0) + \frac{h^2}{2}g''(0) + \frac{h^3}{6}g'''(0) + \frac{h^4}{24}g^{(4)}(\eta_+), \\ g(-h) &= g(0) - hg'(0) + \frac{h^2}{2}g''(0) - \frac{h^3}{6}g'''(0) + \frac{h^4}{24}g^{(4)}(\eta_-). \end{aligned}$$

Adding both equations yields

$$g(h) + g(-h) = 2g(0) + h^2 g''(0) + \frac{h^4}{12}\frac{g^{(4)}(\eta_+) + g^{(4)}(\eta_-)}{2}.$$

Since the fourth derivative $g^{(4)}$ is continuous, we can apply the intermediate value theorem to find $\eta \in [\eta_-, \eta_+]$ with

$$g^{(4)}(\eta) = \frac{g^{(4)}(\eta_+) + g^{(4)}(\eta_-)}{2}$$

and obtain

$$g(h) - 2g(0) + g(-h) = h^2 g''(0) + \frac{h^4}{12} g^{(4)}(\eta).$$

Dividing by h^2 gives us the required equation. ■

Exercise 2.2 (First derivative) *Let $h \in \mathbb{R}_{>0}$ and $g \in C^2[0,h]$. Prove that there is an $\eta \in (0,h)$ such that*

$$\frac{g(h) - g(0)}{h} = g'(0) + \frac{h}{2} g''(\eta).$$

Let now $g \in C^3[-h,h]$. Prove that there is an $\eta \in (-h,h)$ such that

$$\frac{g(h) - g(-h)}{2h} = g'(0) + \frac{h^2}{6} g'''(\eta).$$

Applying Lemma 2.1 to the partial derivatives with respect to x_1 and x_2, we obtain the approximations

$$\frac{2u(x_1,x_2) - u(x_1+h,x_2) - u(x_1-h,x_2)}{h^2} = \frac{\partial^2 u}{\partial x_1^2}(x) + \frac{h^2}{12}\frac{\partial^4 u}{\partial x_1^4}(\eta_1,x_2), \tag{2.2a}$$

$$\frac{2u(x_1,x_2) - u(x_1,x_2+h) - u(x_1,x_2-h)}{h^2} = \frac{\partial^2 u}{\partial x_2^2}(x) + \frac{h^2}{12}\frac{\partial^4 u}{\partial x_2^4}(x_1,\eta_2), \tag{2.2b}$$

with suitable intermediate points $\eta_1 \in [x_1 - h, x_2 + h]$ und $\eta_2 \in [x_2 - h, x_2 + h]$. Adding both equations and dropping the h^2 terms leads to the approximation

$$\Delta_h u(x) = \frac{u(x_1+h,x_2) + u(x_1-h,x_2) + u(x_1,x_2+h) + u(x_1,x_2-h) - 4u(x_1,x_2)}{h^2} \tag{2.3}$$

$$\text{für alle } x \in \Omega,\ h \in H_x$$

of the Laplace operator, where the set

$$H_x := \{h \in \mathbb{R}_{>0} \ :\ x_1 + h \in [0,1],\ x_1 - h \in [0,1],\ x_2 + h \in [0,1],\ x_2 - h \in [0,1]\}$$

describes those step sizes for which the difference quotient can be evaluated without leaving the domain Ω. The approximation (2.3) is frequently called a *five point star*, since the values of u are required in five points in a star-shaped pattern centered at x.

In order to quantify the approximation error, we introduce suitable norms on function spaces.

Reminder 2.3 (Maximum norm) *For real-valued continuous functions on a compact set K, we define the* maximum norm *by*

$$\|u\|_{\infty,K} := \max\{|u(x)| \ :\ x \in K\} \qquad \textit{for all } u \in C(K).$$

For vectors with a general finite index set $\mathcal{I}$, we let

$$\|u\|_\infty := \max\{|u_i| \ :\ i \in \mathcal{I}\} \qquad \textit{for all } u \in \mathbb{R}^{\mathcal{I}}.$$

Lemma 2.4 (Consistency) *If $u \in C^4(\bar\Omega)$ holds, we have*

$$|\Delta_h u(x) - \Delta u(x)| \leq \frac{h^2}{6}|u|_{4,\Omega} \qquad \textit{for all } x \in \Omega,\ h \in H_x, \tag{2.4}$$

where we use the semi-norm

$$|u|_{4,\Omega} := \max\left\{ \left\| \frac{\partial^{\nu+\mu} u}{\partial x_1^\nu \partial x_2^\mu} \right\|_{\infty,\bar\Omega} \ :\ \nu,\mu \in \mathbb{N}_0,\ \nu+\mu = 4 \right\}$$

on the right-hand side that is defined by the maximum norm of the fourth derivatives.

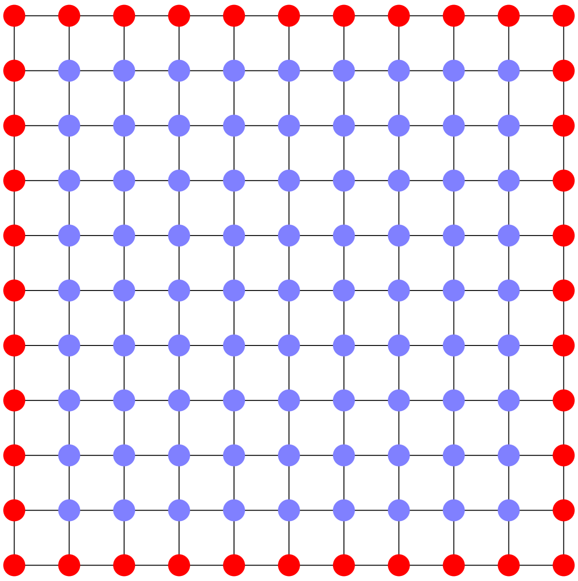

Figure 2.1.: Grid for $N = 9$

Proof. We add the equations (2.2) and bound the fourth derivatives by $|u|_{4,\Omega}$. ∎

Compared to the differential operator Δ, the *difference operator* Δ_h offers the advantage that only values of the function in a small number of discrete points are required. We can use this property to replace the domain Ω by a finite set of points that is far better suited for computers.

Definition 2.5 (Grid) *Let $N \in \mathbb{N}$, and let*

$$\begin{aligned} h &:= \frac{1}{N+1}, \\ \Omega_h &:= \{(ih, jh) \ : \ i, j \in \{1, \ldots, N\}\} \subseteq \Omega, \\ \partial\Omega_h &:= \{(ih, 0), (ih, 1), (0, jh), (1, jh) \ : \ i, j \in \{0, \ldots, N+1\}\} \subseteq \partial\Omega, \\ \bar{\Omega}_h &:= \Omega_h \cup \partial\Omega_h. \end{aligned}$$

We call Ω_h, $\partial\Omega_h$ and $\bar{\Omega}_h$ grids *for the sets Ω, $\partial\Omega$ and $\bar{\Omega}$.*

Restricting the estimate (2.4) to the grid Ω_h yields

$$|-\Delta_h u(x) - f(x)| = |-\Delta_h u(x) + \Delta u(x)| \leq \frac{h^2}{6}\|u\|_{4,\bar{\Omega}} \qquad \text{for all } x \in \Omega,$$

and this property suggests that we look for a solution of the equation $-\Delta_h u = f$, since we may hope that it will approximate the "real" solution u. Since the evaluation of $\Delta_h u$ in $x \in \Omega_h$ requires only values in points of $\bar{\Omega}_h$, we introduce functions that are only defined in these points.

Definition 2.6 (Grid function) *Let Ω_h and $\bar{\Omega}_h$ grids for Ω and $\bar{\Omega}$. The spaces*

$$G(\Omega_h) := \{u_h \ : \ u_h \text{ maps } \Omega_h \text{ to } \mathbb{R}\},$$
$$G(\bar{\Omega}_h) := \{u_h \ : \ u_h \text{ maps } \bar{\Omega}_h \text{ to } \mathbb{R}\}$$

are called spaces of grid functions *from Ω_h and $\bar{\Omega}_h$, respectively, to $\mathbb{R}$. The space*

$$G_0(\bar{\Omega}_h) := \{u_h \in G(\bar{\Omega}_h) \ : \ u_h(x) = 0 \text{ for all } x \in \partial\Omega_h\}$$

is called the space of grid functions with homogeneous Dirichlet boundary conditions*.*

The difference operator Δ_h is obviously a linear mapping from $G(\bar{\Omega}_h)$ to $G(\Omega_h)$, and we can approximate the differential equation (2.1) by the following system of linear equations:

Find a grid function $u_h \in G_0(\bar{\Omega}_h)$ such that

$$-\Delta_h u_h(x) = f(x) \qquad\qquad \text{for all } x \in \Omega_h. \tag{2.5}$$

Since this system of linear equations (each point $x \in \Omega_h$ corresponds to a linear equation that u_h has to satisfy) is defined on the set Ω_h of discrete points instead of the continuous set Ω, we call (2.5) a *discretization* of the potential equation (2.1). In this particular case, all differential operators are replaced by difference quotients involving a finite number of values, giving this approach the name *finite difference method.*

2.2. Stability and convergence

Merely formulating the discrete system (2.5), is not enough, we also have to investigate whether this system can be solved, whether the solution is unique, and whether it approximates the continuous solution u.

If is easy to see that $-\Delta_h$ is a linear mapping from $G_0(\bar{\Omega}_h)$ to $G(\Omega_h)$ and that

$$\dim G_0(\bar{\Omega}_h) = \dim G(\Omega_h) = N^2$$

holds. In order to prove that the system (2.5) has a unique solution, it is enough to prove that $-\Delta_h$ is an injective mapping.

A particularly elegant way of proving this result is to use the following stability result for the maximum norm:

Lemma 2.7 (Maximum principle) *Let $v_h \in G(\bar{\Omega}_h)$ denote a grid function satisfying*

$$-\Delta_h v_h(x) \leq 0 \qquad\qquad \text{for all } x \in \Omega_h.$$

There exists a boundary point $x_0 \in \partial\Omega_h$ such that

$$v_h(x) \leq v_h(x_0) \qquad\qquad \text{for all } x \in \bar{\Omega}_h,$$

i.e., the grid function takes its maximum at the boundary.

Proof. We define the sets of neighbours of points x by

$$N(x) := \{(x_1 - h, x_2), (x_1 + h, x_2), (x_1, x_2 - h), (x_1, x_2 + h)\} \qquad \text{for all } x \in \Omega_h.$$

The distance (with respect to the grid) from a grid point to the boundary is denoted by

$$\delta\colon \bar{\Omega}_h \to \mathbb{N}_0, \qquad x \mapsto \begin{cases} 0 & \text{if } x \in \partial\Omega_h, \\ 1 + \min\{\delta(x') \ : \ x' \in N(x)\} & \text{otherwise.} \end{cases}$$

We denote the maximum of v_h by

$$m := \max\{v_h(x) \ : \ x \in \bar{\Omega}_h\}$$

and intend to prove by induction

$$(v_h(x) = m \wedge \delta(x) \leq d) \Longrightarrow \exists x_0 \in \partial\Omega_h : \ v_h(x_0) = m \tag{2.6}$$

for all $d \in \mathbb{N}_0$ and all $x \in \bar{\Omega}_h$. This implication yields our claim since $\delta(x)$ is finite for all $x \in \bar{\Omega}$.

The base case $d = 0$ of the induction is straightforward: if $x \in \bar{\Omega}_h$ with $v_h(x) = m$ and $\delta(x) = d = 0$ exists, the definition of δ already implies $x \in \partial\Omega_h$, so we can choose $x_0 = x$.

Let now $d \in \mathbb{N}_0$ satisfy (2.6). Let $x \in \bar{\Omega}_h$ be given with $\delta(x) = d + 1$ and $v_h(x) = m$. This implies $x \in \Omega_h$ and we obtain

$$\sum_{x' \in N(x)} h^{-2}(v_h(x) - v_h(x')) = 4h^{-2}v_h(x) - \sum_{x' \in N(x)} h^{-2}v_h(x') = -\Delta_h v_h(x) \leq 0.$$

Since $m = v_h(x)$ is the maximum of v_h, none of the summands on the left side of this inequality can be negative. Since the sum cannot be positive, all summands have to be equal to zero, and this implies

$$m = v_h(x) = v_h(x') \qquad \text{for all } x' \in N(x).$$

Due to $\delta(x) = d + 1$, there has to be a $x' \in N(x)$ with $\delta(x') = d$, and since we have just proven $v_h(x') = m$, we can apply the induction assumption to complete the proof. ■

The maximum principle already guarantees the injectivity of the differen operator $-\Delta_h$ and the existence of a unique solution.

Corollary 2.8 (Unique solution) *The system of linear equations (2.5) has a unique solution.*

Proof. Let $u_h, \tilde{u}_h \in G_0(\bar{\Omega}_h)$ be given with

$$\begin{aligned} -\Delta_h u_h(x) &= f(x) && \text{for all } x \in \Omega_h, \\ -\Delta_h \tilde{u}_h(x) &= f(x) && \text{for all } x \in \Omega_h. \end{aligned}$$

We let $v_h := u_h - \tilde{u}_h$ and obtain

$$\Delta_h v_h(x) = \Delta_h u_h(x) - \Delta_h \tilde{u}_h(x) = -f(x) + f(x) = 0 \qquad \text{for all } x \in \Omega_h.$$

The requirements of Lemma 2.7 are fulfilled, so the grid function v_h has to take its maximum at the boundary $\partial\Omega_h$. Due to $v_h \in G_0(\bar{\Omega}_h)$, we have $v_h|_{\partial\Omega_h} = 0$, and therefore

$$v_h(x) \leq 0 \qquad \text{for all } x \in \Omega_h.$$

We can apply the same argument to the grid function $\tilde{v}_h := \tilde{u}_h - u_h = -v_h$ to obtain

$$v_h(x) = -\tilde{v}_h(x) \geq 0 \qquad \text{for all } x \in \Omega_h,$$

and this yields $v_h = 0$ and $u_h = \tilde{u}_h$. We have proven that Δ_h is injective.

Due to $\dim G(\Omega_h) = \dim G_0(\bar{\Omega}_h)$, the rank-nullity theorem implies that Δ_h also has to be surjective. ■

Since Lemma 2.7 only requires $\Delta_h v_h$ not to be negative in any point $x \in \Omega_h$, we can also use it to obtain the following stability result that guarantees that small perturbations of the right-hand side of (2.5) are not significantly amplified.

Lemma 2.9 (Stability) *Let $u_h \in G_0(\bar{\Omega}_h)$ a grid function with homogeneous Dirichlet boundary conditions. We have*

$$\|u_h\|_{\infty,\Omega_h} \leq \frac{1}{8}\|\Delta_h u_h\|_{\infty,\Omega_h}.$$

Proof. (cf. [7, Theorem 4.4.1]) The key idea of our proof is to consider the function

$$w\colon \bar{\Omega} \to \mathbb{R}_{\geq 0}, \qquad x \mapsto \frac{x_1}{2}(1 - x_1).$$

Since it is quadratic polynomial, we have $|w|_{4,\Omega} = 0$, and we can combine

$$-\Delta w(x) = 1 \qquad \text{for all } x \in \Omega$$

with (2.4) to obtain

$$-\Delta_h w_h(x) = 1 \qquad \text{for all } x \in \Omega_h$$

with the grid function $w_h := w|_{\bar{\Omega}_h} \in G(\bar{\Omega}_h)$.

We denote the minimum and maximum of $-\Delta_h u_h$ by

$$\alpha := \min\{-\Delta_h u_h(x) \ : \ x \in \Omega_h\},$$
$$\beta := \max\{-\Delta_h u_h(x) \ : \ x \in \Omega_h\}$$

and define

$$u_h^+ := w_h \beta.$$

This implies

$$-\Delta_h u_h^+(x) = -\Delta_h w_h(x)\beta = \beta \qquad \text{for all } x \in \Omega_h,$$

so we also have

$$-\Delta_h(u_h - u_h^+)(x) = -\Delta_h u_h(x) - \beta \leq 0 \qquad \text{for all } x \in \Omega_h.$$

Let $x \in \Omega_h$. Lemma 2.7 yields a boundary point $x_0 \in \partial\Omega_h$ such that

$$u_h(x) - u_h^+(x) \leq u_h(x_0) - u_h^+(x_0).$$

Due to the Dirichlet boundary conditions, we have $u_h(x_0) = 0$ and conclude

$$u_h(x) \leq u_h^+(x) - u_h^+(x_0).$$

It is easy to prove $0 \leq w(z) \leq 1/8$ for all $z \in \bar{\Omega}_h$, which implies $u_h^+(x) - u_h^+(x_0) \leq \beta/8$. Since x is arbitrary, we have proven

$$u_h(x) \leq \frac{1}{8}\beta \qquad \text{for all } x \in \Omega_h.$$

Since $-u_h$ is bounded from above by $-\alpha$, we can apply the same arguments to $-u_h$ to get

$$u_h(x) \geq \frac{1}{8}\alpha \qquad \text{for all } x \in \Omega_h.$$

Combining both estimates yields

$$\|u_h\|_{\infty,\Omega_h} \leq \frac{1}{8}\max\{|\alpha|, |\beta|\} = \frac{1}{8}\|\Delta_h u_h\|_{\infty,\Omega_h}.$$

■

Combining this stability result with the consistency result of Lemma 2.4 we can prove the *convergence* of our discretization scheme.

Theorem 2.10 (Convergence) *Let $u \in C^4(\bar{\Omega})$ be the solution of (2.1, and let $u_h \in G_0(\Omega_h)$ be the solution of (2.5). We have*

$$\|u - u_h\|_{\infty,\Omega_h} \leq \frac{h^2}{48}|u|_{4,\Omega}.$$

Proof. Due to (2.1), we have

$$f(x) = -\Delta u(x) \qquad \text{for all } x \in \Omega_h.$$

The consistency result of Lemma 2.4 yields

$$|\Delta_h u(x) - \Delta_h u_h(x)| = |\Delta_h u(x) + f(x)|$$

$$= |\Delta_h u(x) - \Delta u(x)| \leq \frac{h^2}{6} |u|_{4,\Omega} \qquad \text{for all } x \in \Omega_h,$$

which is equivalent to

$$\|\Delta_h (u - u_h)\|_{\infty,\Omega_h} = \|\Delta_h u - \Delta_h u_h\|_{\infty,\Omega_h} \leq \frac{h^2}{6} |u|_{4,\Omega}.$$

Now we can apply the stability result of Lemma 2.9 to get

$$\|u - u_h\|_{\infty,\Omega_h} \leq \frac{1}{8} \|\Delta_h u - \Delta_h u_h\|_{\infty,\Omega_h} \leq \frac{1}{8} \frac{h^2}{6} |u|_{4,\Omega}.$$

■

If we can solve the linear system (2.5), we can expect the solution to converge to approximate $u|_{\Omega_h}$ at a rate of h^2. In order to express the linear system in terms of matrices and vectors instead of general linear operators, we have to introduce suitable bases for the spaces $G_0(\bar{\Omega}_h)$ and $G(\Omega_h)$. A straightforward choice is the basis $(\varphi_y)_{y \in \Omega_h}$ consisting of the functions

$$\varphi_y(x) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } x \in \bar{\Omega}_h,$$

that are equal to one in y and equal to zero everywhere else and obviously form a basis of $G_0(\bar{\Omega}_h)$. Restricting the functions to $G(\Omega_h)$ yields a basis of this space, as well. Expressing $-\Delta_h$ in these bases yields a matrix $\mathbf{L} \in \mathbb{R}^{\Omega_h \times \Omega_h}$ given by

$$(\ell_h)_{x,y} := \begin{cases} 4h^{-2} & \text{if } x = y \\ -h^{-2} & \text{if } |x_1 - y_1| = h, \ x_2 = y_2, \\ -h^{-2} & \text{if } x_1 = y_1, \ |x_2 - y_2| = h, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } x, y \in \Omega_h.$$

Expressing the grid function u_h and f_h in these bases yields vectors $\mathbf{u}_h, \mathbf{f}_h \in \mathbb{R}^{\Omega_h}$ and the discretized potential equation (2.5) takes the form

$$\mathbf{L}_h \mathbf{u}_h = \mathbf{f}_h. \tag{2.7}$$

Since (2.5) has a unique solution, the same holds for (2.7).

The matrix $\mathbf{L}_h$ is particularly benign: a glance at the coefficients yields $\mathbf{L}_h = \mathbf{L}_h^*$, so the matrix is symmetric. Applying the stability result of Lemma 2.9 to subsets $\omega_h \subseteq \Omega_h$ shows that not only $\mathbf{L}_h$ is invertible, but also all of its principal submatrices $\mathbf{L}_h|_{\omega_h \times \omega_h}$. This property guarantees that $\mathbf{L}_h$ possesses an invertible LR factorization that can be used to solve the system (2.7). We can even prove that $\mathbf{L}_h$ is positive definite, so we can use the more efficient Cholesky factorization.

For large values of N, i.e., for high accuracies, this approach is not particularly useful, since it does not take advantage of the special structure of $\mathbf{L}_h$: every row and column contains by definition not more than five non-zero coefficients. Matrices with the property that only a small number of entries per row or column are non-zero are called *sparse*, and this property can be used to carry out matrix-vector multiplications efficiently and even to solve the linear system.

Exercise 2.11 (First derivative) *If we approximate the one-dimensional differential equation*

$$u'(x) = f(x) \qquad \text{for all } x \in (0,1)$$

by the central difference quotient introduced in Exercise 2.2, we obtain a matrix $\mathbf{L} \in \mathbb{R}^{N\times N}$ *given by*

$$\ell_{ij} = \begin{cases} 1/(2h) & \text{if } j = i+1, \\ -1/(2h) & \text{if } j = i-1, \\ 0 & \text{otherwise}, \end{cases} \qquad \text{for all } i,j \in [1:N].$$

Prove that $\mathbf{L}$ *is* not *invertible if* N *is odd.*

Remark 2.12 (General domains) *Finite difference discretization are particularly well-suited for differential equations on "simple" domains like the unit square investigated here. Treating more complicated domains requires us to use more involved techniques like the* Shortley-Weller discretization *and may significantly increase the complexity of the resulting algorithms.*

2.3. Diagonal dominance and invertibility

The finite difference approach can be applied to treat more general partial differential equations: we simply have to replace all differential operators by suitable difference quotients. While the consistency of these schemes can usually be proven by using suitable Taylor expansions, the stability poses a challenge.

We investigate linear systems of equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.8}$$

with a matrix $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$, a right-hand side $\mathbf{b} \in \mathbb{R}^{\mathcal{I}}$ and a solution $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$. A generalization of the stability Lemma 2.9 would look like

$$\|\mathbf{x}\|_\infty \leq C\|\mathbf{A}\mathbf{x}\|_\infty \qquad \text{for all } \mathbf{x} \in \mathbb{R}^{\mathcal{I}}$$

with a constant $C \in \mathbb{R}_{\geq 0}$. This inequality can only hold if $\mathbf{A}$ is injective, i.e., invertible, and we can rewrite it in the form

$$\|\mathbf{A}^{-1}\mathbf{b}\|_\infty \leq C\|\mathbf{b}\|_\infty \qquad \text{for all } \mathbf{b} \in \mathbb{R}^{\mathcal{I}}$$

by substituting $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

We recall that any norm $\|\cdot\|$ for $\mathbb{R}^{\mathcal{I}}$ induces the operator norm

$$\|\mathbf{A}\| := \sup\left\{\frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \ :\ \mathbf{x} \in \mathbb{R}^{\mathcal{I}} \setminus \{0\}\right\} \qquad \text{for all } \mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}. \tag{2.9}$$

Stability therefore simply means that we have to be able to find an upper bound for $\|\mathbf{A}^{-1}\|_\infty$ that is independent of the mesh parameter h.

Lemma 2.13 (Neumann series) *Let $\|\cdot\|$ be a norm for $\mathbb{R}^{\mathcal{I}}$, and let Let $\mathbf{X} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$.*

If we have $\|\mathbf{X}\| < 1$, the matrix $\mathbf{I} - \mathbf{X}$ is invertible with

$$\sum_{\ell=0}^{\infty} \mathbf{X}^\ell = (\mathbf{I} - \mathbf{X})^{-1}, \qquad \|(\mathbf{I} - \mathbf{X})^{-1}\| \leq \frac{1}{1 - \|\mathbf{X}\|}.$$

Proof. Let $\|\mathbf{X}\| < 1$. We define the partial sums

$$\mathbf{Y}_m := \sum_{\ell=0}^{m} \mathbf{X}^\ell \qquad \text{for all } m \in \mathbb{N}_0.$$

In order to prove that $(\mathbf{Y}_m)_{m=0}^\infty$ is a Cauchy sequency, we first observe

$$\begin{aligned} \|\mathbf{Y}_m - \mathbf{Y}_n\| &= \left\|\sum_{\ell=n+1}^{m} \mathbf{X}^\ell\right\| \leq \sum_{\ell=n+1}^{m} \|\mathbf{X}\|^\ell = \|\mathbf{X}\|^{n+1} \sum_{\ell=0}^{m-n-1} \|\mathbf{X}\|^\ell \\ &\leq \|\mathbf{X}\|^{n+1} \sum_{\ell=0}^{\infty} \|\mathbf{X}\|^\ell = \frac{\|\mathbf{X}\|^{n+1}}{1 - \|\mathbf{X}\|} \qquad \text{for all } n, m \in \mathbb{N}_0 \text{ with } n < m. \end{aligned}$$

Given $\epsilon \in \mathbb{R}_{>0}$, we can find $n_0 \in \mathbb{N}$ with $\|\mathbf{X}\|^{n_0+1} \leq (1 - \|\mathbf{X}\|)\epsilon$, and this implies

$$\|\mathbf{Y}_m - \mathbf{Y}_n\| \leq \frac{\|\mathbf{X}\|^{n+1}}{1 - \|\mathbf{X}\|} \leq \frac{\|\mathbf{X}\|^{n_0+1}}{1 - \|\mathbf{X}\|} \leq \epsilon \qquad \text{for all } n, m \in \mathbb{N}_0,\ n_0 \leq n < m.$$

We conclude that $(\mathbf{Y}_m)_{m=0}^\infty$ is a Cauchy sequency and therefore has a limit

$$\mathbf{Y} := \lim_{m\to\infty} \mathbf{Y}_m = \sum_{\ell=0}^{\infty} \mathbf{X}^\ell$$

satisfying

$$\|\mathbf{Y}\| = \left\|\sum_{\ell=0}^{\infty} \mathbf{X}^\ell\right\| \leq \sum_{\ell=0}^{\infty} \|\mathbf{X}\|^\ell = \frac{1}{1 - \|\mathbf{X}\|}.$$

Due to

$$\begin{aligned} (\mathbf{I} - \mathbf{X})\mathbf{Y} &= (\mathbf{I} - \mathbf{X}) \sum_{\ell=0}^{\infty} \mathbf{X}^\ell = \sum_{\ell=0}^{\infty} \mathbf{X}^\ell - \sum_{\ell=0}^{\infty} \mathbf{X}^{\ell+1} = \sum_{\ell=0}^{\infty} \mathbf{X}^\ell - \sum_{\ell=1}^{\infty} \mathbf{X}^\ell = \mathbf{I}, \\ \mathbf{Y}(\mathbf{I} - \mathbf{X}) &= \sum_{\ell=0}^{\infty} \mathbf{X}^\ell(\mathbf{I} - \mathbf{X}) = \sum_{\ell=0}^{\infty} \mathbf{X}^\ell - \sum_{\ell=0}^{\infty} \mathbf{X}^{\ell+1} = \sum_{\ell=0}^{\infty} \mathbf{X}^\ell - \sum_{\ell=1}^{\infty} \mathbf{X}^\ell = \mathbf{I}, \end{aligned}$$

we finally obtain $\mathbf{Y} = (\mathbf{I} - \mathbf{X})^{-1}$. ■

Exercise 2.14 (Generalized convergence criterion) *Let $\|\cdot\|$ be a norm for $\mathbb{R}^{\mathcal{I}}$ and let $\mathbf{X} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$. Assume that there is a $k \in \mathbb{N}$ such that $\|\mathbf{X}^k\| < 1$. Prove*

$$\sum_{\ell=0}^{\infty} \mathbf{X}^\ell = (\mathbf{I} - \mathbf{X})^{-1}, \qquad \|(\mathbf{I} - \mathbf{X})^{-1}\| \leq \frac{\sum_{m=0}^{k-1} \|\mathbf{X}^m\|}{1 - \|\mathbf{X}^k\|}.$$

In order to be able to apply Lemma 2.13, we have to be able to find an upper bound for the operator norm. In the case of the maximum norm, this is particularly simple.

Lemma 2.15 (Maximum norm) *We have*

$$\|\mathbf{X}\|_\infty = \max\left\{ \sum_{j\in\mathcal{I}} |x_{ij}| \ : \ i \in \mathcal{I} \right\} \qquad \text{for all } \mathbf{X} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}.$$

Proof. Let $\mathbf{X} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ and set

$$\mu := \max\left\{ \sum_{j\in\mathcal{I}} |x_{ij}| \ : \ i \in \mathcal{I} \right\}.$$

Let $\mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ and $i \in \mathcal{I}$. We have

$$|(\mathbf{X}\mathbf{y})_i| = \left| \sum_{j\in\mathcal{I}} x_{ij} y_j \right| \leq \sum_{j\in\mathcal{I}} |x_{ij}|\,|y_j| \leq \sum_{j\in\mathcal{I}} |x_{ij}|\,\|\mathbf{y}\|_\infty \leq \mu \|\mathbf{y}\|_\infty$$

and conclude $\|\mathbf{X}\|_\infty \leq \mu$.

Now we fix $i \in \mathcal{I}$ such that

$$\mu = \sum_{j\in\mathcal{I}} |x_{ij}|.$$

If we introduce the vector $\mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ given by

$$y_j := \begin{cases} -1 & \text{if } x_{ij} < 0 \\ 1 & \text{otherwise} \end{cases} \qquad \text{for all } j \in \mathcal{I},$$

we find $\|\mathbf{y}\|_\infty = 1$ and

$$\mu = \sum_{j\in\mathcal{I}} |x_{ij}| = \sum_{j\in\mathcal{I}} x_{ij} y_j = (\mathbf{X}\mathbf{y})_i \leq \|\mathbf{X}\mathbf{y}\|_\infty = \frac{\|\mathbf{X}\mathbf{y}\|_\infty}{\|\mathbf{y}\|} \leq \|\mathbf{X}\|_\infty.$$

∎

Using the maximum norm and the Neumann series, we can find a simple criterion that allows us to check whether a given matrix is invertible: the diagonal elements have to be large enough.

Definition 2.16 (Diagonally dominant matrices) *A matrix* $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}}$ *with* $\mathcal{I} \subseteq \mathcal{J}$ *is called* weakly diagonally dominant *if*

$$\sum_{\substack{j\in\mathcal{J}\\ j\neq i}} |a_{ij}| \leq |a_{ii}| \qquad \textit{for all } i \in \mathcal{I}.$$

It is called strictly diagonally dominant *if*

$$\sum_{\substack{j\in\mathcal{J}\\ j\neq i}} |a_{ij}| < |a_{ii}| \qquad \textit{for all } i \in \mathcal{I}.$$

Using the Neumann series, it is possible to prove that strictly diagonally dominant matrices are invertible.

Lemma 2.17 (Strictly diagonally dominant) *Let* $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ *be strictly diagonally dominant. Then* $\mathbf{A}$ *is invertible.*

Proof. Since $\mathbf{A}$ is strictly diagonally dominant, we have

$$a_{ii} \neq 0 \qquad \text{for all } i \in \mathcal{I},$$

so the diagonal part $\mathbf{D} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ of $\mathbf{A}$, given by

$$d_{ij} = \begin{cases} a_{ii} & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in \mathcal{I},$$

is invertible. The matrix

$$\mathbf{M} := \mathbf{I} - \mathbf{D}^{-1}\mathbf{A} \tag{2.10}$$

satisfies

$$m_{ii} = 1 - \frac{a_{ii}}{a_{ii}} = 0 \qquad \text{for all } i \in \mathcal{I}.$$

Since $\mathbf{A}$ is strictly diagonally dominant, we also have

$$\sum_{j\in\mathcal{I}} |m_{ij}| = \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |m_{ij}| = \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} \frac{|a_{ij}|}{|a_{ii}|} = \frac{1}{|a_{ii}|} \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |a_{ij}| < 1 \qquad \text{for all } i \in \mathcal{I},$$

and we can conclude $\|\mathbf{M}\|_\infty < 1$ by Lemma 2.15.

Now Lemma 2.13 yields that $\mathbf{I} - \mathbf{M} = \mathbf{D}^{-1}\mathbf{A}$ is invertible, and this implies that the matrix $\mathbf{A}$ itself also has to be invertible. ■

Remark 2.18 (Jacobi iteration) *Lemma 2.17 is, in fact, not only a proof of the* existence *of the inverse* $\mathbf{A}^{-1}$*, but also suggests a practical algorithm: in order to solve the linear system (2.8), we choose an arbitrary vector* $\mathbf{x}^{(0)}$*, and consider the sequence* $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ *given by*

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \mathbf{D}^{-1}(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}) \qquad \text{for all } m \in \mathbb{N}_0.$$

The difference between these vectors and the solution $\mathbf{x}$ *satisfies*

$$\begin{aligned} \mathbf{x}^{(m+1)} - \mathbf{x} &= \mathbf{x}^{(m)} - \mathbf{x} - \mathbf{D}^{-1}(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}) \\ &= \mathbf{x}^{(m)} - \mathbf{x} - \mathbf{D}^{-1}\mathbf{A}(\mathbf{x}^{(m)} - \mathbf{x}) = \mathbf{M}(\mathbf{x}^{(m)} - \mathbf{x}) \qquad \text{for all } m \in \mathbb{N}_0. \end{aligned}$$

Due to $\|\mathbf{M}\|_\infty < 1$*, we obtain*

$$\lim_{m\to\infty} \mathbf{x}^{(m)} = \mathbf{x},$$

i.e., we can compute the solution of the linear system by iteratively multiplying by $\mathbf{A}$ *and dividing by the diagonal elements. If the matrix-vector multiplication can be realized efficiently, one step of the iteration takes only a small amount of time.*

This algorithm is know as the Jacobi iteration*.*

2.4. Convergence of the Neumann series

We have seen that *strictly* diagonally dominant matrices are invertible and that we can approximate the inverse by the Neumann series and the solution of the linear system (2.8) by the Jacobi iteration.

Unfortunately, the matrices associated with partial differential equations are usually not strictly diagonally dominant: any reasonable difference quotient will yield the value zero if applied to the constant function, and this implies

$$\sum_{j\in\mathcal{I}} a_{ij} = 0$$

for all grid points $i \in \mathcal{I}$ that are not adjacent to the boundary. Obviously, this means

$$|a_{ii}| = \left| \sum_{\substack{j\in\mathcal{I} \\ j\neq i}} a_{ij} \right| \leq \sum_{\substack{j\in\mathcal{I} \\ j\neq i}} |a_{ij}|,$$

so the best we can hope for is a *weakly* diagonally dominant matrix, and the simple example

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

indicates that weakly diagonally dominant matrices may be not invertible. If we want to ensure that $\mathbf{A}^{-1}$ exists, we have to include additional conditions.

The proof of Lemma 2.17 relies on the fact that the Neumann series for the matrix $\mathbf{M}$ converges. Lemma 2.13 states that this is the case if $\|\mathbf{M}\| < 1$ holds, but this is only a *sufficient* condition, not a necessary one: for any $x \in \mathbb{R}$,

$$\mathbf{M}_x = \begin{pmatrix} 0 & x \\ 0 & 0 \end{pmatrix}$$

satisfies $\mathbf{M}_x^2 = \mathbf{0}$, so the Neumann series for this matrix always converges. On the other hand, given any norm $\|\cdot\|$, we can find an $x \in \mathbb{R}$ with $\|\mathbf{M}_x\| \geq 1$.

Definition 2.19 (Spectral radius) *Let* $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$. $\lambda \in \mathbb{C}$ *is called an* eigenvalue *of* $\mathbf{X}$ *if an* eigenvector $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ *exists such that*

$$\mathbf{X}\mathbf{e} = \lambda\mathbf{e}.$$

The set

$$\sigma(\mathbf{X}) := \{\lambda \in \mathbb{C} \ : \ \lambda \textit{ is an eigenvalue of } \mathbf{X}\}$$

is called the spectrum *of* $\mathbf{X}$*. The maximum of the eigenvalues' absolute values*

$$\varrho(\mathbf{X}) := \max\{|\lambda| \ : \ \lambda \in \sigma(\mathbf{X})\}$$

is called the spectral radius *of* $\mathbf{X}$*.*

Lemma 2.20 (Necessary condition) *Let* $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$*. If the sequence* $(\mathbf{X}^\ell)_{\ell=0}^\infty$ *converges to zero, we have* $\varrho(\mathbf{X}) < 1$*.*

Proof. By contraposition.

Let $\varrho(\mathbf{X}) \geq 1$. Then we can find an eigenvalue $\lambda \in \sigma(\mathbf{X})$ with $|\lambda| \geq 1$. Let $\mathbf{e} \in \mathbb{R}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ be a matching eigenvector. We have

$$\mathbf{X}^\ell\mathbf{e} = \lambda^\ell\mathbf{e}, \qquad \|\mathbf{X}^\ell\mathbf{e}\| = |\lambda|^\ell\|\mathbf{e}\| \geq \|\mathbf{e}\| \qquad \text{for all } \ell \in \mathbb{N}_0,$$

and this implies that $(\mathbf{X}^\ell)_{\ell=0}^\infty$ cannot converge to zero. ■

The Neumann series can only converge if $(\mathbf{X}^\ell)_{\ell=0}^\infty$ converges to zero, so $\varrho(\mathbf{X}) < 1$ is a *necessary* condition for its convergence. We will now prove that it is also sufficient, i.e., that the convergence of the Neumann series can be characterized by the spectral radius.

Theorem 2.21 (Schur decomposition) *Let* $\mathbf{X} \in \mathbb{C}^{n\times n}$*. There are an upper triangular matrix* $\mathbf{R} \in \mathbb{C}^{n\times n}$ *and a unitary matrix* $\mathbf{Q} \in \mathbb{C}^{n\times n}$ *such that*

$$\mathbf{Q}^*\mathbf{X}\mathbf{Q} = \mathbf{R}.$$

Proof. By induction.

Base case: For $n = 1$, any matrix $\mathbf{X} \in \mathbb{C}^{1\times 1}$ already is upper triangular, so we can choose $\mathbf{Q} = \mathbf{I}$.

Inductive step: Let $n \in \mathbb{N}$ be such that our claim holds for all matrices $\mathbf{X} \in \mathbb{C}^{n\times n}$. Let $\mathbf{X} \in \mathbb{C}^{(n+1)\times(n+1)}$.

By the fundamental theorem of algebra, the characteristic polynomial $p_X(t) = \det(t\mathbf{I} - \mathbf{X})$ has at least one zero $\lambda \in \mathbb{C}$. Since then $\lambda\mathbf{I} - \mathbf{X}$ is singular, we can find an eigenvector $\mathbf{e} \in \mathbb{C}^{n+1}$, and we can use scaling to ensure $\|\mathbf{e}\|_2 = 1$.

Let $\mathbf{Q}_0 \in \mathbb{C}^{(n+1)\times(n+1)}$ be the Householder reflection with $\mathbf{Q}_0\delta = \mathbf{e}$, where δ denotes the first canonical unit vector. We find

$$\mathbf{Q}_0^*\mathbf{X}\mathbf{Q}_0 = \begin{pmatrix} \lambda & \mathbf{R}_0 \\ & \widehat{\mathbf{X}} \end{pmatrix}$$

for $\mathbf{R}_0 \in \mathbb{C}^{1\times n}$ and $\widehat{\mathbf{X}} \in \mathbb{C}^{n\times n}$.

Now we can apply the induction assumption to find an upper triangular matrix $\widehat{\mathbf{R}} \in \mathbb{C}^{n\times n}$ and a unitary matrix $\widehat{\mathbf{Q}} \in \mathbb{C}^{n\times n}$ such that

$$\widehat{\mathbf{Q}}^*\widehat{\mathbf{X}}\widehat{\mathbf{Q}} = \widehat{\mathbf{R}}.$$

We let

$$\mathbf{Q} := \mathbf{Q}_0 \begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}} \end{pmatrix}, \qquad \mathbf{R} := \begin{pmatrix} \lambda & \mathbf{R}_0 \\ & \mathbf{R} \end{pmatrix},$$

observe that $\mathbf{Q}$ is a product of unitary matrices and therefore unitary itself, and conclude

$$\mathbf{Q}^*\mathbf{X}\mathbf{Q} = \begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}}^* \end{pmatrix} \mathbf{Q}_0^*\mathbf{X}\mathbf{Q}_0 \begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}} \end{pmatrix} = \begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}}^* \end{pmatrix} \begin{pmatrix} \lambda & \mathbf{R}_0 \\ & \widehat{\mathbf{X}} \end{pmatrix} \begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}} \end{pmatrix} = \begin{pmatrix} \lambda & \mathbf{R}_0 \\ & \widehat{\mathbf{R}} \end{pmatrix} = \mathbf{R}.$$

Since $\widehat{\mathbf{R}}$ is upper triangular, so is $\mathbf{R}$. ■

Using the Schur decomposition, we can investigate the relationship between the spectral radius and matrix norms.

Lemma 2.22 (Spectral radius and operator norms) *Let $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$. We have*

$$\varrho(\mathbf{X}) \leq \|\mathbf{X}\|$$

for any operator norm induced by a norm $\|\cdot\|$ for $\mathbb{C}^{\mathcal{I}}$.

Given an $\epsilon \in \mathbb{R}_{>0}$, we can find a norm $\|\cdot\|_{X,\epsilon}$ such that the corresponding operator norm satisfies

$$\|\mathbf{X}\|_{X,\epsilon} \leq \varrho(\mathbf{X}) + \epsilon.$$

Proof. We may assume $\mathcal{I} = [1 : n]$ without loss of generality.

Let $\|\cdot\|$ be a norm for $\mathbb{C}^n$. Let $\lambda \in \sigma(\mathbf{X})$, and let $\mathbf{e} \in \mathbb{C}^n$ be a corresponding eigenvector. We have

$$\|\mathbf{X}\mathbf{e}\| = \|\lambda\mathbf{e}\| = |\lambda|\ \|\mathbf{e}\|,$$

and the definition (2.9) of the operator norm yields $\|\mathbf{X}\| \geq |\lambda|$, i.e., $\|\mathbf{X}\| \geq \varrho(\mathbf{X})$.

Let now $\epsilon \in \mathbb{R}_{>0}$. Due to Theorem 2.21, we can find a unitary matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{n\times n}$ such that

$$\mathbf{Q}^*\mathbf{X}\mathbf{Q} = \mathbf{R},$$

and since unitary matrices leave the Euclidean norm invariant, we have

$$\|\mathbf{X}\|_2 = \|\mathbf{R}\|_2.$$

We split $\mathbf{R} \in \mathbb{C}^{n\times n}$ into the diagonal $\mathbf{D} \in \mathbb{C}^{n\times n}$ and the upper triangular part $\mathbf{N}$, given by

$$d_{ij} = \begin{cases} r_{ii} & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \qquad n_{ij} = \begin{cases} r_{ij} & \text{if } i < j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i,j \in [1:n].$$

We have $\mathbf{R} = \mathbf{D} + \mathbf{N}$ and $\|\mathbf{D}\|_2 = \varrho(\mathbf{R}) = \varrho(\mathbf{X})$, so we only have to take care of $\mathbf{N}$.

For a given $\delta \in \mathbb{R}_{>0}$, we can define the diagonal matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ by

$$s_{ij} = \begin{cases} \delta^i & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i,j \in [1:n].$$

We observe $\mathbf{S}^{-1}\mathbf{D}\mathbf{S} = \mathbf{D}$ and

$$(\mathbf{S}^{-1}\mathbf{N}\mathbf{S})_{ij} = \delta^{j-i} n_{ij} \qquad \text{for all } i,j \in [1:n].$$

We choose δ small enough to ensure $\|\mathbf{S}^{-1}\mathbf{N}\mathbf{S}\|_2 \le \epsilon$.

We define the norm

$$\|\mathbf{y}\|_{X,\epsilon} := \|\mathbf{S}^{-1}\mathbf{Q}^*\mathbf{y}\|_2 \qquad \text{for all } \mathbf{y} \in \mathbb{C}^{\mathcal{I}}$$

and observe

$$\begin{aligned} \|\mathbf{X}\mathbf{y}\|_{X,\epsilon} &= \|\mathbf{S}^{-1}\mathbf{Q}^*\mathbf{X}\mathbf{y}\|_2 = \|\mathbf{S}^{-1}\mathbf{Q}^*\mathbf{X}\mathbf{Q}\mathbf{S}(\mathbf{S}^{-1}\mathbf{Q}^*\mathbf{y})\|_2 \le \|\mathbf{S}^{-1}\mathbf{Q}^*\mathbf{X}\mathbf{Q}\mathbf{S}\|_2\|\mathbf{S}^{-1}\mathbf{Q}^*\mathbf{y}\|_2 \\ &= \|\mathbf{S}^{-1}\mathbf{R}\mathbf{S}\|_2\|\mathbf{y}\|_{X,\epsilon} \qquad \text{for all } \mathbf{y} \in \mathbb{C}^{\mathcal{I}}, \end{aligned}$$

which implies $\|\mathbf{X}\|_{X,\epsilon} \le \|\mathbf{S}^{-1}\mathbf{R}\mathbf{S}\|_2$.

Due to $\mathbf{R} = \mathbf{D} + \mathbf{N}$, we can use the triangle inequality to obtain

$$\|\mathbf{X}\|_{X,\epsilon} = \|\mathbf{S}^{-1}(\mathbf{D}+\mathbf{N})\mathbf{S}\|_2 \le \|\mathbf{S}^{-1}\mathbf{D}\mathbf{S}\|_2 + \|\mathbf{S}^{-1}\mathbf{N}\mathbf{S}\|_2 \le \|\mathbf{D}\|_2 + \epsilon = \varrho(\mathbf{X}) + \epsilon,$$

completing the proof. ■

Corollary 2.23 (Neumann series) *Let $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$. The Neumann series converges if and only if $\varrho(\mathbf{X}) < 1$. In this case, $\mathbf{I} - \mathbf{X}$ is invertible and we have*

$$\sum_{\ell=0}^{\infty} \mathbf{X}^\ell = (\mathbf{I} - \mathbf{X})^{-1}.$$

Proof. If the Neumann series converges, we have

$$\lim_{\ell\to\infty} \mathbf{X}^\ell = \mathbf{0}.$$

By Lemma 2.20, this implies $\varrho(\mathbf{X}) < 1$.

Let now $\varrho(\mathbf{X}) < 1$, and let $\epsilon := (1 - \varrho(\mathbf{X}))/2$. By Lemma 2.22, we can find a norm $\|\cdot\|_{X,\epsilon}$ such that

$$\|\mathbf{X}\|_{X,\epsilon} \leq \varrho(\mathbf{X}) + \epsilon = \varrho(\mathbf{X}) + (1 - \varrho(\mathbf{X}))/2 = \frac{\varrho(\mathbf{X}) + 1}{2} < 1.$$

Applying Lemma 2.13 with this norm, we conclude that the Neumann series converges to $(\mathbf{I} - \mathbf{X})^{-1}$. ■

2.5. Irreducibly diagonally dominant matrices

In order to apply Corollary 2.23, we need a criterion for estimating the spectral radius of a given matrix. A particularly elegant tool are *Gershgorin discs.*

Theorem 2.24 (Gershgorin discs) *Let $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$. For every index $i \in \mathcal{I}$, the* Gershorin disc *is given by*

$$\mathcal{D}_{X,i} := \Big\{ z \in \mathbb{C} \ : \ |z - x_{ii}| < \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}| \Big\}.$$

We have

$$\sigma(\mathbf{X}) \subseteq \bigcup_{i\in\mathcal{I}} \overline{\mathcal{D}_{X,i}},$$

i.e., every eigenvalue $\lambda \in \sigma(\mathbf{X})$ is contained in the closure of at least one of the Gershgorin discs.

Proof. [10, Theorem 4.6] Let $\lambda \in \sigma(\mathbf{X})$. Let $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ be an eigenvector for λ and $\mathbf{X}$.

We fix $i \in \mathcal{I}$ with

$$|e_j| \leq |e_i| \qquad\qquad \text{for all } j \in \mathcal{I}.$$

Due to $\mathbf{e} \neq \mathbf{0}$, we have $|e_i| > 0$.

Since $\mathbf{e}$ is an eigenvector, we have

$$\lambda e_i = (\mathbf{X}\mathbf{e})_i = \sum_{j\in\mathcal{I}} x_{ij} e_j,$$

and the triangle inequality yields

$$(\lambda - x_{ii}) e_i = \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} x_{ij} e_j,$$

$$|\lambda - x_{ii}|\,|e_i| \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|\,|e_j| \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|\,|e_i|,$$
$$|\lambda - x_{ii}| \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|,$$

i.e., $\lambda \in \overline{\mathcal{D}_{X,i}}$. ■

Exercise 2.25 (Diagonally dominant) *Let $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ be a matrix with non-zero diagonal elements, let $\mathbf{D}$ be its diagonal part, and let $\mathbf{M} := \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$.*

Assume that $\mathbf{A}$ is weakly diagonally dominant. Prove $\varrho(\mathbf{M}) \leq 1$ by Theorem 2.24

Assume that $\mathbf{A}$ is strictly diagonally dominant. Prove $\varrho(\mathbf{M}) < 1$ by Theorem 2.24.

Exercise 2.26 (Invertibility) *Let $\epsilon \in \mathbb{R}_{>0}$, let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be given by*

$$\mathbf{A} = \begin{pmatrix} 3 & \epsilon & & \\ 1/\epsilon & \ddots & \ddots & \\ & \ddots & \ddots & \epsilon \\ & & 1/\epsilon & 3 \end{pmatrix}.$$

Prove $\sigma(\mathbf{A}) \subseteq [1,5]$ and conclude that $\mathbf{A}$ is invertible.

Hints: *All eigenvalues of symmetric matrices are real.*

What is the effect of the similarity transformation with the matrix $\mathbf{S}$ used in the proof of Lemma 2.22 on the matrix $\mathbf{A}$?

Theorem 2.24 states that any eigenvalue of a matrix $\mathbf{X}$ is contained in at least one *closed* Gershgorin disc $\overline{\mathcal{D}_{X,i}}$. In the case of weakly diagonally dominant matrices, we find $\varrho(\mathbf{M}) \leq 1$, but for convergence of the Neumann series we require $\varrho(\mathbf{M}) < 1$, i.e., we need a condition that ensures that no eigenvalue lies on the boundary of the Gershgorin disc.

Definition 2.27 (Irreducible matrix) *Let $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$. We define the sets of* neighbours *by*

$$N(i) := \{j \in \mathcal{I} \ : \ x_{ij} \neq 0\} \qquad \text{for all } i \in \mathcal{I}$$

*(cf. the proof of Lemma 2.7) and the sets of m-*th generation neighbours *by*

$$N_m(i) := \begin{cases} \{i\} & \text{if } m = 0, \\ \bigcup_{j\in N_{m-1}(i)} N(j) & \text{otherwise} \end{cases} \qquad \text{for all } m \in \mathbb{N}_0,\ i \in \mathcal{I}.$$

The matrix $\mathbf{X}$ is called irreducible *if for all $i,j \in \mathcal{I}$ there is an $m \in \mathbb{N}_0$ with $j \in N_m(i)$.*

In the context of finite difference methods, an irreducible matrix corresponds to a grid that allows us to reach any point by traveling from points to their left, right, top, or bottom neighbours. In the case of the unit square and the discrete Laplace operator, this property is obviously guaranteed.

For irreducible matrices, we can obtain the following refined result:

Lemma 2.28 (Gershgorin for irreducible matrices) *Let* $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ *be irreducible, and let the Gershgorin discs be defined as in Theorem 2.24.*

If an eigenvalue $\lambda \in \sigma(\mathbf{X})$ *is not an element of any* open *Gershgorin disc, i.e.,*

$$\lambda \notin \mathcal{D}_{X,i} \qquad \textit{for all } i \in \mathcal{I},$$

it is an element of the boundary of all *Gershgorin discs, i.e., we have*

$$\lambda \in \partial\mathcal{D}_{X,i} \qquad \textit{for all } i \in \mathcal{I}.$$

Proof. [10, Theorem 4.7] Let $\lambda \in \sigma(\mathbf{X})$ be an element of the boundary of the union of all Gershgorin discs, and let $\mathbf{e} \in \mathbb{C}^{\mathcal{I}}$ be a corresponding eigenvector of $\mathbf{X}$.

In a preparatory step, we fix $i \in \mathcal{I}$ with

$$|e_j| \leq |e_i| \qquad \text{for all } j \in \mathcal{I}.$$

As in the proof of Theorem 2.24 we find

$$|\lambda - x_{ii}|\,|e_i| \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|\,|e_j|, \qquad |\lambda - x_{ii}| \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|. \tag{2.11}$$

Our assumption implies that λ cannot be an element of the interior of any Gershgorin disc, so it has to be an element of the boundary of $\mathcal{D}_{X,i}$, i.e.,

$$|\lambda - x_{ii}| \geq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|,$$

and combining this equation with the left estimate in (2.11) yields

$$\sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|\,|e_i| \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|\,|e_j|, \qquad 0 \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |x_{ij}|\,(|e_j| - |e_i|).$$

Due to our choice of $i \in \mathcal{I}$, we have $|e_j| - |e_i| \leq 0$ for all $j \in \mathcal{I}$ and conclude $|e_j| = |e_i|$ for all $j \in \mathcal{I}$ with $j \neq i$ and $x_{ij} \neq 0$, i.e., for all neighbours $j \in N(i)$.

We will now prove $|e_j| = |e_i|$ for all $j \in N_m(i)$ and all $m \in \mathbb{N}_0$ by induction.

Base case: For $m = 0$, we have $N_0(i) = \{i\}$ and the claim is trivial.

Induction step: Let $m \in \mathbb{N}_0$ be such that $|e_j| = |e_i|$ holds for all $j \in N_m(i)$. Let $k \in N_{m+1}(i)$. By definition, there is a $j \in N_m(i)$ such that $k \in N(j)$. Due to the induction assumption, we have $|e_j| = |e_i|$, and by the previous argument we obtain $|e_k| = |e_j| = |e_i|$.

This means that (2.11) holds for all $i \in \mathcal{I}$, and this is equivalent to $\lambda \in \overline{\mathcal{D}_{X,i}}$. Due to $\lambda \notin \mathcal{D}_{X,i}$, we obtain $\lambda \in \partial\mathcal{D}_{X,i}$. ■

Definition 2.29 (Irreducibly diagonally dominant) *Let $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$. We call the matrix* $\mathbf{A}$ irreducibly diagonally dominant *if it is irreducible and weakly diagonally dominant and if there is an index $i \in \mathcal{I}$ with*

$$\sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |a_{ij}| < |a_{ii}|.$$

Lemma 2.30 (Invertible diagonal) *Let $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ be a weakly diagonally dominant, and let $\#\mathcal{I} > 1$. If $\mathbf{A}$ is irreducible, we have $a_{ii} \neq 0$ for all $i \in \mathcal{I}$, i.e., the diagonal of $\mathbf{A}$ is invertible.*

Proof. By contraposition. We assume that there is an index $i \in \mathcal{I}$ with $a_{ii} = 0$. Since $\mathbf{A}$ is weakly diagonally dominant, this implies $a_{ij} = 0$ for all $j \in \mathcal{I}$, i.e., $N(i) = \emptyset$. We obtain $N_1(i) = N(i) = \emptyset$, and a straightforward induction yields $N_m(i) = \emptyset$ for all $m \in \mathbb{N}$. If $\#\mathcal{I} > 1$ holds, we can find $j \in \mathcal{I} \setminus \{i\}$ and conclude $j \notin N_m(i)$ for all $m \in \mathbb{N}_0$, so $\mathbf{A}$ cannot be irreducible. ∎

Corollary 2.31 (Irreducibly diagonally dominant) *Let $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ be irreducibly diagonally dominant, and let $\mathbf{M} := \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$ with the diagonal $\mathbf{D}$ of $\mathbf{A}$.*

The matrix $\mathbf{A}$ is invertible and we have

$$\mathbf{A}^{-1} = \left(\sum_{\ell=0}^{\infty} \mathbf{M}^{\ell}\right) \mathbf{D}^{-1}.$$

Proof. Due to Lemma 2.30, the diagonal matrix $\mathbf{D}$ is invertible and $\mathbf{M}$ is well-defined.

We have already seen that

$$m_{ij} = \begin{cases} 0 & \text{if } i = j,\\ -a_{ij}/a_{ii} & \text{otherwise} \end{cases} \qquad\qquad \text{holds for all } i, j \in \mathcal{I},$$

so $\mathbf{M}$ is irreducible, since $\mathbf{A}$ is.

For every $i \in \mathcal{I}$ we have

$$\sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |m_{ij}| = \frac{1}{|a_{ii}|} \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |a_{ij}| \leq 1,$$

since $\mathbf{A}$ is weakly diagonally dominant. Due to $m_{ii} = 0$, the Gershgorin disc $\mathcal{D}_{M,i}$ is a subset of the disc with radius one around zero. This implies $\varrho(\mathbf{M}) \leq 1$.

We now have to prove $\varrho(\mathbf{M}) < 1$. Due to Definition 2.29, there is an index $i \in \mathcal{I}$ such that

$$\sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |a_{ij}| < |a_{ii}|, \qquad\qquad \alpha := \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} \frac{|a_{ij}|}{|a_{ii}|} < 1,$$

so the i-th Gershgorin disc $\mathcal{D}_{M,i}$ has a radius of $\alpha < 1$. Let $\lambda \in \sigma(\mathbf{M})$. If $|\lambda| \leq \alpha < 1$ holds, we are done. If $|\lambda| > \alpha$ holds, we have $\lambda \notin \partial\mathcal{D}_{X,i}$, and Lemma 2.28 implies that there exists at least one *open* Gershgorin disc $\mathcal{D}_{X,j}$ with $j \in \mathcal{I}$ and $\lambda \in \mathcal{D}_{X,j}$. Since this is an open disc around zero of radius at most one, we conclude $|\lambda| < 1$.

We conclude $\varrho(\mathbf{M}) < 1$, so the Neumann series converges to

$$\sum_{\ell=0}^{\infty} \mathbf{M}^\ell = (\mathbf{I} - \mathbf{M})^{-1} = (\mathbf{D}^{-1}\mathbf{A})^{-1} = \mathbf{A}^{-1}\mathbf{D}.$$

Multiplying by $\mathbf{D}^{-1}$ yields the final result. ■

2.6. Discrete maximum principle

Let us now return our attention to the investigation of finite difference discretization schemes. We denote the set of interior grid points by Ω_h, the set of boundary points by $\partial\Omega_h$, and the set of all grid points by $\bar{\Omega}_h$.

The discretization leads to a system

$$\mathbf{L}\mathbf{u} = \mathbf{f}, \qquad \mathbf{u}|_{\partial\Omega_h} = \mathbf{g} \tag{2.12}$$

of linear equations with the matrix $\mathbf{L} \in \mathbb{R}^{\Omega_h \times \bar{\Omega}_h}$, the right-hand side $\mathbf{f} \in \mathbb{R}^{\Omega_h}$, the boundary values $\mathbf{g} \in \mathbb{R}^{\partial\Omega_h}$, and the solution $\mathbf{u} \in \mathbb{R}^{\Omega_h}$.

We can separate the boundary values from the unknown values by introducing $\mathbf{A} := \mathbf{L}|_{\Omega_h \times \Omega_h}$ and $\mathbf{B} := \mathbf{L}|_{\Omega_h \times \partial\Omega_h}$. The system (2.12) takes the form

$$\mathbf{A}\mathbf{u}|_{\Omega_h} + \mathbf{B}\mathbf{u}|_{\partial\Omega_h} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{u}|_{\Omega_h} \\ \mathbf{u}|_{\partial\Omega_h} \end{pmatrix} = \mathbf{L}\mathbf{u} = \mathbf{f},$$

and due to $\mathbf{u}|_{\partial\Omega_h} = \mathbf{g}$, we obtain

$$\mathbf{A}\mathbf{u}|_{\Omega_h} = \mathbf{f} - \mathbf{B}\mathbf{g}. \tag{2.13}$$

In the model problem, we can apply the maximum principle introduced in Lemma 2.7 to vanishing boundary conditions $\mathbf{g} = \mathbf{0}$ and find that the coefficients of $\mathbf{u}$ are non-positive if the same holds for the coefficients of $\mathbf{A}\mathbf{u} \leq \mathbf{0}$.

Definition 2.32 (Positive matrices and vectors) *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{\mathcal{I} \times \mathcal{J}}$ *and* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$. *We define*

$$\begin{aligned}
\mathbf{x} > \mathbf{y} &\iff \forall i \in \mathcal{I} \ : \ x_i > y_i, \\
\mathbf{x} \geq \mathbf{y} &\iff \forall i \in \mathcal{I} \ : \ x_i \geq y_i, \\
\mathbf{A} > \mathbf{B} &\iff \forall i \in \mathcal{I}, j \in \mathcal{J} \ : \ a_{ij} > b_{ij}, \\
\mathbf{A} \geq \mathbf{B} &\iff \forall i \in \mathcal{I}, j \in \mathcal{J} \ : \ a_{ij} \geq b_{ij}.
\end{aligned}$$

Using these notations, Lemma 2.7 can be written as

$$\mathbf{Lu} \le \mathbf{0} \Rightarrow \mathbf{u} \le \mathbf{0} \qquad \text{for all } \mathbf{u} \in \mathbb{R}^{\mathcal{I}}.$$

In order to preserve this property in the general case, we would like to ensure $\mathbf{A}^{-1} \ge \mathbf{0}$. Due to Lemma 2.13, we have

$$\mathbf{A}^{-1} = \left(\sum_{\ell=0}^{\infty} \mathbf{M}^{\ell} \right) \mathbf{D}^{-1}, \qquad \mathbf{M} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A},$$

where $\mathbf{D}$ again denotes the diagonal part of $\mathbf{A}$. If we can ensure $\mathbf{M} \ge \mathbf{0}$ and $\mathbf{D} > \mathbf{0}$, this representation implies $\mathbf{A}^{-1} \ge \mathbf{0}$.

Due to

$$m_{ij} = \begin{cases} -a_{ij}/a_{ii} & \text{if } i \neq j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in \Omega_h, \tag{2.14}$$

we should ensure $a_{ij} \le 0$ for all $i, j \in \mathcal{I}$ with $i \neq j$ and $a_{ii} > 0$ for all $i \in \mathcal{I}$.

Definition 2.33 (Z-matrix) *A matrix* $\mathbf{A} \in \mathbb{R}^{\Omega_h \times \bar{\Omega}_h}$ *is called a* Z-matrix *if*

$$\begin{aligned} a_{ii} &> 0 && \textit{for all } i \in \Omega_h, \\ a_{ij} &\le 0 && \textit{for all } i \in \Omega_h,\ j \in \bar{\Omega}_h,\ i \neq j. \end{aligned}$$

If $\mathbf{A}$ is a Z-matrix, we have $\mathbf{M} \ge \mathbf{0}$. If the Neumann series for $\mathbf{M}$ converges, this implies $\mathbf{A}^{-1} \ge \mathbf{0}$. For an irreducibly diagonally dominant matrix $\mathbf{A}$, we can even obtain a stronger result.

Lemma 2.34 (Positive power) *Let* $\mathbf{A} \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ *be a matrix with* $\mathbf{A} \ge \mathbf{0}$.

The matrix $\mathbf{A}$ *is irreducible if and only if for every pair* $i, j \in \mathcal{I}$ *there is an* $m \in \mathbb{N}_0$ *with* $(\mathbf{A}^m)_{ij} > 0$.

Proof. We first prove

$$(\mathbf{A}^m)_{ij} > 0 \iff j \in N_m(i) \tag{2.15}$$

by induction for $m \in \mathbb{N}_0$.

Base case: Due to $\mathbf{A}^0 = \mathbf{I}$, we have $(\mathbf{A}^0)_{ij} \neq 0$ if and only if $i = j$.

Induction assumption: Let $m \in \mathbb{N}_0$ be chosen such that (2.15) holds for all $i, j \in \mathcal{I}$.

Induction step: Let $i, j \in \mathcal{I}$, and let $\mathbf{B} := \mathbf{A}^m$. We have

$$(\mathbf{A}^{m+1})_{ij} = (\mathbf{BA})_{ij} = \sum_{k \in \mathcal{I}} b_{ik} a_{kj} = \sum_{\substack{k \in \mathcal{I} \\ j \in N(k)}} b_{ik} a_{kj}$$

Assume first $(\mathbf{A}^{m+1})_{ij} > 0$. Then there has to be at least on $k \in \mathcal{I}$ with $b_{ik} > 0$ and $a_{kj} > 0$. By the induction assumption, the first inequality implies $k \in N_m(i)$. The second inequality implies $j \in N(k)$, and we conclude

$$j \in \bigcup_{k \in N_m(i)} N(k) = N_{m+1}(i).$$

Now assume $j \in N_{m+1}(i)$. By definition we find $k \in N_m(i)$ with $j \in N(k)$. By the induction assumption, we have $b_{ik} > 0$, and by the definition of neighbours we have $a_{kj} > 0$. Due to $\mathbf{A} \geq \mathbf{0}$ and $\mathbf{B} \geq \mathbf{0}$, we obtain

$$(\mathbf{A}^{m+1})_{ij} \geq a_{ik} b_{kj} > 0,$$

completing the induction.

Since $\mathbf{A}$ is irreducible if and only if for every pair $i, j \in \mathcal{I}$ there is an $m \in \mathbb{N}_0$ with $j \in N_m(i)$, our proof is complete. ∎

Theorem 2.35 (Positive inverse) *Let* $\mathbf{A}$ *be an irreducibly diagonally dominant Z-matrix. Then* $\mathbf{A}$ *is invertible with* $\mathbf{A}^{-1} > \mathbf{0}$.

Proof. Since $\mathbf{A}$ is a Z-matrix, all of its diagonal elements are strictly positive, so $\mathbf{M} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$ is well-defined. We have already seen that $\mathbf{M} \geq \mathbf{0}$ holds.

Since $\mathbf{A}$ is irreducibly diagonally dominant, the Neumann series for $\mathbf{M}$ fulfills

$$\mathbf{A}^{-1}\mathbf{D} = (\mathbf{I} - \mathbf{M})^{-1} = \sum_{\ell=0}^{\infty} \mathbf{M}^{\ell},$$

and due to $\mathbf{D} \geq \mathbf{0}$, this implies $\mathbf{A}^{-1} \geq \mathbf{0}$.

Since $\mathbf{A}$ is irreducible, so is $\mathbf{M}$. Let $i, j \in \mathcal{I}$. By Lemma 2.34, we find $m \in \mathbb{N}_0$ with $(\mathbf{M}^m)_{ij} > 0$ and conclude

$$\left(\sum_{\ell=0}^{\infty} \mathbf{M}^{\ell}\right)_{ij} \geq (\mathbf{M}^m)_{ij} > 0,$$

i.e., we have $(\mathbf{A}^{-1}\mathbf{D})_{ij} > \mathbf{0}$. Due to $\mathbf{D} \geq 0$, this implies $(\mathbf{A}^{-1})_{ij} > 0$. ∎

Remark 2.36 (M-matrix) *A Z-matrix* $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ *is called an* M-matrix *if* $\mathbf{A}^{-1} \geq \mathbf{0}$.

Theorem 2.35 states that an irreducibly diagonally dominant Z-matrix is an M-matrix.

Lemma 2.37 (Harmonic extension) *Let* $\mathbf{L} \in \mathbb{R}^{\Omega_h \times \bar{\Omega}_h}$ *be a Z-matrix such that* $\mathbf{A} := \mathbf{L}|_{\Omega_h \times \Omega_h}$ *is irreducibly diagonally dominant. There is a vector* $\mathbf{u}_0 \in \mathbb{R}^{\bar{\Omega}_h}$ *such that*

$$\mathbf{L}\mathbf{u}_0 = \mathbf{0}, \tag{2.16a}$$

$$\mathbf{u}_0|_{\partial\Omega_h} = \mathbf{1}_{\partial\Omega_h}, \tag{2.16b}$$

where $\mathbf{1}_{\partial\Omega_h}$ *denotes the vector in* $\mathbb{R}^{\partial\Omega_h}$ *with every component equal to one. This vector satisfies*

$$\mathbf{u} \leq \mathbf{1}_{\bar{\Omega}_h}.$$

Proof. Due to (2.13), (2.16) is equivalent to

$$\mathbf{A}\mathbf{u}_0|_{\Omega_h} = -\mathbf{B}\mathbf{1}|_{\partial\Omega_h}.$$

Due to Theorem 2.35, this equation has a unique solution.

Since $\mathbf{L}$ is a Z-matrix and weakly diagonally dominant, we have

$$(\mathbf{L1})_i = \ell_{ii} + \sum_{\substack{j\in\bar\Omega_h \\ j\neq i}} \ell_{ij} = |\ell_{ii}| - \sum_{\substack{j\in\bar\Omega_h \\ j\neq i}} |\ell_{ij}| \geq 0 \qquad \text{for all } i \in \mathcal{I}.$$

This implies

$$\mathbf{L}(\mathbf{1} - \mathbf{u}_0) \geq \mathbf{0},$$

and

$$\mathbf{1}|_{\partial\Omega_h} = \mathbf{u}_0|_{\partial\Omega_h},$$

yields

$$\mathbf{A}(\mathbf{1} - \mathbf{u}_0)|_{\Omega_h} = \mathbf{A}(\mathbf{1} - \mathbf{u}_0)|_{\Omega_h} + \mathbf{B}(\mathbf{1} - \mathbf{u}_0)|_{\partial\Omega_h} = \mathbf{L}(\mathbf{1} - \mathbf{u}_0) \geq \mathbf{0}.$$

Due to $\mathbf{A}^{-1} > \mathbf{0}$, we find

$$\mathbf{1} - \mathbf{u}_0 \geq \mathbf{0}.$$

■

Theorem 2.38 (Discrete maximum principle) *Let $\mathbf{L} \in \mathbb{R}^{\Omega_h\times\bar\Omega_h}$ be an irreducibly diagonally dominant Z-matrix.*

Let $\mathbf{u} \in \mathbb{R}^{\bar\Omega_h}$ satisfy

$$\mathbf{Lu} \leq \mathbf{0}.$$

Then there is a boundary index $j \in \partial\Omega_h$ such that

$$u_i \leq u_j \qquad \textit{for all } i \in \bar\Omega_h.$$

Proof. We denote the maximum of $\mathbf{u}$ on the boundary by

$$\beta := \max\{u_i \ : \ i \in \partial\Omega_h\}.$$

Let $\mathbf{u}_0 \in \mathbb{R}^{\bar\Omega_h}$ be the function introduced in Lemma 2.37 and define

$$\widehat{\mathbf{u}} := \mathbf{u} - \beta\mathbf{u}_0.$$

Due to (2.16), we have

$$\mathbf{L}\widehat{\mathbf{u}} = \mathbf{L}(\mathbf{u} - \beta\mathbf{u}_0) = \mathbf{Lu} \leq \mathbf{0}$$

and

$$\widehat{u}_i = u_i - \beta \leq 0 \qquad \text{for all } j \in \partial\Omega_h.$$

With $\mathbf{A} = \mathbf{L}|_{\Omega_h\times\Omega_h}$ and $\mathbf{B} = \mathbf{L}|_{\Omega_h\times\partial\Omega_h}$, we find

$$\mathbf{0} \geq \mathbf{L}\widehat{\mathbf{u}} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{u}}|_{\Omega_h} \\ \widehat{\mathbf{u}}|_{\partial\Omega_h} \end{pmatrix} = \mathbf{A}\widehat{\mathbf{u}}|_{\Omega_h} + \mathbf{B}\widehat{\mathbf{u}}|_{\partial\Omega_h}.$$

Since $\mathbf{L}$ is a Z-matrix, we have $\mathbf{B} \leq \mathbf{0}$ and therefore $\mathbf{B}\widehat{\mathbf{u}} \geq \mathbf{0}$.

Due to $\mathbf{A}^{-1} \geq \mathbf{0}$, we find

$$\widehat{\mathbf{u}}|_{\Omega_h} \leq \widehat{\mathbf{u}}|_{\Omega_h} + \mathbf{A}^{-1}\mathbf{B}\widehat{\mathbf{u}} = \mathbf{A}^{-1}\mathbf{L}\widehat{\mathbf{u}} \leq \mathbf{0},$$

and conclude

$$\mathbf{u} = \widehat{\mathbf{u}} + \beta\mathbf{u}_0 \leq \beta\mathbf{u}_0.$$

Due to Lemma 2.37, each component of $\mathbf{u}_0$ is bounded by one, and therefore each component of $\mathbf{u}$ is bounded by β. ■

2.7. Stability, consistency, and convergence

Let us consider a partial differential equation

$$\mathcal{L}u(x) = f(x) \qquad \text{for all } x \in \Omega,$$

on a domain Ω. We prescibe *Dirichlet boundary conditions*, i.e., we require

$$u(x) = g(x) \qquad \text{for all } x \in \partial\Omega.$$

Here $f : \Omega \to \mathbb{R}$ and $g : \partial\Omega \to \mathbb{R}$ are suitable functions, $\mathcal{L}$ is a linear differential operator, and $u : \Omega \to \mathbb{R}$ is the solution.

We approximate Ω by a grid Ω_h and the boundary $\partial\Omega$ by $\partial\Omega_h$, and let $\bar{\Omega}_h := \Omega_h \cup \partial\Omega_h$.

As before, we define the spaces

$$\begin{aligned} G(\Omega_h) &:= \{u : \Omega_h \to \mathbb{R}\}, \\ G(\bar{\Omega}_h) &:= \{u : \bar{\Omega}_h \to \mathbb{R}\}, \\ G_0(\bar{\Omega}_h) &:= \{u \in G(\bar{\Omega}_h) \ : \ u|_{\partial\Omega_h} = 0\} \end{aligned}$$

and consider a linear finite difference operator

$$\mathcal{L}_h : G(\bar{\Omega}_h) \to G(\Omega_h)$$

that approximates $\mathcal{L}$. The finite difference approximation $u_h \in G(\bar{\Omega}_h)$ of the solution u is then given by

$$\mathcal{L}_h u_h(x) = f(x) \qquad \text{for all } x \in \Omega_h, \tag{2.17a}$$

$$u_h(x) = g(x) \qquad \text{for all } x \in \partial\Omega_h. \tag{2.17b}$$

The functions u_h, $f|_{\Omega_h}$ and $g|_{\partial\Omega_h}$ can be interpreted as vectors $\mathbf{u} \in \mathbb{R}^{\bar{\Omega}_h}$, $\mathbf{f} \in \mathbb{R}^{\Omega_h}$ and $\mathbf{g} \in \mathbb{R}^{\partial\Omega_h}$, and the linear operator $\mathcal{L}_h$ corresponds to a matrix $\mathbf{L} \in \mathbb{R}^{\Omega_h \times \bar{\Omega}_h}$. We find

$$\mathbf{L}\mathbf{u} = \mathbf{f}, \qquad \mathbf{u}|_{\partial\Omega_h} = \mathbf{g} \tag{2.18}$$

and now have to prove that this system has a unique solution $\mathbf{u}$ that converges to the solution u of the original differential equation.

Using Theorem 2.38, we can apply the same arguments as in the proof of Lemma 2.9 to establish stability of general finite difference schemes.

Corollary 2.39 (Stability) *Let $\mathbf{L} \in \mathbb{R}^{\Omega_h \times \bar{\Omega}_h}$ be an irreducibly diagonally dominant Z-matrix, and let $\mathcal{L}_h : G(\bar{\Omega}_h) \to G(\Omega_h)$ denote the corresponding finite difference operator.*

Let $w_h \in G(\bar{\Omega}_h)$ be a grid function satisfying

$$\mathcal{L}_h w_h(x) \geq 1 \qquad \text{for all } x \in \Omega_h,$$

and let

$$\gamma := \max\{|w_h(x) - w_h(y)| \ : \ x, y \in \bar{\Omega}_h\}.$$

Then we have

$$\|u_h\|_{\infty,\Omega_h} \leq \gamma \|\mathcal{L}_h u_h\|_{\infty,\Omega_h} \qquad \text{for all } u_h \in G_0(\bar{\Omega}_h). \tag{2.19}$$

Proof. Let $u_h \in G_0(\bar{\Omega}_h)$. We fix

$$\beta := \max\{|\mathcal{L}_h u_h(x)| \ : \ x \in \Omega_h\} = \|\mathcal{L}_h u_h\|_{\infty,\Omega_h}.$$

Then we have

$$\begin{aligned} \mathcal{L}_h(u_h - \beta w_h)(x) &= \mathcal{L}_h u_h(x) - \beta \mathcal{L}_h w_h(x) \leq \mathcal{L}_h u_h(x) - \beta \leq 0 && \text{for all } x \in \Omega_h, \\ \mathcal{L}_h(-u_h - \beta w_h)(x) &= -\mathcal{L}_h u_h(x) - \beta \mathcal{L}_h w_h(x) \leq -\mathcal{L}_h u_h(x) - \beta \leq 0 && \text{for all } x \in \Omega_h. \end{aligned}$$

Theorem 2.38 yields boundary indices $y, z \in \partial\Omega_h$ such that

$$\begin{aligned} u_h(x) - \beta w_h(x) &\leq u_h(y) - \beta w_h(y), \\ -u_h(x) - \beta w_h(x) &\leq -u_h(z) - \beta w_h(z) && \text{for all } x \in \Omega_h. \end{aligned}$$

Due to $u_h|_{\partial\Omega_h} = 0$, we find

$$\begin{aligned} u_h(x) &\leq \beta(w_h(x) - w_h(y)) \leq \beta\gamma, \\ -u_h(x) &\leq \beta(w_h(x) - w_h(z)) \leq \beta\gamma && \text{for all } x \in \Omega_h, \end{aligned}$$

and this implies $\|u_h\|_{\infty,\Omega_h} \leq \beta\gamma = \gamma\|\mathcal{L}_h u_h\|_{\infty,\Omega_h}$. ∎

Corollary 2.40 (Error estimate) *Let $\gamma \in \mathbb{R}_{\geq 0}$ be a constant such that (2.19) holds.*

Let $\mathcal{L}_h$ be consistent of order p *with $\mathcal{L}$ and the solution u, i.e., let*

$$\|\mathcal{L}u - \mathcal{L}_h u\|_{\infty,\Omega_h} \leq C_{cn} h^p$$

hold for constants $C_{cn} \in \mathbb{R}_{\geq 0}$, $p \in \mathbb{N}$, and the mesh width $h \in \mathbb{R}_{>0}$.

Then we have

$$\|u - u_h\|_{\infty,\Omega_h} \leq C_{cn} \gamma h^p.$$

Proof. By definition, we have

$$\mathcal{L}u(x) = f(x) = \mathcal{L}_h u_h(x) \qquad \text{for all } x \in \Omega_h.$$

Using (2.19) yields

$$\|u_h - u\|_{\infty,\Omega_h} \leq \gamma\|\mathcal{L}_h u_h - \mathcal{L}_h u\|_{\infty,\Omega_h} = \gamma\|\mathcal{L}u - \mathcal{L}_h u\|_{\infty,\Omega_h} \leq C_{\mathrm{cn}} \gamma h^p.$$

∎

2.8. Analysis in Hilbert spaces

Until now, we have worked with the maximum norm to establish consistency, stability, and convergence. In the following chapters, it is advantageous to formulate our statements in terms of norms corresponding to *Hilbert spaces.*

Definition 2.41 (Inner product) *Let $\mathcal{V}$ be an $\mathbb{R}$-vector space. A mapping $a\colon \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ is called a* bilinear form *if*

$$a(v+\alpha w, u) = a(v,u) + \alpha a(w,u) \qquad \text{for all } u,v,w \in \mathcal{V},\ \alpha \in \mathbb{R}, \tag{2.20a}$$

$$a(v, u+\alpha w) = a(v,u) + \alpha a(v,w) \qquad \text{for all } u,v,w \in \mathcal{V},\ \alpha \in \mathbb{R}. \tag{2.20b}$$

A bilinear form a is called positive definite *if*

$$a(u,u) > 0 \qquad \text{for all } u \in \mathcal{V}\setminus\{0\}, \tag{2.20c}$$

and it is called symmetric *if*

$$a(u,v) = a(v,u) \qquad \text{for all } u,v \in \mathcal{V}. \tag{2.20d}$$

A symmetric positive definite bilinear form is called an inner product *for the space $\mathcal{V}$.*

Lemma 2.42 (Cauchy-Schwarz) *Let $a\colon \mathcal{V}\times\mathcal{V}\to\mathbb{R}$ be an inner product. We have*

$$|a(v,u)|^2 \leq a(v,v)a(u,u) \qquad \text{for all } u,v \in \mathcal{V}.$$

Both sides are equal if and only if u and v are linearly dependent.

Proof. Let $u,v\in\mathcal{V}$. If $a(v,v)=0$, we let $\alpha\in\mathbb{R}$ and use (2.20a) and (2.20b) to find

$$\begin{aligned} 0 &\leq a(\tfrac{1}{\alpha}v - \alpha u, \tfrac{1}{\alpha}v - \alpha u) \\ &= \tfrac{1}{\alpha}a(v, \tfrac{1}{\alpha}v - \alpha u) - \alpha a(u, \tfrac{1}{\alpha}v - \alpha u) \\ &= \tfrac{1}{\alpha^2}a(v,v) - a(v,u) - a(u,v) + \alpha^2 a(u,u). \end{aligned}$$

With (2.20d), we conclude

$$2a(v,u) \leq \tfrac{1}{\alpha^2}a(v,v) + \alpha^2 a(u,u) = \alpha^2 a(u,u).$$

Since this inequality holds for arbitrary values of α, we have $a(v,u)\leq 0$. Replacing v by $-v$, we obtain $-a(v,u) = a(-v,u) \leq 0$, and therefore $|a(v,u)| = 0$.

Let now $a(v,v)\neq 0$. For all $\alpha\in\mathbb{R}$ we can apply (2.20a), (2.20b), and (2.20d) to obtain

$$\begin{aligned} 0 &\leq a(u-\alpha v, u-\alpha v) \\ &= a(u, u-\alpha v) - \alpha a(v, u-\alpha v) \\ &= a(u,u) - \alpha a(u,v) - \alpha a(v,u) + \alpha^2 a(v,v) \end{aligned}$$

$$= a(u,u) - 2\alpha a(v,u) + \alpha^2 a(v,v).$$

Due to $a(v,v) > 0$, we can minimize the last term by choosing

$$\alpha := \frac{a(v,u)}{a(v,v)}$$

and obtain

$$0 \leq a(u,u) - 2\frac{a(v,u)^2}{a(v,v)} + \frac{a(v,u)^2}{a(v,v)^2} a(v,v) = a(u,u) - \frac{a(v,u)^2}{a(v,v)}.$$

Multiplying by $a(v,v)$ yields

$$0 \leq a(v,v)a(u,u) - a(v,u)^2,$$

and this is the Cauchy-Schwarz inequality. If $a(v,u)^2 = a(v,v)a(u,u)$ holds, we have

$$0 \leq a(u - \alpha v, u - \alpha v) = a(v,v)a(u,u) - a(v,u)^2 = 0,$$

i.e., $a(u - \alpha v, u - \alpha v) = 0$. Due to (2.20c), this implies $u - \alpha v = 0$, so u and v are linear dependent. ■

Remark 2.43 (Positive semidefinite) *A bilinear form $a \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is called* positive semidefinite*, if*

$$a(u,u) \geq 0 \qquad \text{for all } u \in \mathcal{V}.$$

The proof of Lemma 2.42 remains valid except for the final statement if a is only symmetric and positive semidefinite.

Corollary 2.44 (Hilbert norm) *Let $a \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be an inner product.*

$$\|u\|_a := \sqrt{a(u,u)} \qquad \text{for all } u \in \mathcal{V}$$

is a norm for the space $\mathcal{V}$. We call it the Hilbert norm *corresponding to the inner product. Using this norm, the Cauchy-Schwarz inequality takes the short form*

$$|a(v,u)| \leq \|v\|_a \|u\|_a \qquad \text{for all } u, v \in \mathcal{V}. \tag{2.21}$$

Proof. Let $u \in \mathcal{V}$. Due to (2.20a), we have

$$a(0,u) = a(u-u,u) = a(u,u) - a(u,u) = 0$$

and therefore $\|0\|_a = \sqrt{a(0,0)} = 0$. If $\|u\|_a = 0$ holds, we have

$$0 = \|u\|_a^2 = a(u,u),$$

and (2.20c) yields $u = 0$.

For $\alpha \in \mathbb{R}$, (2.20a) and (2.20b) yield

$$\|\alpha u\|_a = \sqrt{a(\alpha u, \alpha u)} = \sqrt{\alpha^2 a(u,u)} = |\alpha| \, \|u\|_a.$$

Let $v \in \mathcal{V}$. (2.20a), (2.20b), (2.20d), and Lemma 2.42 yield

$$\begin{aligned} \|u+v\|_a^2 &= a(u+v, u+v) = a(u,u) + 2a(v,u) + a(v,v) \\ &\leq a(u,u) + 2\sqrt{a(u,u)a(v,v)} + a(v,v) = (\sqrt{a(u,u)} + \sqrt{a(v,v)})^2 \\ &= (\|u\|_a + \|v\|_a)^2, \end{aligned}$$

so we also have established the triangle inequality. ∎

Definition 2.45 (Banach space) *Let $\mathcal{V}$ be an $\mathbb{R}$-vector space with a norm $\|\cdot\|_V$. A sequence $(u_n)_{n=0}^\infty$ in $\mathcal{V}$ is called a* Cauchy sequence*, if for every $\epsilon \in \mathbb{R}_{>0}$ there is an $n_0 \in \mathbb{N}_0$ such that*

$$\|u_n - u_m\|_V \leq \epsilon \qquad \text{for all } n, m \in \mathbb{N}_0 \text{ with } n, m \geq n_0.$$

If every Cauchy sequence converges, i.e., if for every Cauchy sequence $(u_n)_{n=0}^\infty$ there is a $u \in \mathcal{V}$ with

$$\lim_{n \to \infty} \|u - u_n\|_V = 0,$$

the space $\mathcal{V}$ is called a Banach space.

Definition 2.46 (Hilbert space) *Let $\mathcal{V}$ be an $\mathbb{R}$-vector space, and let $a \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be an inner product for $\mathcal{V}$. If $\mathcal{V}$ is a Banach space with respect to the Hilbert norm $\|\cdot\|_a$, we call it a* Hilbert space. *In this case, we denote the norm by $\|u\|_V := \|u\|_a$ and the inner product by $\langle v, u\rangle_V := a(v,u)$ for all $u, v \in \mathcal{V}$.*

In order to take advantage of the properties of Hilbert spaces, we have to equip the grid functions introduced in Definition 2.6 with a suitable inner product. We imitate the L^2-inner product on the space of square-integrable functions:

Definition 2.47 (Hilbert space of grid functions) *Let $N \in \mathbb{N}$, and let h and Ω_h be as in Definition 2.5. Let $G_0(\bar{\Omega}_h)$ be the space of grid functions with homogeneous Dirichlet boundary conditions. With the inner product*

$$\langle v, u\rangle_{\Omega_h} := h^2 \sum_{x \in \Omega_h} v(x)u(x) \qquad \text{for all } v, u \in G_0(\bar{\Omega}_h),$$

$G_0(\bar{\Omega}_h)$ is a Hilbert space with the Hilbert norm given by

$$\|u\|_{\Omega_h} := h \left(\sum_{x \in \Omega_h} u(x)^2 \right)^{1/2} \qquad \text{for all } u \in G_0(\bar{\Omega}_h).$$

In order to obtain convergence with respect to the Hilbert norm $\|\cdot\|_{\Omega_h}$, we need consistency and stability estimates. For the consistency, we can simply rely on the estimate provided by Lemma 2.4.

Lemma 2.48 (Consistency) *Let $u \in C^4(\bar{\Omega})$. We have*

$$\|\Delta u - \Delta_h u\|_{\Omega_h} \leq \frac{h^2}{6}|u|_{4,\Omega}.$$

Proof. We use (2.4) to find

$$\begin{aligned}\|\Delta u - \Delta_h u\|_{\Omega_h}^2 &= h^2 \sum_{x\in\Omega_h} (\Delta u(x) - \Delta_h u(x))^2 \leq h^2 \sum_{x\in\Omega_h} \frac{h^4}{36}|u|_{4,\Omega}^2 \\ &= h^2 N^2 \frac{h^4}{36}|u|_{4,\Omega}^2 = \frac{N^2}{(N+1)^2}\frac{h^4}{36}|u|_{4,\Omega}^2 \leq \frac{h^4}{36}|u|_{4,\Omega}^2.\end{aligned}$$

■

Replacing the stability estimate is a little more challenging, since it involves the maximum norm on the right-hand side, where we would like to see the Hilbert norm instead. Instead of working with the discrete maximum principle (cf. Theorem 2.38), we can rely on the Cauchy-Schwarz inequality for the Euclidean inner product.

Lemma 2.49 (Stability) *Let $u_h \in G_0(\bar{\Omega}_h)$. We have*

$$\|u_h\|_{\Omega_h}^2 \leq \frac{1}{2}\langle u_h, -\Delta_h u_h\rangle_{\Omega_h},$$

and this implies

$$\|u_h\|_{\Omega_h} \leq \frac{1}{2}\|\Delta_h u_h\|_{\Omega_h}.$$

Proof. The Cauchy-Schwarz inequality (2.21) applied to the Euclidean inner product reads

$$\left(\sum_{k=1}^n x_k y_k\right)^2 \leq \left(\sum_{k=1}^n x_k^2\right)\left(\sum_{k=1}^n y_k^2\right) \qquad \text{for all } x, y \in \mathbb{R}^n,\ n \in \mathbb{N}. \tag{2.22}$$

Let $x \in \Omega_h$. By definition, we find $i, j \in [1:N]$ with $x = (ih, jh)$. Due to $u_h(0, jh) = 0$, we can use a telescoping sum to obtain

$$u_h(x) = u_h(ih, jh) - u_h(0, jh) = \sum_{k=1}^{i} u_h(kh, jh) - u_h((k-1)h, jh),$$

and the Cauchy-Schwarz inequality (2.22) yields

$$u_h(x)^2 = \left(\sum_{k=1}^{i} u_h(kh, jh) - u_h((k-1)h, jh)\right)^2$$

$$\le \left(\sum_{k=1}^{i} 1\right) \left(\sum_{k=1}^{i} (u_h(kh, jh) - u_h((k-1)h, jh))^2\right)$$
$$= i \sum_{k=1}^{i} (u_h(kh, jh) - u_h((k-1)h, jh))^2.$$

This sum only involves differences between neighbouring grid points, just like the discrete Laplacian. We denote the set of neighbours again by

$$\begin{aligned} N(x) := \{y \in \bar{\Omega}_h \ &: (x_1 = y_1 \wedge |x_2 - y_2| = h) \\ &\vee (x_2 = y_2 \wedge |x_1 - y_1| = h)\} \qquad \text{for all } x \in \bar{\Omega}_h \end{aligned}$$

and write the discrete Laplacian as

$$-\Delta_h u_h(x) = h^{-2} \sum_{y \in N(x)} u_h(x) - u_h(y) \qquad \text{for all } x \in \Omega_h.$$

Taking advantage of the homogeneous Dirichlet conditions, the symmetry $y \in N(x) \iff x \in N(y)$, and employing a change of variables, we find

$$\begin{aligned}
\langle u_h, -\Delta_h u_h\rangle_{\Omega_h} &= \sum_{x \in \Omega_h} u_h(x) \sum_{y \in N(x)} u_h(x) - u_h(y) \\
&= \sum_{x \in \bar{\Omega}_h} u_h(x) \sum_{y \in N(x)} u_h(x) - u_h(y) \\
&= \sum_{\substack{x,y \in \bar{\Omega}_h \\ y \in N(x)}} u_h(x)(u_h(x) - u_h(y)) \\
&= \frac{1}{2} \sum_{\substack{x,y \in \bar{\Omega}_h \\ y \in N(x)}} u_h(x)(u_h(x) - u_h(y)) + \frac{1}{2} \sum_{\substack{x,y \in \bar{\Omega}_h \\ x \in N(y)}} u_h(x)(u_h(x) - u_h(y)) \\
&= \frac{1}{2} \sum_{\substack{x,y \in \bar{\Omega}_h \\ y \in N(x)}} u_h(x)(u_h(x) - u_h(y)) - \frac{1}{2} \sum_{\substack{x,y \in \bar{\Omega}_h \\ x \in N(y)}} u_h(x)(u_h(y) - u_h(x)) \\
&= \frac{1}{2} \sum_{\substack{x,y \in \bar{\Omega}_h \\ y \in N(x)}} u_h(x)(u_h(x) - u_h(y)) - \frac{1}{2} \sum_{\substack{x,y \in \bar{\Omega}_h \\ y \in N(x)}} u_h(y)(u_h(x) - u_h(y)) \\
&= \frac{1}{2} \sum_{\substack{x,y \in \bar{\Omega}_h \\ y \in N(x)}} (u_h(x) - u_h(y))^2 \\
&\ge \frac{1}{2} \sum_{x \in \Omega_h} (u_h(x) - u_h(x_1 - h, x_2))^2 + \frac{1}{2} \sum_{x \in \Omega_h} (u_h(x_1 - h, x_2) - u_h(x))^2 \\
&= \sum_{x \in \Omega_h} (u_h(x) - u_h(x_1 - h, x_2))^2.
\end{aligned}$$

We can complete the proof by combining both estimates and using Gauss' summation formula:

$$\begin{aligned}
\sum_{x\in\Omega_h} u_h(x)^2 &\leq \sum_{i=1}^N \sum_{j=1}^N u_h(ih,jh)^2 \leq \sum_{i=1}^N \sum_{j=1}^N i \sum_{k=1}^i (u_h(kh,jh) - u_h((k-1)h,jh))^2 \\
&\leq \sum_{i=1}^N \sum_{j=1}^N i \sum_{k=1}^N (u_h(kh,jh) - u_h((k-1)h,jh))^2 \\
&= \frac{N(N+1)}{2} \sum_{k=1}^N \sum_{j=1}^N (u_h(kh,jh) - u_h((k-1)h,jh))^2 \\
&\leq \frac{N(N+1)}{2} \langle u_h, -\Delta_h u_h\rangle_{\Omega_h} \leq \frac{(N+1)^2}{2} \langle u_h, -\Delta_h u_h\rangle_{\Omega_h} \\
&= \frac{1}{2h^2} \langle u_h, -\Delta_h u_h\rangle_{\Omega_h},
\end{aligned}$$

and multiplying by h^2 yields the first estimate.

From this, we immediately obtain

$$\|u_h\|_{\Omega_h}^2 \leq \frac{1}{2}\langle u_h, -\Delta_h u_h\rangle_{\Omega_h} \leq \frac{1}{2}\|u_h\|_{\Omega_h} \|\Delta_h u_h\|_{\Omega_h}$$

using the Cauchy-Schwarz inequality (2.21), and dividing by $\|u_h\|_{\Omega_h}$ yields the second estimate. ∎

We can proceed as in the proof of Theorem 2.10 to find

$$\|u_h - u|_{\Omega_h}\|_{\Omega_h} \leq \frac{h^2}{12}|u|_{4,\Omega},$$

i.e., the grid functions will converge to the solution with respect to the Hilbert norm at the same rate as with respect to the maximum norm.

The larger constant in the estimate is due to the larger constant in the stability estimate. Using more sophisticated techniques, it is actually possible to prove

$$16\|u_h\|_{\Omega_h}^2 \leq \langle u_h, -\Delta_h u_h\rangle_{\Omega_h},$$

and the constant grows to $2\pi^2$ as the mesh is refined, the smallest eigenvalue of the Laplace operator on the unit square Ω.

3. Finite difference methods for parabolic equations

Partial differential equations like Poisson's equation are typically used to describe systems that do not change over time, e.g., the electrostatic field corresponding to a fixed charge distribution or equilibrium states of mechanical systems.

Now we focus on time-dependent partial differential equations, starting with *parabolic* equations that can be approached similarly to ordinary differential equations.

3.1. Heat equation

A classical example for a parabolic equation is the *heat equation.* For the two-dimensional unit square $\Omega = (0,1)^2$, it takes the form

$$\frac{\partial u}{\partial t}(t,x) = g(t,x) + \Delta_x u(t,x) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \Omega, \tag{3.1a}$$

where Δ_x is the Laplace operator applied only to the x variable. As in the previous chapter, we have to add boundary conditions to ensure the uniqueness of the solution. We once again choose homogeneous Dirichlet conditions

$$u(t,x) = 0 \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \partial\Omega. \tag{3.1b}$$

We also have to provide initial conditions

$$u(a,x) = u_0(x) \qquad \text{for all } x \in \Omega. \tag{3.1c}$$

The value $u(t,x)$ can be interpreted as the temperature at time $t \in \mathbb{R}_{\geq 0}$ in the point $x \in \Omega$. The function g describes where and when heat is created: a positive value $g(t,x)$ means that at time $t \in \mathbb{R}_{\geq 0}$ the point $x \in \Omega$ is being heated, while a negative value means that it is being cooled.

If g is constant with respect to time, i.e., if there is a function $g_\infty : \Omega \to \mathbb{R}$ such that

$$g(t,x) = g_\infty(x) \qquad \text{for all } t \in \mathbb{R},$$

it is possible to prove that the solution u will converge to a function $u_\infty \in C^2(\Omega)$ that solves the Poisson equation

$$-\Delta u_\infty(x) = g_\infty(x) \qquad \text{for all } x \in \Omega,$$

$$u_\infty(x) = 0 \qquad \text{for all } x \in \partial\Omega.$$

This limit is called the *equilibrium solution*, and we can approximate it by the techniques we have already discussed.

In order to handle the time dependence of the solution, we interprete u and g as functions in time mapping to functions in space, i.e., we let

$$\widehat{u}(t)(x) := u(t,x), \qquad \widehat{g}(t)(x) := g(t,x) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \Omega.$$

By introducing the space

$$C^\infty_{\partial\Omega}(\bar{\Omega}) := \{u \in C(\bar{\Omega}) \ :\ u|_\Omega \in C^\infty(\Omega),\ u|_{\partial\Omega} = 0\}$$

and extending the Laplace operator to

$$\Delta v(x) := \begin{cases} \frac{\partial^2 v}{\partial x_1^2}(x) + \frac{\partial^2 v}{\partial x_2^2}(x) & \text{if } x \in \Omega, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } v \in C^\infty_{\partial\Omega}(\bar{\Omega}),\ x \in \bar{\Omega},$$

we can write (3.1) as the ordinary differential equation

$$\widehat{u}(0) = \widehat{u}_0, \qquad \widehat{u}'(t) = \widehat{g}(t) + \Delta\widehat{u}(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}, \tag{3.2}$$

with the initial value $\widehat{u}_0 \in C^\infty_{\partial\Omega}(\bar{\Omega})$, the heating function $\widehat{g} \in C(\mathbb{R}_{\geq 0}, C^\infty_{\partial\Omega}(\bar{\Omega}))$, and the solution $\widehat{u} \in C^1(\mathbb{R}_{\geq 0}, C^\infty_{\partial\Omega}(\bar{\Omega}))$.

3.2. Method of lines

The idea of the *method of lines* is to replace the spatial differential operator by an approximation. We choose the finite difference discretization we have already employed for the Poisson equation: we let $N \in \mathbb{N}$, let $h := 1/(N+1)$, replace the domain Ω by the grid Ω_h and the space $C^\infty_{\partial\Omega}(\bar{\Omega})$ by $G_0(\bar{\Omega}_h)$, and approximate the differential operator Δ by

$$\Delta_h : G_0(\bar{\Omega}_h) \to G_0(\bar{\Omega}_h)$$

defined as

$$\Delta_h v(x) = \frac{v(x_1+h,x_2) + v(x_1-h,x_2) + v(x_1,x_2+h) + v(x_1,x_2-h) - 4v(x)}{h^2}$$

for all $x \in \Omega_h$ and extended by

$$\Delta_h v(x) = 0$$

for all boundary points $x \in \partial\Omega_h$. Replacing $\widehat{u}$, $\widehat{g}$ and $\widehat{u}_0$ by

$$u_h(t) := \widehat{u}(t)|_{\bar{\Omega}_h}, \qquad g_h(t) := \widehat{g}(t)|_{\bar{\Omega}_h}, \qquad u_{0,h} := \widehat{u}_0|_{\bar{\Omega}_h} \qquad \text{for all } t \in \mathbb{R}_{\geq 0},$$

we obtain the approximation

$$u_h(0) = u_{0,h}, \qquad u_h'(t) = g_h(t) + \Delta_h u_h(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}, \tag{3.3}$$

and this is an ordinary differential equation in the finite-dimensional space $G_0(\bar{\Omega}_h)$.

If we introduce the function

$$f\colon \mathbb{R}_{\geq 0} \times G_0(\bar{\Omega}_h) \to G_0(\bar{\Omega}_h), \qquad (t, y_h) \mapsto g_h(t) + \Delta_h y_h, \tag{3.4}$$

we can write (3.3) in the standard form

$$u_h(0) = u_{0,h}, \qquad u_h'(t) = f(t, u_h(t)) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}. \tag{3.5}$$

Lemma 3.1 (Unique solution) *The ordinary differential equation (3.3) has a unique solution* $u_h \in C^1(\mathbb{R}_{\geq 0}, G_0(\bar{\Omega}_h))$.

Proof. Since $G_0(\bar{\Omega}_h)$ is finite-dimensional, the mapping Δ_h is continuous, i.e., there is a constant C_Δ such that

$$\|\Delta_h y_h\|_\infty \leq C_\Delta \|y_h\|_\infty \qquad \text{for all } y_h \in G_0(\bar{\Omega}_h).$$

We find

$$\begin{aligned} \|f(t, y_h) - f(t, z_h)\|_\infty &= \|\Delta_h y_h - \Delta_h z_h\|_\infty = \|\Delta_h (y_h - z_h)\|_\infty \\ &\leq C_\Delta \|y_h - z_h\|_\infty \qquad \text{for all } y_h, z_h \in G_0(\bar{\Omega}_h), \end{aligned}$$

so the function f is Lipschitz continuous in the second parameter.

We can apply the Picard-Lindeløf theorem to conclude that (3.5) has a unique solution, and this is equivalent to (3.3) having a unique solution. ■

In this proof, the existence of C_Δ is the consequence of the fact that Δ_h is a linear mapping between finite-dimensional spaces. This is a fairly general approach and does not provide us with any information regarding the behaviour of the Lipschitz constant.

In the case of our model problem, we can fortunately compute all eigenvalues and eigenvectors of Δ_h, and this gives us better insight into the behaviour of the system.

Lemma 3.2 (Eigenvalues) *We define*

$$\begin{aligned} \lambda_{h,\nu} &:= 4h^{-2}(\sin^2(\pi\nu_1 h/2) + \sin^2(\pi\nu_2 h/2) && \textit{for all } \nu \in [1:N]^2, \\ e_{h,\nu}(x) &:= 2\sin(\pi\nu_1 x_1)\sin(\pi\nu_2 x_2) && \textit{for all } \nu \in [1:N]^2,\ x \in \bar{\Omega}_h. \end{aligned}$$

Then we have

$$\begin{aligned} -\Delta_h e_{h,\nu} &= \lambda_{h,\nu} e_{h,\nu} && \textit{for all } \nu \in [1:N]^2, \\ \langle e_{h,\nu}, e_{h,\mu}\rangle_{\Omega_h} &= \begin{cases} 1 & \textit{if } \nu = \mu, \\ 0 & \textit{otherwise} \end{cases} && \textit{for all } \nu, \mu \in [1:N]^2. \end{aligned}$$

Proof. Let $\nu \in [1:N]^2$, and let $x \in \Omega_h$. We have

$$-\Delta_h e_{h,\nu}(x) = \frac{2e_{h,\nu}(x) - e_{h,\nu}(x_1 + h, x_2) - e_{h,\nu}(x_1 - h, x_2)}{h^2}$$

$$+\frac{2e_{h,\nu}(x) - e_{h,\nu}(x_1, x_2 + h) - e_{h,\nu}(x_1, x_2 - h)}{h^2}.$$

Using the trigonometric identity $\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$, we obtain

$$\begin{aligned}
2e_{h,\nu}(x) - e_{h,\nu}(x_1 + h, x_2) - e_{h,\nu}(x_1 - h, x_2)& \\
= 4\sin(\pi\nu_1 x_1)\sin(\pi\nu_2 x_2)& \\
- 2\sin(\pi\nu_1 x_1 + \pi\nu_1 h)\sin(\pi\nu_2 x_2)& \\
- 2\sin(\pi\nu_1 x_1 - \pi\nu_1 h)\sin(\pi\nu_2 x_2)& \\
= 2\big(2\sin(\pi\nu_1 x_1)& \\
- \sin(\pi\nu_1 x_1)\cos(\pi\nu_1 h) - \cos(\pi\nu_1 x_1)\sin(\pi\nu_1 h)& \\
- \sin(\pi\nu_1 x_1)\cos(-\pi\nu_1 h) - \cos(\pi\nu_1 x_1)\sin(-\pi\nu_1 h)\big)\sin(\pi\nu_2 x_2)& \\
= 2\big(2\sin(\pi\nu_1 x_1)& \\
- \sin(\pi\nu_1 x_1)\cos(\pi\nu_1 h) - \cos(\pi\nu_1 x_1)\sin(\pi\nu_1 h)& \\
- \sin(\pi\nu_1 x_1)\cos(\pi\nu_1 h) + \cos(\pi\nu_1 x_1)\sin(\pi\nu_1 h)\big)\sin(\pi\nu_2 x_2)& \\
= 4\big(1 - \cos(\pi\nu_1 h)\big)\sin(\pi\nu_1 x_1)\sin(\pi\nu_2 x_2)& \\
= 2\big(1 - \cos(\pi\nu_1 h)\big)e_{h,\nu}(x).&
\end{aligned}$$

Using the trigonometric identity $\cos(\alpha) = 1 - 2\sin^2(\alpha/2)$, we obtain

$$\begin{aligned}
2e_{h,\nu}(x) - e_{h,\nu}(x_1 + h, x_2) - e_{h,\nu}(x_1 - h, x_2)& \\
= 2\big(1 - 1 + 2\sin^2(\pi\nu_1 h/2)\big)e_{h,\nu}(x) = 4\sin^2(\pi\nu_1 h/2)e_{h,\nu}(x).&
\end{aligned}$$

Applying the same reasoning to the second variable x_2, we get

$$\begin{aligned}
2e_{h,\nu}(x) - e_{h,\nu}(x_1, x_2 + h) - e_{h,\nu}(x_1, x_2 - h)& \\
= 2\big(1 - 1 + 2\sin^2(\pi\nu_2 h/2)\big)e_{h,\nu}(x) = 4\sin^2(\pi\nu_2 h/2)e_{h,\nu}(x),&
\end{aligned}$$

and adding both equations finally yields

$$-\Delta_h e_{h,\nu}(x) = 4h^{-2}(\sin^2(\pi\nu_1 h/2) + \sin^2(\pi\nu_2 h/2))e_{h,\nu}(x) = \lambda_{h,\nu}e_{h,\nu}(x).$$

In order to establish the orthogonality of the eigenvectors, we make use of the trigonometric identity

$$\sum_{k=1}^{N}\sin(\pi\nu kh)\sin(\pi\mu kh) = \begin{cases}(N+1)/2 & \text{if } \nu = \mu,\\ 0 & \text{otherwise}\end{cases} \qquad \text{for all } \nu, \mu \in [1:N].$$

Let $\nu, \mu \in [1:N]^2$. We have

$$\langle e_{h,\nu}, e_{h,\mu}\rangle_{\Omega_h} = h^2\sum_{x\in\Omega_h} e_{h,\nu}(x)e_{h,\mu}(x) = h^2\sum_{i,j=1}^{N} e_{h,\nu}(ih, jh)e_{h,\mu}(ih, jh)$$

$$\begin{aligned}
&= 4h^2 \sum_{i,j=1}^{N} \sin(\pi\nu_1 ih)\sin(\pi\nu_2 jh)\sin(\pi\mu_1 ih)\sin(\pi\mu_2 jh)\\
&= 4h^2 \left(\sum_{i=1}^{N} \sin(\pi\nu_1 ih)\sin(\pi\mu_1 ih)\right)\left(\sum_{j=1}^{N} \sin(\pi\nu_2 jh)\sin(\pi\mu_2 jh)\right)\\
&= 4h^2 \begin{cases}(N+1)^2/4 & \text{if } \nu=\mu,\\ 0 & \text{otherwise}\end{cases}\\
&= \begin{cases}1 & \text{if } \nu=\mu,\\ 0 & \text{otherwise.}\end{cases}
\end{aligned}$$

■

We can see that that largest eigenvalue of $-\Delta_h$ is given by

$$\begin{aligned}
\lambda_{\max} := \lambda_{h,(N,N)} &= 2\widehat{\lambda}_{h,N} = 8h^{-2}\sin^2(\pi Nh/2)\\
&= 8h^{-2}\sin^2\left(\frac{\pi}{2}\frac{N}{N+1}\right) \approx 8h^{-2}\sin^2(\pi/2) = 8h^{-2}
\end{aligned}$$

for large values of N. We have already seen in Lemma 2.22 that the spectral radius is a lower bound for any operator norm, therefore the Lipschitz constant C_Δ of the ordinary differential equation (3.5) cannot be smaller than $\lambda_{\max} \approx 8h^{-2} \approx 8N^2$.

Since the Lipschitz constant plays an important role in many stability and convergence estimates, the fact that it grows as we refine the finite difference grid is rather inconvenient.

Fortunately, the equation (3.5) has a redeeming quality: since all eigenvalues of $-\Delta_h$ are strictly positive, the finite difference operator is positive definite, i.e., we have

$$\langle -\Delta_h u_h, u_h\rangle_2 > 0 \qquad \text{for all } u_h \in G_0(\bar\Omega_h)\setminus\{0\}. \tag{3.6}$$

In fact, we even have

$$\langle -\Delta_h u_h, u_h\rangle_2 \geq \lambda_{\min}\|u_h\|_2^2 \qquad \text{for all } u_h \in G_0(\bar\Omega_h),$$

where $\lambda_{\min} := \lambda_{h,(1,1)} \approx 2\pi^2$ denotes the minimal eigenvalue of $-\Delta_h$.

For the right-hand side f introduced in (3.4), this implies

$$\begin{aligned}
&\langle f(t,u_h) - f(t,v_h), u_h - v_h\rangle_2\\
&\qquad = \langle \Delta_h(u_h - v_h), u_h - v_h\rangle_2 \leq -\lambda_{\min}\|u_h - v_h\|_2^2 \qquad \text{for all } u_h, v_h \in G_0(\bar\Omega_h).
\end{aligned}$$

Lemma 3.3 (Perturbations) *Let $\mathcal{V}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{V}}$. Let $f \in C(\mathbb{R}\times\mathcal{V},\mathcal{V})$ satisfy*

$$\langle f(t,v) - f(t,w), v - w\rangle_{\mathcal{V}} \leq -\lambda\|v-w\|_{\mathcal{V}}^2 \qquad \textit{for all } t \in \mathbb{R}_{\geq 0},\ v,w \in \mathcal{V}$$

for a suitable constant $\lambda \in \mathbb{R}_{\geq 0}$. Let $y, z \in C^1(\mathbb{R}_{\geq 0}, \mathcal{V})$ satisfy the ordinary differential equations

$$y'(t) = f(t, y(t)), \qquad z'(t) = f(t, z(t)) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

Then we have

$$\|y(t) - z(t)\|_{\mathcal{V}} \leq e^{-\lambda t}\|y(0) - z(0)\|_{\mathcal{V}} \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

Proof. We consider the function

$$\gamma \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}, \qquad t \mapsto \|y(t) - z(t)\|_{\mathcal{V}}^2.$$

It is continuously differentiable with

$$\begin{aligned}\gamma'(t) &= 2\langle y'(t) - z'(t), y(t) - z(t)\rangle_2 \\ &= 2\langle f(t, y(t)) - f(t, z(t)), y(t) - z(t)\rangle_2 \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.\end{aligned}$$

Due to our assumption, we have

$$\begin{aligned}\gamma'(t) &= 2\langle f(t, y(t)) - f(t, z(t)), y(t) - z(t)\rangle_2 \\ &\leq -2\lambda\|y(t) - z(t)\|_{\mathcal{V}}^2 = -2\lambda\gamma(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.\end{aligned}$$

We have to prove $\gamma(t) \leq e^{-2\lambda t}\gamma(0)$.

To this end, we follow the proof of [11, Satz 9.IX] and introduce

$$\widehat{\gamma} \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}, \qquad t \mapsto e^{-2\lambda t}\gamma(0)$$

and $\omega := \widehat{\gamma} - \gamma$. Our goal is to prove $\omega \geq 0$.

Since ω is continuous, every point $t \in \mathbb{R}_{\geq 0}$ with $\omega(t) < 0$ would be surrounded by an interval $[a, b]$ such that $\omega|_{[a,b]} \leq 0$. Since $\omega(0) = 0$ holds, we can enlarge the interval to ensure $\omega(a) = 0$. We have

$$\omega'(t) = \widehat{\gamma}'(t) - \gamma'(t) \geq -2\lambda\widehat{\gamma}(t) + 2\lambda\gamma(t) = -2\lambda\omega(t) \geq 0 \qquad \text{for all } t \in [a, b].$$

Due to the fundamental theorem of calculus and $\omega(a) = 0$, this implies $\omega|_{[a,b]} \geq 0$ and therefore $\omega|_{[a,b]} = 0$. We conclude that there can be no $t \in \mathbb{R}_{\geq 0}$ with $\omega(t) < 0$.

We have proven

$$\|y(t) - z(t)\|_{\mathcal{V}}^2 = \gamma(t) \leq \widehat{\gamma}(t) = e^{-2\lambda t}\gamma(0) = e^{-2\lambda t}\|y(0) - z(0)\|_{\mathcal{V}}^2 \qquad \text{for all } t \in \mathbb{R}_{\geq 0},$$

and taking the square root yields the required estimate. ■

In our case, this lemma states that the ordinary differential equation (3.5) is *very* stable with regard to perturbations of the initial value, even if the Lipschitz constant C_Δ is large.

Remark 3.4 (Limit $t \to \infty$) *In our model problem, we have $\lambda_{min} > 0$, and Lemma 3.3 yields that all solutions of (3.5) have to converge to the same limit for $t \to \infty$, no matter what the initial value at $t = a$ is.*

If g_h is fixed, i.e., if $g_h(t) = g_{h,\infty}$ holds for a function $g_{h,\infty} \in G_0(\bar{\Omega}_h)$, we can even compute this limit: if $u_{h,\infty} \in G_0(\bar{\Omega}_h)$ solves

$$-\Delta_h u_{h,\infty} = g_{h,\infty},$$

we have

$$f(t, u_{h,\infty}) = g_{h,\infty} + \Delta_h u_{h,\infty} = 0 \qquad\qquad \text{for all } t \in \mathbb{R},$$

so the constant function $t \mapsto u_{h,\infty}$ is a solution of (3.5). Due to Lemma 3.3, all solutions have to converge to $u_{h,\infty}$ for $t \to \infty$.

3.3. Time-stepping methods

Let us take a look at approximation methods for general initial value problems. Let $\mathcal{V}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$.

The initial value problem (3.5) is of the following form:

Let $y_0 \in \mathcal{V}$, and let $f \in C(\mathbb{R}_{\geq 0} \times \mathcal{V}, \mathcal{V})$. Find $y \in C^1(\mathbb{R}_{\geq 0}, \mathcal{V})$ such that

$$y(0) = y_0, \qquad y'(t) = f(t, y(t)) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}. \tag{3.7}$$

We consider time-stepping methods for finding an approximate solution. The basic idea is to replace the continuous time interval $\mathbb{R}_{\geq 0}$ by discrete points

$$0 = t_0 < t_1 < t_2 < \dots$$

and try to approximate $y(t_i)$ for these points. To keep the presentation simple, we fix a step size $\delta \in \mathbb{R}_{>0}$ and let

$$t_i := i\delta \qquad\qquad \text{for all } i \in \mathbb{N}_0.$$

A *single-step method* is defined by a *time-step function*

$$\Psi : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathcal{V} \to \mathcal{V}$$

that takes the current time t_i, the time step δ and the current value $y(t_i)$ and computes an approximation of $y(t_{i+1})$. The resulting sequence of approximate solutions is given by

$$\widetilde{y}(0) := y_0, \qquad \widetilde{y}(t_{i+1}) := \Psi(t_i, \delta, \widetilde{y}(t_i)) \qquad \text{for all } i \in \mathbb{N}_0.$$

In order to construct Ψ, we can take our cue from the fundamental theorem of calculus: due to (3.7), we have

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} y'(s)\,ds = \int_{t_i}^{t_{i+1}} f(s, y(s))\,ds,$$

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s))\,ds,$$

so it seems straightforward to look for a quadrature formula to approximate the integral.

Unfortunately, this quadrature formula cannot evaluate the integrand in the interval $[t_i, t_{i+1}]$, since only $y(t_i)$ is at our disposal.

A simple solution is to use a quadrature formula that only uses this one value.

Lemma 3.5 (Rectangle rule) *Let $g \in C^1([a,b], \mathcal{V})$. There are $\eta_a, \eta_b \in [a,b]$ such that*

$$\left\| \int_a^b g(s)\,ds - (b-a)g(a) \right\| \leq \frac{(b-a)^2}{2} \|g'(\eta_a)\|,$$
$$\left\| \int_a^b g(s)\,ds - (b-a)g(b) \right\| \leq \frac{(b-a)^2}{2} \|g'(\eta_b)\|.$$

Proof. For the first statement, we consider the function

$$\varphi\colon [a,b] \to \mathbb{R}, \qquad s \mapsto s - b.$$

Using partial integration and $\varphi' = 1$, we find

$$\begin{aligned} \int_a^b g(s)\,ds - (b-a)g(a) &= \int_a^b g(s) - g(a)\,ds = \int_a^b \varphi'(s)(g(s) - g(a))\,ds, \\ &= \Big[\varphi(s)(g(s) - g(a))\Big]_{s=a}^b - \int_a^b \varphi(s) g'(s)\,ds = -\int_a^b \varphi(s) g'(s)\,ds, \end{aligned}$$

and the mean value theorem yields $\eta_a \in [a,b]$ such that

$$\begin{aligned} \left\| \int_a^b g(s)\,ds - (b-a)g(a) \right\| &\leq \int_a^b |\varphi(s)| \|g'(s)\|\,ds \\ &= \|g'(\eta_a)\| \int_a^b b - s\,ds = \|g'(\eta_a)\| \frac{(b-a)^2}{2}. \end{aligned}$$

The second statement can be proven by using the same arguments for $\varphi(s) = s - a$. ■

Applying the rectangle quadrature rule to the left point of interval $[t_i, t_{i+1}]$, we find

$$\int_{t_i}^{t_{i+1}} f(s, y(s))\,ds \approx \delta f(t_i, y(t_i))$$

and therefore

$$\Psi(t, \delta, y(t)) = y(t) + \delta f(t, y(t)).$$

This defines the *explicit Euler method.* For our model problem (3.5), we have

$$\begin{aligned} u_h(t_{i+1}) &= u_h(t_i) + \delta(g_h(t_i) + \Delta_h u_h(t_i)), \\ \Psi(t_i, \delta, x) &= x + \delta g_h(t_i) + \delta \Delta_h x, \end{aligned}$$

so performing one time step requires only linear combinations of grid functions and one evaluation of the finite difference operator.

We can also use the right point of $[t_i, t_{i+1}]$ as a quadrature point. This yields

$$\int_{t_i}^{t_{i+1}} f(s, y(s))\, ds \approx \delta f(t_{i+1}, y(t_{i+1})).$$

We find

$$\begin{aligned} y(t_{i+1}) &\approx y(t_i) + \delta f(t_{i+1}, y(t_{i+1})), \\ y(t_{i+1}) - \delta f(t_{i+1}, y(t_{i+1})) &\approx y(t_i) \end{aligned}$$

and have to solve this equation to obtain $y(t_{i+1})$. This approach is known as the *implicit Euler method.*

For our model problem (3.5), we find

$$\begin{aligned} y(t_{i+1}) - \delta f(t_{i+1}, y(t_{i+1})) = u_h(t_{i+1}) - \delta g_h(t_{i+1}) - \delta\Delta_h u_h(t_{i+1}) &\approx u_h(t_i), \\ (I - \delta\Delta_h) u_h(t_{i+1}) &\approx u_h(t_i) + \delta g_h(t_{i+1}), \\ \Psi(t, \delta, x) &= (I - \delta\Delta_h)^{-1}(x + \delta g_h(t_{i+1})), \end{aligned}$$

so performing one time step requires us to solve the finite difference equation. This is computationally considerably more expensive than the explicit Euler method, but it has particularly attractive properties regarding parabolic equations.

Since both Euler methods approximate the integral essentially by the integral of a constant function, they are only first-order accurate. In order to reach a higher accuracy, we can approximate the integral by the trapezoidal rule.

Lemma 3.6 (Trapezoidal rule) *Let $g \in C^2([a,b], \mathcal{V})$. There is an $\eta \in [a,b]$ such that*

$$\left\| \int_a^b g(s)\, ds - \frac{b-a}{2}(g(a) + g(b)) \right\| \leq \frac{(b-a)^3}{12} \|g''(\eta)\|.$$

Proof. We consider the function

$$\varphi\colon [a,b] \to \mathbb{R}, \qquad s \mapsto \frac{(s-a)(s-b)}{2},$$

satisfying $\varphi'' = 1$ and $\varphi(a) = \varphi(b) = 0$. The linear function interpolating g in a and b is given by

$$p\colon [a,b] \to \mathcal{V}, \qquad s \mapsto \frac{b-s}{b-a} g(a) + \frac{s-a}{b-a} g(b),$$

and its integral coincides with the quadrature rule

$$\int_a^b p(s)\, ds = \int_a^b \frac{b-s}{b-a}\, ds\ g(a) + \int_a^b \frac{s-a}{b-a}\, ds\ g(b) = \frac{b-a}{2} g(a) + \frac{b-a}{2} g(b).$$

Using partial integration twice, we find

$$\begin{aligned}\int_a^b g(s)\,ds - \frac{b-a}{2}(g(a)-g(b)) &= \int_a^b g(s)-p(s)\,ds = \int_a^b \varphi''(s)(g(s)-p(s))\,ds\\ &= \Big[\varphi'(s)(g(s)-p(s))\Big]_{s=a}^b - \int_a^b \varphi'(s)(g'(s)-p'(s))\,ds\\ &= -\int_a^b \varphi'(s)(g'(s)-p'(s))\,ds\\ &= -\Big[\varphi(s)(g'(s)-p'(s))\Big]_{s=a}^b + \int_a^b \varphi(s)g''(s)\,ds = \int_a^b \varphi(s)g''(s)\,ds.\end{aligned}$$

By the mean value theorem we find $\eta \in [a,b]$ such that

$$\begin{aligned}\left\| \int_a^b g(s)\,ds - \frac{b-a}{2}(g(a)-g(b))\right\| &\leq \int_a^b |\varphi(s)|\,\|g''(s)\|\,ds\\ &= -\|g''(\eta)\| \int_a^b \varphi(s)\,ds = \frac{(b-a)^3}{12}\|g''(\eta)\|.\end{aligned}$$

■

Approximating the integral by the trapezoidal rule yields

$$\int_{t_i}^{t_{i+1}} f(s,y(s))\,ds \approx \frac{\delta}{2}(f(t_i,y(t_i)) + f(t_{i+1},y(t_{i+1}))).$$

This approach requires us to solve

$$\begin{aligned}y(t_{i+1}) &\approx y(t_i) + \frac{\delta}{2}f(t_i,y(t_i)) + \frac{\delta}{2}f(t_{i+1},y(t_{i+1})),\\ y(t_{i+1}) - \frac{\delta}{2}f(t_{i+1},y(t_{i+1})) &\approx y(t_i) + \frac{\delta}{2}f(t_i,y(t_i))\end{aligned}$$

to obtain an approximation of $y(t_{i+1})$. The resulting algorithm is known as the *Crank-Nicolson method* [5]. Since the trapezoidal rule integrates linear functions exactly, we can expect second-order convergence.

For our model problem, we find

$$\begin{aligned}u_h(t_{i+1}) - \frac{\delta}{2}g_h(t_{i+1}) - \frac{\delta}{2}\Delta_h u_h(t_{i+1}) &\approx u_h(t_i) + \frac{\delta}{2}g_h(t_i) + \frac{\delta}{2}\Delta_h u_h(t_i),\\ \left(I - \frac{\delta}{2}\Delta_h\right) u_h(t_{i+1}) &\approx u_h(t_i) + \delta\frac{g_h(t_i)+g_h(t_{i+1})}{2} + \frac{\delta}{2}\Delta_h u_h(t_i),\end{aligned}$$

so the function Ψ is given by

$$\Psi(t,\delta,x) = \left(I - \frac{\delta}{2}\Delta_h\right)^{-1}\left(x + \delta\frac{g_h(t_i)+g_h(t_{i+1})}{2} + \frac{\delta}{2}\Delta_h x\right). \tag{3.8}$$

3.4. Consistency, stability, convergence

We have to ensure that the time-stepping methods approximate the solution of the ordinary differential equation sufficiently well. As in the previous chapter, we split the error analysis into two parts: we investigate the *consistency* of the time-stepping methods, i.e., how well the algorithm approximates the solution during only one step, and we establish the *stability* of the methods, i.e., how sensitive they react to perturbations. Combining consistency and stability allows us to establish the *convergence* of the methods.

In order to obtain the required results, particularly estimates regarding the stability of the methods, we have to take advantage of the methods' properties. A standard assumption is that the right-hand side function f is *Lipschitz-continuous* with respect to its third parameter, i.e., that there is a constant $L_f \in \mathbb{R}_{\geq 0}$ such that

$$\|f(t,v) - f(t,w)\| \leq L_f \|v - w\| \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ v,w \in \mathcal{V}, \tag{3.9}$$

In our case, this condition is problematic, since the best possible Lipschitz constant is given by $L_f = \|\Delta_h\| \sim h^{-2}$, i.e., the Lipschitz constant will grow rapidly when we refine the finite difference grid.

That is why we also consider an alternative condition that is frequently sufficient to obtain the required results: we assume that the differential equation is *contracting*, i.e., that

$$\langle f(t,v) - f(t,w), v - w \rangle \leq 0 \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ v,w \in \mathcal{V}. \tag{3.10}$$

In our case, we have $f(t,v) - f(t,w) = \Delta_h(v - w)$, and Lemma 3.2 yields that $-\Delta_h$ is positive definite, so Δ_h has to be negative definite, i.e., (3.10) holds.

Definition 3.7 (Consistency) *Let $p \in \mathbb{N}$. A time-stepping method, characterized by its time-step function Ψ, is* consistent of p-th order *with a solution y of the initial value problem (3.7), if there are constants $C_{cn} \in \mathbb{R}_{\geq 0}$ and $\delta_{max} \in \mathbb{R}_{>0} \cup \{\infty\}$ such that*

$$\|y(t+\delta) - \Psi(t,\delta,y(t))\| \leq C_{cn}\delta^{p+1} \qquad \textit{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in (0,\delta_{max}).$$

Lemma 3.8 (Explicit Euler method) *Let $y \in C^2(\mathbb{R}_{\geq 0}, \mathcal{V})$ be a solution of (3.7). The explicit Euler method satisfies*

$$\|y(t+\delta) - \Psi(t,\delta,y(t))\| \leq \frac{\delta^2}{2}\|y''\|_{\infty,[t,t+\delta]} \qquad \textit{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{>0},$$

i.e., the explicit Euler method is consistent of first order.

Proof. Let $t \in \mathbb{R}_{\geq 0}$ and $\delta \in \mathbb{R}_{>0}$. By the fundamental theorem of calculus, we have

$$y(t+\delta) = y(t) + \int_t^{t+\delta} y'(s)\,ds,$$
$$\Psi(t,\delta,y(t)) = y(t) + \delta f(t,y(t)) = y(t) + \delta y'(t).$$

With Lemma 3.5, we find $\eta \in [a, b]$ such that

$$\|y(t+\delta) - \Psi(t, \delta, y(t))\| \leq \frac{\delta^2}{2} \|y''(\eta)\|.$$

Taking the maximum yields our estimate. ■

Lemma 3.9 (Implicit Euler method) *Let $y \in C^2(\mathbb{R}_{\geq 0}, \mathcal{V})$ be a solution of (3.7). If there is a Lipschitz constant $L_f \in \mathbb{R}_{\geq 0}$ such that (3.9) holds, the implicit Euler method satisfies*

$$\|y(t+\delta) - \Psi(t, \delta, y(t))\| \leq \frac{\delta^2}{2(1 - L_f \delta)} \|y''\|_{\infty,[t,t+\delta]} \quad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in (0, 1/L_f).$$

If instead we have (3.10), we obtain the stronger result

$$\|y(t+\delta) - \Psi(t, \delta, y(t))\| \leq \frac{\delta^2}{2} \|y''\|_{\infty,[t,t+\delta]} \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{>0}.$$

In both cases, the implicit Euler method is consistent of first order.

Proof. Left as an exercise, e.g., by following the lines of the proof of Lemma 3.10. ■

Lemma 3.10 (Crank-Nicolson method) *Let $y \in C^3(\mathbb{R}_{\geq 0}, \mathcal{V})$ be a solution of (3.7). If there is a constant $L_f \in \mathbb{R}_{\geq 0}$ such that the Lipschitz condition (3.9) holds, the Crank-Nicolson method satisfies*

$$\|y(t+\delta) - \Psi(t, \delta, y(t))\| \leq \frac{\delta^3}{12(1 - L_f \delta/2)} \|y''\|_{\infty,[t,t+\delta]} \quad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in (0, 2/L_f).$$

If instead we have (3.10), we obtain the stronger result

$$\|y(t+\delta) - \Psi(t, \delta, y(t))\| \leq \frac{\delta^3}{12} \|y''\|_{\infty,[t,t+\delta]} \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{>0}.$$

In both cases, the Crank-Nicolson method is consistent of second order.

Proof. Let $t \in \mathbb{R}_{\geq 0}$ and $\delta \in \mathbb{R}_{>0}$. We denote the error by

$$e := y(t+\delta) - \tilde{y}(t+\delta), \qquad \text{where } \tilde{y}(t+\delta) := \Psi(t, \delta, y(t)).$$

By the fundamental theorem of calculus, we have

$$y(t+\delta) = y(t) + \int_t^{t+\delta} y'(s)\, ds,$$

and the definition of the Crank-Nicolson method yields

$$\tilde{y}(t+\delta) = y(t) + \frac{\delta}{2}\big(f(t, y(t)) + f(t+\delta, \tilde{y}(t+\delta))\big)$$

$$\begin{aligned}
&= y(t) + \frac{\delta}{2}\big(f(t,y(t)) + f(t+\delta, y(t+\delta))\big) \\
&\qquad + \frac{\delta}{2}\big(f(t+\delta,\tilde{y}(t+\delta)) - f(t+\delta,y(t+\delta))\big) \\
&= y(t) + \frac{\delta}{2}\big(y'(t) + y'(t+\delta)\big) + \frac{\delta}{2}\big(f(t+\delta,\tilde{y}(t+\delta)) - f(t+\delta,y(t+\delta))\big).
\end{aligned}$$

We can split the error into the quadrature error

$$e_q := \int_t^{t+\delta} y'(s)\,ds - \frac{\delta}{2}(y'(t) + y'(t+\delta))$$

and the approximation error

$$e_a := \frac{\delta}{2}\big(f(t+\delta,y(t+\delta)) - f(t+\delta,\tilde{y}(t+\delta))\big).$$

For the quadrature error, Lemma 3.6 yields

$$\|e_q\| \leq \frac{\delta^3}{12}\|y'''\|_{\infty,[t,t+\delta]}.$$

Let now (3.9) hold, and let $\delta < 2/L_f$. Then we have

$$\begin{aligned}
\|e_a\| &= \frac{\delta}{2}\big\|f(t+\delta,y(t+\delta)) - f(t+\delta,\tilde{y}(t+\delta))\big\| \\
&\leq \frac{\delta}{2}L_f\|y(t+\delta) - \tilde{y}(t+\delta)\| = \frac{\delta L_f}{2}\|e\|, \\
\|e\| \leq \|e_q\| + \|e_a\| &\leq \frac{\delta^3}{12}\|y'''\|_{\infty,[t,t+\delta]} + \frac{\delta L_f}{2}\|e\|,
\end{aligned}$$

and

$$(1 - \delta L_f/2)\|e\| \leq \frac{\delta^3}{12}\|y'''\|_{\infty,[t,t+\delta]}$$

yields the first estimate.

For the second estimate, let (3.10) hold instead of (3.9). Since $\mathcal{V}$ is a Hilbert space, we have

$$\begin{aligned}
\|e\|^2 &= \langle e, e\rangle = \langle e_q + e_a, e\rangle \\
&= \langle e_q, e\rangle + \frac{\delta}{2}\langle f(t+\delta,y(t+\delta)) - f(t+\delta,\tilde{y}(t+\delta)), y(t+\delta) - \tilde{y}(t+\delta)\rangle \\
&\leq \langle e_q, e\rangle \leq \|e_q\|\,\|e\| \leq \frac{\delta^3}{12}\|y'''\|_{\infty,[t,t+\delta]}\|e\|
\end{aligned}$$

by the Cauchy-Schwarz inequality, and dividing by $\|e\|$ yields the desired estimate. ■

Remark 3.11 (Global consistency) *The estimates of Lemmas 3.8, 3.9, and 3.10 involve the maximum norm of derivatives of the solution y in $[t, t+\delta]$, while our Definition 3.7 requires the constant C_{cn} to be independent of t and δ.*

We can fix this issue by considering only the approximation of y in a compact time interval $[a, b]$. Assuming that the relevant derivatives of y are continuous, the maximum norm in $[t, t+\delta]$ can be bounded by the maximum norm in $[a, b]$, and this is indeed a constant that does not depend on t and δ.

Consistency allows us to control the error introduced by one step of our algorithm. Since the next step starts with an approximation instead of the exact solution, we have to investigate how the errors introduced at different steps propagate with time.

Definition 3.12 (Stability) *A time-stepping method, characterized by its time-step function Ψ, is* stable *if there are constants $L_\Psi \in \mathbb{R}$ and $\delta_{max} \in \mathbb{R}_{>0} \cup \{\infty\}$ such that*

$$\|\Psi(t,\delta,v) - \Psi(t,\delta,w)\| \leq (1 + L_\Psi\delta)\|v - w\| \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in (0, \delta_{max}),\ v, w \in \mathcal{V}.$$

The method is unconditionally stable *if we can choose $L_\Psi = 0$.*

Lemma 3.13 (Explicit Euler method) *Assume that $f \in C(\mathbb{R}_{\geq 0} \times \mathcal{V}, \mathcal{V})$ is Lipschitz-continuous in the second argument, i.e., that there is a constant $L_f \in \mathbb{R}_{\geq 0}$ such that*

$$\|f(t,v) - f(t,w)\| \leq L_f\|v - w\| \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ v, w \in \mathcal{V}.$$

Then the explicit Euler method satisfies

$$\|\Psi(t,\delta,v) - \Psi(t,\delta,w)\| \leq (1 + L_f\delta)\|v - w\| \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{>0},\ v, w \in \mathcal{V},$$

i.e., it is stable.

Proof. Let $t \in \mathbb{R}_{\geq 0}$, $\delta \in \mathbb{R}_{\geq 0}$, and $v, w \in \mathcal{V}$. We have

$$\begin{aligned} \widetilde{v} &:= \Psi(t,\delta,v) = v + \delta f(t,v), \\ \widetilde{w} &:= \Psi(t,\delta,w) = w + \delta f(t,w) \end{aligned}$$

and find

$$\begin{aligned} \|\widetilde{v} - \widetilde{w}\| &= \|v - w + \delta(f(t,v) - f(t,w))\| \leq \|v - w\| + \delta\|f(t,v) - f(t,w)\| \\ &\leq \|v - w\| + L_f\delta\|v - w\| = (1 + L_f\delta)\|v - w\|. \end{aligned}$$

■

For the model problem, we can obtain the following more precise estimate.

Lemma 3.14 (Explicit Euler, heat equation) *Let Ψ denote the time-step function of the explicit Euler method for our model problem (3.3), and let*

$$C_\Psi := \max\{1, |1 - \delta\lambda_{max}|\}.$$

We have

$$\|\Psi(t,\delta,u_h) - \Psi(t,\delta,v_h)\|_{\Omega_h} \leq C_\Psi \|u_h - v_h\|_{\Omega_h} \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{\geq 0}, \\ u_h, v_h \in G_0(\bar{\Omega}_h).$$

Proof. Let $t \in \mathbb{R}_{\geq 0}$, $\delta \in \mathbb{R}_{\geq 0}$, and $u_h, v_h \in G_0(\bar{\Omega})$. Due to Lemma 3.2, We can find $(\alpha_\nu)_{\nu\in[1:N]^2}$ such that

$$u_h - v_h = \sum_{\nu\in[1:N]^2} \alpha_\nu e_{h,\nu}.$$

We let

$$\begin{aligned} \widetilde{u}_h &:= \Psi(t,\delta,u_h) = u_h + \delta(g_h(t) + \Delta_h u_h), \\ \widetilde{v}_h &:= \Psi(t,\delta,v_h) = v_h + \delta(g_h(t) + \Delta_h v_h) \end{aligned}$$

and observe

$$\begin{aligned} \widetilde{u}_h - \widetilde{v}_h &= u_h - v_h + \delta\Delta_h(u_h - v_h) = (I + \delta\Delta_h)(u_h - v_h) \\ &= \sum_{\nu\in[1:N]^2} (I + \delta\Delta_h)\alpha_\nu e_{h,\nu} = \sum_{\nu\in[1:N]^2} (1 - \delta\lambda_{h,\nu})\alpha_\nu e_{h,\nu}, \\ \|\widetilde{u}_h - \widetilde{v}_h\|^2 &= \langle \widetilde{u}_h - \widetilde{v}_h, \widetilde{u}_h - \widetilde{v}_h \rangle \\ &= \sum_{\nu\in[1:N]^2} \sum_{\mu\in[1:N]^2} (1 - \delta\lambda_{h,\nu})(1 - \delta\lambda_{h,\mu})\alpha_\nu\alpha_\mu \langle e_{h,\nu}, e_{h,\mu} \rangle \\ &= \sum_{\nu\in[1:N]^2} (1 - \delta\lambda_{h,\nu})^2 \alpha_\nu^2. \end{aligned}$$

Since $s \mapsto (1-s)^2$ is convex, we have

$$(1 - \delta\lambda_{h,\nu})^2 \leq \max\{(1 - \delta\lambda_{\min})^2, (1 - \delta\lambda_{\max})^2\} \leq \max\{1, |1 - \delta\lambda_{\max}|\}^2 = C_\Psi^2$$

and conclude

$$\begin{aligned} \|\widetilde{u}_h - \widetilde{v}_h\|^2 &= \sum_{\nu\in[1:N]^2} (1 - \delta\lambda_{h,\nu})^2 \alpha_\nu^2 \leq C_\Psi^2 \sum_{\nu\in[1:N]^2} \alpha_\nu^2 \\ &= C_\Psi^2 \sum_{\nu\in[1:N]^2} \sum_{\mu\in[1:N]^2} \alpha_\nu\alpha_\mu \langle e_{h,\nu}, e_{h,\mu} \rangle \\ &= C_\Psi^2 \langle u_h - v_h, u_h - v_h \rangle = C_\Psi^2 \|u_h - v_h\|^2. \end{aligned}$$

■

This estimate is quite sharp: if we choose $u_h = e_{h,(N,N)}$ and $v_h = 0$, we have

$$\|\Psi(t,\delta,u_h) - \Psi(t,\delta,v_h)\|_{\Omega_h} = |1 - \delta\lambda_{\max}|\,\|u_h - v_h\|_{\Omega_h}.$$

In particular, if we have $\delta\lambda_{\max} > 1$, the solution computed by the explicit Euler method will change its sign at each step. For $\delta\lambda_{\max} > 2$, the approximate solution will diverge rapidly for $t \to \infty$, although we have seen in Lemma 3.3 that the exact solution converges.

In order to obtain a stable method, we have to ensure

$$\delta \leq 1/\lambda_{\max} \approx h^2/8. \tag{3.11}$$

This is called the *Courant-Friedrichs-Lewy condition* (abbreviated as *CFL condition*) for our discretization of the heat equation [4]. Bounds like this are common for explicit time-stepping schemes for parabolic or hyperbolic partial differential equations.

Lemma 3.15 (Implicit Euler method) *Assume that (3.10) holds. Then the implicit Euler method satisfies*

$$\|\Psi(t,\delta,v) - \Psi(t,\delta,w)\| \leq \|v - w\| \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{\geq 0},\ v,w \in \mathcal{V},$$

i.e., it is unconditionally stable.

Proof. Let $t \in \mathbb{R}_{\geq 0}$, $\delta \in \mathbb{R}_{\geq 0}$ and $v,w \in \mathcal{V}$. We define

$$\widetilde{v} := \Psi(t,\delta,v), \qquad \widetilde{w} := \Psi(t,\delta,w)$$

and obtain

$$\widetilde{v} = v + \delta f(t,\widetilde{v}), \qquad \widetilde{w} = w + \delta f(t,\widetilde{w}).$$

Using (3.10) and the Cauchy-Schwarz inequality, we find

$$\begin{aligned}\|\widetilde{v} - \widetilde{w}\|^2 &= \langle \widetilde{v} - \widetilde{w}, \widetilde{v} - \widetilde{w}\rangle = \langle v + \delta f(t+\delta,\widetilde{v}) - w - \delta f(t+\delta,\widetilde{w}), \widetilde{v} - \widetilde{w}\rangle \\ &= \langle v - w, \widetilde{v} - \widetilde{w}\rangle + \delta\langle f(t+\delta,\widetilde{v}) - f(t+\delta,\widetilde{w}), \widetilde{v} - \widetilde{w}\rangle \\ &\leq \langle v - w, \widetilde{v} - \widetilde{w}\rangle \leq \|v - w\|\,\|\widetilde{v} - \widetilde{w}\|,\end{aligned}$$

and dividing by $\|\widetilde{v} - \widetilde{w}\|$ yields our result. ∎

For the implicit Euler method, we can also obtain a sharper estimate if we consider our model problem.

Lemma 3.16 (Implicit Euler, heat equation) *Let Ψ denote the time-step function of the implicit Euler method for the model problem (3.3). We have*

$$\|\Psi(t,\delta,u_h) - \Psi(t,\delta,v_h)\|_{\Omega_h} \leq \frac{1}{1+\delta\lambda_{min}}\|u_h - v_h\|_{\Omega_h} \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{\geq 0},\ u_h, v_h \in G_0(\bar{\Omega}_h).$$

Proof. Let $t \in \mathbb{R}_{\geq 0}$, $\delta \in \mathbb{R}_{\geq 0}$, and $u_h, v_h \in G_0(\bar{\Omega}_h)$. Due to Lemma 3.2, we can find $(\alpha_\nu)_{\nu\in[1:N]^2}$ such that

$$u_h - v_h = \sum_{\nu\in[1:N]^2} \alpha_\nu e_{h,\nu}.$$

We let

$$\begin{aligned}\widetilde{u}_h &:= \Psi(t,\delta,u_h) = u_h + \delta(g_h(t) + \Delta_h \widetilde{u}_h),\\ \widetilde{v}_h &:= \Psi(t,\delta,v_h) = v_h + \delta(g_h(t) + \Delta_h \widetilde{v}_h)\end{aligned}$$

and observe

$$\begin{aligned}\widetilde{u}_h - \widetilde{v}_h &= u_h - v_h + \delta\Delta_h(\widetilde{u}_h - \widetilde{v}_h),\\ (I - \delta\Delta_h)(\widetilde{u}_h - \widetilde{v}_h) &= u_h - v_h,\\ \widetilde{u}_h - \widetilde{v}_h &= (I-\delta\Delta_h)^{-1}(u_h - v_h) = \sum_{\nu\in[1:N]^2} (I-\delta\Delta_h)^{-1}\alpha_\nu e_{h,\nu}\\ &= \sum_{\nu\in[1:N]^2} \frac{1}{1+\delta\lambda_{h,\nu}}\alpha_\nu e_{h,\nu},\\ \|\widetilde{u}_h - \widetilde{v}_h\|^2 &= \langle \widetilde{u}_h - \widetilde{v}_h, \widetilde{u}_h - \widetilde{v}_h\rangle\\ &= \sum_{\nu\in[1:N]^2}\sum_{\mu\in[1:N]^2} \frac{1}{1+\delta\lambda_{h,\nu}}\frac{1}{1+\delta\lambda_{h,\mu}}\alpha_\nu\alpha_\mu\langle e_{h,\nu}, e_{h,\mu}\rangle\\ &= \sum_{\nu\in[1:N]^2} \frac{1}{(1+\delta\lambda_{h,\nu})^2}\alpha_\nu^2 \leq \sum_{\nu\in[1:N]^2} \frac{1}{(1+\delta\lambda_{\min})^2}\alpha_\nu^2\\ &= \frac{1}{(1+\delta\lambda_{\min})^2}\sum_{\nu\in[1:N]^2}\sum_{\mu\in[1:N]^2} \alpha_\nu\alpha_\mu\langle e_{h,\nu}, e_{h,\mu}\rangle\\ &= \frac{1}{(1+\delta\lambda_{\min})^2}\|u_h - v_h\|^2.\end{aligned}$$

■

We can see that the implicit Euler method is stable with $L_\Psi = 0$ for the model problem (3.3). If we assume $\delta \leq 1$, we even have

$$\frac{1}{1+\delta\lambda_{\min}} = 1 - \frac{\lambda_{\min}}{1+\delta\lambda_{\min}}\delta \leq 1 - \frac{\lambda_{\min}}{1+\lambda_{\min}}\delta,$$

i.e., the method is stable with $L_\Psi = -\lambda_{\min}/(1+\lambda_{\min})$.

Finding a general stability result for the Crank-Nicolson method requires special considerations that would lead too far, so we focus only on the model problem.

Lemma 3.17 (Crank-Nicolson, heat equation) *Let Ψ denote the time-step function of the implicit Euler method for the model problem (3.3). We have*

$$\|\Psi(t,\delta,u_h) - \Psi(t,\delta,v_h)\|_{\Omega_h} \leq \|u_h - v_h\|_{\Omega_h} \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{\geq 0},$$

$$u_h, v_h \in G_0(\bar{\Omega}_h),$$

i.e., the Crank-Nicolson method is unconditionally stable.

Proof. Let $t \in \mathbb{R}_{\geq 0}$, $\delta \in \mathbb{R}_{\geq 0}$, and $u_h, v_h \in G_0(\bar{\Omega}_h)$. Due to Lemma 3.2, we can find $(\alpha_\nu)_{\nu \in [1:N]^2}$ such that

$$u_h - v_h = \sum_{\nu \in [1:N]^2} \alpha_\nu e_{h,\nu}.$$

We let

$$\begin{aligned} \widetilde{u}_h &:= \Psi(t, \delta, u_h) = u_h + \frac{\delta}{2}(g_h(t) + \Delta_h u_h) + \frac{\delta}{2}(g_h(t+\delta) + \Delta_h \widetilde{u}_h), \\ \widetilde{v}_h &:= \Psi(t, \delta, v_h) = v_h + \frac{\delta}{2}(g_h(t) + \Delta_h v_h) + \frac{\delta}{2}(g_h(t+\delta) + \Delta_h \widetilde{u}_h), \end{aligned}$$

and observe

$$\begin{aligned} \widetilde{u}_h - \widetilde{v}_h &= u_h - v_h + \frac{\delta}{2}\Delta_h(u_h - v_h) + \frac{\delta}{2}\Delta_h(\widetilde{u}_h - \widetilde{v}_h), \\ \Bigl(I - \frac{\delta}{2}\Delta_h\Bigr)(\widetilde{u}_h - \widetilde{v}_h) &= \Bigl(I + \frac{\delta}{2}\Delta_h\Bigr)(u_h - v_h), \\ \widetilde{u}_h - \widetilde{v}_h &= \Bigl(I - \frac{\delta}{2}\Delta_h\Bigr)^{-1}\Bigl(I + \frac{\delta}{2}\Delta_h\Bigr)(u_h - v_h), \\ \widetilde{u}_h - \widetilde{v}_h &= \Bigl(I - \frac{\delta}{2}\Delta_h\Bigr)^{-1}\Bigl(I + \frac{\delta}{2}\Delta_h\Bigr) \sum_{\nu \in [1:N]^2} \alpha_\nu e_{h,\nu} \\ &= \sum_{\nu \in [1:N]^2} \frac{1 - \delta\lambda_{h,\nu}/2}{1 + \delta\lambda_{h,\nu}/2} \alpha_\nu e_{h,\nu}, \\ \|\widetilde{u}_h - \widetilde{v}_h\|^2 &= \sum_{\nu \in [1:N]^2} \Bigl(\frac{1 - \delta\lambda_{h,\nu}/2}{1 + \delta\lambda_{h,\nu}/2}\Bigr)^2 \alpha_\nu^2. \end{aligned}$$

We have to investigate the function

$$\gamma \colon \mathbb{R}_{\geq 0} \to \mathbb{R}, \qquad s \mapsto \frac{1-s}{1+s}.$$

Due to

$$\gamma'(s) = \frac{-(1+s) - (1-s)}{(1+s)^2} = \frac{-2}{(1+s)^2} \qquad \text{for all } s \in \mathbb{R}_{\geq 0},$$

the function is monotonic decreasing. We have

$$\gamma(0) = 1, \qquad \gamma(1) = 0, \qquad \lim_{s \to \infty} \gamma(s) = -1$$

and conclude

$$\gamma(s)^2 \leq 1 \qquad \text{for all } s \in \mathbb{R}_{\geq 0}.$$

This means

$$\|\widetilde{u}_h - \widetilde{v}_h\|^2 = \sum_{\nu\in[1:N]^2} \gamma(\delta\lambda_{h,\nu}/2)^2 \alpha_\nu^2 \leq \sum_{\nu\in[1:N]^2} \alpha_\nu^2 = \|u_h - v_h\|^2.$$

■

Remark 3.18 (Oscillations) *The function γ introduced in the proof of Lemma 3.17 is negative for arguments greater than one.*

This means that the sign of the ν-th eigenvector component changes if $\delta\lambda_{h,\nu} > 2$ holds, i.e., we have an unconditionally stable method that still may produce oscillations for high-frequency eigenvectors if the step size δ is too large.

Exercise 3.19 (Midpoint rule) *We can define another time-stepping scheme based on the midpoint quadrature rule: for any $g \in C^2([a,b],\mathcal{V})$, we can find $\eta \in [a,b]$ such that*

$$\Big\| \int_a^b g(s)\,ds - (b-a)g\Big(\frac{a+b}{2}\Big)\Big\| \leq \frac{(b-a)^3}{24}\|g''(\eta)\|.$$

Based on the approximation

$$y(t+\delta) = y(t) + \int_t^{t+\delta} f(s,y(s))\,ds \approx y(t) + \delta f\Big(t+\delta/2, \frac{y(t)+y(t+\delta)}{2}\Big),$$

we define Ψ as the solution — if it exists — of

$$\Psi(t,\delta,v) = v + \delta f\Big(t+\delta/2, \frac{v+\Psi(t,\delta,v)}{2}\Big) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{\geq 0},\ v \in \mathcal{V}.$$

Prove that

- *this time-stepping scheme is unconditionally stable if (3.10) holds, and that*
- *it coincides with the Crank-Nicolson method if there is a linear operator $\mathcal{A} : \mathcal{V} \to \mathcal{V}$ such that $f(t,x) = \mathcal{A}x$ for all $t \in \mathbb{R}_{\geq 0}$ and $x \in \mathcal{V}$.*

In particular, for our model problem the midpoint rule and the trapezoidal rule lead to the same unconditionally stable second-order consistent time-stepping method if g_h is constant.

We can now proceed to prove convergence of our time-stepping method by combining consistency and stability.

Theorem 3.20 (Convergence) *Let $y \in C^1(\mathbb{R}_{\geq 0},\mathcal{V})$ be a solution of (3.7), and let Ψ be the time-step function of a time-stepping method.*

Let this method be consistent of p-th order with the solution, and let it also be stable. We denote the corresponding constants with C_{cn}, L_Ψ and δ_{max}.

If $L_\Psi \neq 0$, the approximations $\widetilde{y}(t_i)$ computed by the method satisfy

$$\|y(t_i) - \widetilde{y}(t_i)\| \leq C_{cn} \frac{e^{L_\Psi t_i} - 1}{L_\Psi} \delta^p \tag{3.12}$$

for all $i \in \mathbb{N}_0$.

If $L_\Psi = 0$, we have

$$\|y(t_i) - \widetilde{y}(t_i)\| \leq C_{cn} t_i \delta^p \qquad \text{for all } i \in \mathbb{N}_0.$$

Proof. We handle the case $L_\Psi \neq 0$ by induction.

Base case: For $i = 0$, we have $y(t_0) = \widetilde{y}(t_0)$ by definition.

Induction assumption: Let $i \in \mathbb{N}_0$ be such that (3.12) holds.

Induction step: We have

$$\begin{aligned}\|y(t_{i+1}) - \widetilde{y}(t_{i+1})\| &= \|y(t_{i+1}) - \Psi(t_i, \delta, y(t_i)) + \Psi(t_i, \delta, y(t_i)) - \Psi(t_i, \delta, \widetilde{y}(t_i))\| \\ &\leq \|y(t_i + \delta) - \Psi(t_i, \delta, y(t_i))\| + \|\Psi(t_i, \delta, y(t_i)) - \Psi(t_i, \delta, \widetilde{y}(t_i))\|.\end{aligned}$$

We can bound the first term by the consistency condition and the second by the stability condition to find

$$\|y(t_{i+1}) - \widetilde{y}(t_{i+1})\| \leq C_{\text{cn}} \delta^{p+1} + (1 + L_\Psi \delta) \|y(t_i) - \widetilde{y}(t_i)\|,$$

and the induction assumption yields

$$\|y(t_{i+1}) - \widetilde{y}(t_{i+1})\| \leq C_{\text{cn}} \delta^{p+1} + (1 + L_\Psi \delta) C_{\text{cn}} \frac{e^{L_\Psi t_i} - 1}{L_\Psi} \delta^p$$

A Taylor-expansion of $s \to e^s$ around zero reveals that for each $s \in \mathbb{R}$, there is an $\eta \in \mathbb{R}$ such that

$$e^s = 1 + s + e^\eta \frac{s^2}{2}.$$

In particular, we have $1 + s \leq e^s$ for all $s \in \mathbb{R}$. In our case, we find $1 + L_\Psi \delta \leq e^{L_\Psi \delta}$ and conclude

$$\begin{aligned}\|y(t_{i+1}) - \widetilde{y}(t_{i+1})\| &\leq C_{\text{cn}} \left(\delta + \frac{(1 + L_\Psi \delta)(e^{L_\Psi t_i} - 1)}{L_\Psi} \right) \delta^p \\ &\leq C_{\text{cn}} \frac{L_\Psi \delta + e^{L_\Psi \delta} e^{L_\Psi t_i} - 1 - L_\Psi \delta}{L_\Psi} \delta^p \\ &= C_{\text{cn}} \frac{e^{L_\Psi t_{i+1}} - 1}{L_\Psi} \delta^p.\end{aligned}$$

Let us now consider the case $L_\Psi = 0$. Let $i \in \mathbb{N}_0$. By definition, stability with $L_\Psi = 0$ implies stability with any constant $\widehat{L}_\Psi > 0$. Introducing the function

$$C \colon \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}, \qquad L \mapsto C_{\text{cn}} \frac{e^{L t_i} - 1}{L} \delta^p,$$

the first part of our proof can be written as

$$\|y(t_i) - \widetilde{y}(t_i)\| \leq C(L) \qquad \text{for all } L \in \mathbb{R}_{>0}.$$

By L'Hôpital's rule, we have

$$\lim_{L\to 0} C(L) = C_{\mathrm{cn}} \frac{\frac{\partial}{\partial L}(e^{Lt_i} - 1)|_{L=0}}{\frac{\partial}{\partial L} L|_{L=0}} \delta^p = C_{\mathrm{cn}} \frac{t_i e^{0t_i}}{1} \delta^p = C_{\mathrm{cn}} t_i \delta^p,$$

and this yields

$$\|y(t_i) - \widetilde{y}(t_i)\| \leq C_{\mathrm{cn}} t_i \delta^p.$$

■

3.5. Influence of the spatial discretization

We have established that we can write the heat equation in the form

$$u(0) = u_0, \qquad u'(t) = g(t) + \Delta u(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}$$

and that approximating u by grid functions and Δ by the finite difference operator Δ_h yields

$$u_h(0) = u_{0,h}, \qquad u_h'(t) = g_h(t) + \Delta_h u_h(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

We have also proven that the discretized problem can be solved by stable time-stepping methods like the implicit Euler method or the Crank-Nicolson method.

Until now, we have neglected to investigate the error introduced by the spatial discretization, i.e.,

$$e_h(t) := u(t)|_{\bar{\Omega}_h} - u_h(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

Our approach is to represent e_h as the solution of an initial value problem, so we compute the derivative

$$\begin{aligned} e_h'(t) &= u'(t)|_{\bar{\Omega}_h} - u_h'(t) \\ &= g(t)|_{\bar{\Omega}_h} + (\Delta u(t))|_{\bar{\Omega}_h} - g_h(t) - \Delta_h u_h(t) \\ &= (\Delta u(t) - \Delta_h u(t))|_{\bar{\Omega}_h} + \Delta_h(u(t) - u_h(t)) \\ &= (\Delta u(t) - \Delta_h u(t))|_{\bar{\Omega}_h} + \Delta_h e_h(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}. \end{aligned}$$

The first term on the right-hand side corresponds to the spatial consistency error

$$v_h(t) := (\Delta u(t) - \Delta_h u(t))|_{\bar{\Omega}_h} \qquad \text{for all } t \in \mathbb{R}_{\geq 0}$$

incurred for the exact solution u. According to Lemma 2.1, we expect

$$\|v_h(t)\|_{\infty,\Omega_h} \leq Ch^2 \qquad \text{for all } t \in \mathbb{R}_{\geq 0}$$

with a suitable constant $C \in \mathbb{R}_{\geq 0}$. Now the error is the solution of

$$e_h(0) = u_0|_{\bar{\Omega}_h} - u_{0,h}, \qquad e_h'(t) = v_h(t) + \Delta_h e_h(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}, \tag{3.13}$$

and this initial value problem can be used to derive error bounds.

To prepare our proof, we require a very simple version of the stability result obtained in Corollary 2.39.

Lemma 3.21 (Stability of $I - \delta\Delta_h$) *Let $\delta \in \mathbb{R}_{\geq 0}$. We have*

$$\|w_h\|_{\infty,\Omega_h} \leq \|(I - \delta\Delta_h)w_h\|_{\infty,\Omega_h} \qquad \textit{for all } w_h \in G_0(\bar{\Omega}_h).$$

Proof. Let $w_h \in G_0(\bar{\Omega}_h)$, and let $x \in \bar{\Omega}_h$ satisfy

$$w_h(y) \leq w_h(x) \qquad \text{for all } y \in \bar{\Omega}_h. \tag{3.14}$$

If x is a boundary point, i.e., $x \in \partial\Omega_h$, we have $w_h(x) = 0$ and therefore $w_h \leq 0$.

Otherwise, i.e., if $x \in \Omega_h$ holds, we have

$$(I - \delta\Delta_h)w_h(x) = w_h(x) + \delta h^{-2} \sum_{y \in N(x)} (w_h(x) - w_h(y)),$$

where $N(x) \subseteq \bar{\Omega}_h$ denotes the neighbours of x in the grid. Due to (3.14), we find $w_h(x) - w_h(y) \geq 0$ for all $y \in N(x)$ and conclude

$$(I - \delta\Delta_h)w_h(x) \geq w_h(x).$$

In summary, w_h is bounded by $\max\{0, (I - \delta\Delta_h)w_h(x)\}$, and therefore also by the maximum norm $\|(I - \delta\Delta_h)w_h\|_{\infty,\Omega_h}$.

Applying the same argument to $-w_h$, we obtain

$$\|w_h\|_{\infty,\Omega_h} \leq \|(I - \delta\Delta_h)w_h\|_{\infty,\Omega_h} \qquad \text{for all } w_h \in G_0(\bar{\Omega}_h).$$

■

Theorem 3.22 (Method of lines) *Let $v_h \in C(\mathbb{R}_{\geq 0}, G_0(\bar{\Omega}_h))$. We consider the solution $e_h \in C^1(\mathbb{R}_{\geq 0}, G_0(\Omega_h))$ of (3.13), i.e.,*

$$e_h(0) = u_0|_{\bar{\Omega}_h} - u_{0,h}, \qquad e_h'(t) = v_h(t) + \Delta_h e_h(t) \qquad \textit{for all } t \in \mathbb{R}_{\geq 0}.$$

We have

$$\|e_h(t)\|_{\infty,\Omega_h} \leq \|u_0|_{\Omega_h} - u_{0,h}\|_{\infty,\Omega_h} + \int_0^t \|v_h(s)\|_{\infty,\Omega_h}\, ds \qquad \textit{for all } t \in \mathbb{R}_{\geq 0}.$$

Proof. In order to use Lemma 3.21, we essentially imitate the structure of the implicit Euler method.

Let $t \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}$, and

$$\delta := t/n, \qquad t_i := i\delta \qquad \text{for all } i \in [0:n].$$

Using the fundamental theorem of calculus, we find

$$\begin{aligned} e_h(t_i) &= e_h(t_{i-1}) + \int_{t_{i-1}}^{t_i} e_h'(s)\,ds = e_h(t_{i-1}) + \int_{t_{i-1}}^{t_i} v_h(s) + \Delta_h e_h(s)\,ds \\ &= e_h(t_{i-1}) + \int_{t_{i-1}}^{t_i} v_h(s)\,ds + \delta\Delta_h e_h(t_i) + \int_{t_{i-1}}^{t_i} \Delta_h e_h(s) - \Delta_h e_h(t_i)\,ds, \end{aligned}$$

and subtracting $\delta\Delta_h e_h(t_i)$ yields

$$\begin{aligned} (I-\delta\Delta_h)e_h(t_i) &= e_h(t_{i-1}) + \int_{t_{i-1}}^{t_i} v_h(s)\,ds + \int_{t_{i-1}}^{t_i} \Delta_h e_h(s) - \Delta_h e_h(t_i)\,ds, \\ \|(I-\delta\Delta_h)e_h(t_i)\|_{\infty,\Omega_h} &\leq \|e_h(t_{i-1})\|_{\infty,\Omega_h} + \int_{t_{i-1}}^{t_i} \|v_h(s)\|_{\infty,\Omega_h} \\ &\quad + \int_{t_{i-1}}^{t_i} \|\Delta_h e_h(s) - \Delta_h e_h(t_i)\|_{\infty,\Omega_h}\,ds. \end{aligned}$$

The first two terms look similar to the ones appearing in the final result, and we can get rid of the third term by taking advantage of the continuity of e_h.

Let $\epsilon \in \mathbb{R}_{>0}$. Since $s \mapsto e_h(s)$ is a continuous function, the same holds for the function $s \mapsto \Delta_h e_h(s)$. Since $[0,t]$ is a compact interval, this function is also uniformly continuous, so we can find $\delta_\epsilon \in \mathbb{R}_{>0}$ such that

$$\|\Delta_h e_h(s_1) - \Delta_h e_h(s_2)\|_{\infty,\Omega_h} \leq \epsilon \qquad \text{for all } s_1, s_2 \in [0,t] \text{ with } |s_1 - s_2| \leq \delta_\epsilon.$$

We choose n large enough to ensure $\delta = t/n \leq \delta_\epsilon$ and therefore $\|\Delta_h e_h(s) - \Delta_h e_h(t_i)\|_{\infty,\Omega_h} \leq \epsilon$ for all $s \in [t_{i-1}, t_i]$, so we can conclude

$$\begin{aligned} \|(I-\delta\Delta_h)e_h(t_i)\|_{\infty,\Omega_h} &\leq \|e_h(t_{i-1})\|_{\infty,\Omega_h} + \int_{t_{i-1}}^{t_i} \|v_h(s)\|_{\infty,\Omega_h}\,ds + \int_{t_{i-1}}^{t_i} \epsilon\,ds \\ &= \|e_h(t_{i-1})\|_{\infty,\Omega_h} + \int_{t_{i-1}}^{t_i} \|v_h(s)\|_{\infty,\Omega_h}\,ds + \delta\epsilon. \end{aligned}$$

Now we can employ Lemma 3.21 to obtain

$$\begin{aligned} \|e_h(t_i)\|_{\infty,\Omega_h} &\leq \|(I-\delta\Delta_h)e_h(t_i)\|_{\infty,\Omega_h} \\ &\leq \|e_h(t_{i-1})\|_{\infty,\Omega_h} + \int_{t_{i-1}}^{t_i} \|v_h(s)\|_{\infty,\Omega_h} + \delta\epsilon \qquad \text{for all } i \in [1:n]. \end{aligned}$$

A straightforward induction yields

$$\|e_h(t_i)\|_{\infty,\Omega_h} \leq \|e_h(t_0)\|_{\infty,\Omega_h} + \int_0^{t_i} \|v_h(s)\|_{\infty,\Omega_h}\,ds + i\delta\epsilon \qquad \text{for all } i \in [0:n],$$

and in particular

$$\begin{aligned} \|e_h(t)\|_{\infty,\Omega_h} = \|e_h(t_n)\|_{\infty,\Omega_h} &\leq \|e_h(t_0)\|_{\infty,\Omega_h} + \int_{t_0}^{t_n} \|v_h(s)\|_{\infty,\Omega_h}\,ds + n\delta\epsilon \\ &= \|e_h(0)\|_{\infty,\Omega_h} + \int_0^t \|v_h(s)\|_{\infty,\Omega_h}\,ds + t\epsilon. \end{aligned}$$

Since this estimate holds for all $\epsilon \in \mathbb{R}_{>0}$, the proof is complete. ■

If we use the Hilbert norm $\|\cdot\|$ instead of the maximum norm, we can obtain a similar result by a particularly simple argument.

Lemma 3.23 (Approximation error) *Let $e_h \in C^1(\mathbb{R}_{\geq 0}, \mathbb{R}_{\geq 0})$ be a solution of (3.13). We have*

$$\|e_h(t)\| \leq \|u_0|_{\bar\Omega_h} - u_{0,h}\| + \int_0^t \|v_h(s)\|\,ds \qquad \textit{for all } t \in \mathbb{R}_{\geq 0}.$$

Proof. We consider the function

$$\gamma\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}, \qquad t \mapsto \|e_h(t)\|.$$

Since the Hilbert norm is given by the inner product, we find

$$\begin{aligned} \gamma(t) &= \|e_h(t)\| = \langle e_h(t), e_h(t)\rangle^{1/2}, \\ \gamma'(t) &= \frac{1}{2}\frac{2\langle e_h'(t), e_h(t)\rangle}{\langle e_h(t), e_h(t)\rangle^{1/2}} = \frac{\langle v_h(t) + \Delta_h e_h(t), e_h(t)\rangle}{\|e_h(t)\|} \qquad \text{for all } t \in \mathbb{R}_{\geq 0} \end{aligned}$$

by the product rule. Using (3.6) and the Cauchy-Schwarz inequality, we find

$$\begin{aligned} \gamma'(t) &= \frac{\langle v_h(t) + \Delta_h e_h(t), e_h(t)\rangle}{\|e_h(t)\|} = \frac{\langle v_h(t), e_h(t)\rangle + \langle \Delta_h e_h(t), e_h(t)\rangle}{\|e_h(t)\|} \\ &\leq \frac{\langle v_h(t), e_h(t)\rangle}{\|e_h(t)\|} \leq \frac{\|v_h(t)\|\,\|e_h(t)\|}{\|e_h(t)\|} = \|v_h(t)\| \qquad \text{for all } t \in \mathbb{R}_{\geq 0}, \end{aligned}$$

and the fundamental theorem of calculus yields

$$\begin{aligned} \gamma(t) &= \gamma(0) + \int_0^t \gamma'(s)\,ds \leq \gamma(0) + \int_0^t \|v_h(s)\|\,ds \\ &= \|u_0|_{\bar\Omega_h} - u_{0,h}\| + \int_0^t \|v_h(s)\|\,ds \qquad \text{for all } t \in \mathbb{R}_{\geq 0}. \end{aligned}$$

■

4. Finite difference methods for hyperbolic equations

Parabolic equations like the heat equation typically show the behaviour outlined in Lemma 3.3: the initial conditions become less and less important as time progresses, and if the driving terms (g in the case of the heat equation) are constant, the solution converges to a limit for $t \to \infty$.

Many processes do not have this property, e.g., electromagnetic waves keep traveling and never reach a steady state. Some of these processes can be described by *hyperbolic* partial differential equations.

4.1. Transport equations

We consider transport equations as a first example. Let $u \in C^1(\mathbb{R} \times \mathbb{R})$ describe a *density* of a fluid, i.e., the amount of fluid in an interval $[a, b] \subseteq \mathbb{R}$ at a time $t \in \mathbb{R}$ is given by

$$m_{a,b}(t) = \int_a^b u(t, x)\,dx \qquad\qquad \text{for all } t \in \mathbb{R}.$$

The transport of fluid is described by a *flux function* $f \in C(\mathbb{R} \times \mathbb{R})$ that assigns each point $x \in \mathbb{R}$ and each time $t \in \mathbb{R}$ the rate at which fluid flows in the positive direction. The amount of fluid in $[a, b]$ changes accordingly, i.e., we have

$$\frac{\partial}{\partial t} m_{a,b}(t) = f(t, a) - f(t, b) \qquad\qquad \text{for all } t \in \mathbb{R}.$$

The change in fluid is the balance between inflow at a and outflow at b.

We would like to obtain an equation for the density, so we have to get rid of the integral in the definition of $m_{a,b}$. By the mean value theorem for integrals, we can find $\eta \in [a, b]$ such that

$$f(t, a) - f(t, b) = \frac{\partial}{\partial t} m_{a,b}(t) = \int_a^b \frac{\partial u}{\partial t}(t, x)\,dx = (b - a)\frac{\partial u}{\partial t}(t, \eta),$$

and dividing by $b - a$ yields

$$-\frac{f(t, b) - f(t, a)}{b - a} = \frac{\partial u}{\partial t}(t, \eta).$$

If f is differentiable, we can consider $a, b \to x$ and obtain

$$-\frac{\partial f}{\partial x}(t, x) = \frac{\partial u}{\partial t}(t, x). \tag{4.1}$$

A standard assumption is that the flux f depends on the density u. A simple example is given by

$$f(t,x) = \alpha u(t,x) \qquad \text{for all } t,x \in \mathbb{R},$$

where $\alpha \in \mathbb{R}$ is a suitable constant. This choice leads to the *linear transport equation*

$$\alpha \frac{\partial u}{\partial x}(t,x) + \frac{\partial u}{\partial t}(t,x) = 0 \qquad \text{for all } t,x \in \mathbb{R}.$$

For a more interesting example, we consider

$$f(t,x) = \frac{1}{2}u(t,x)^2, \qquad \frac{\partial f}{\partial x}(t,x) = u(t,x)\frac{\partial u}{\partial x}(t,x) \qquad \text{for all } t,x \in \mathbb{R}.$$

This leads to the nonlinear *Burgers' equation* [2]

$$u(t,x)\frac{\partial u}{\partial x}(t,x) + \frac{\partial u}{\partial t}(t,x) = 0 \qquad \text{for all } t,x \in \mathbb{R}.$$

4.2. Method of characteristics

For the potential equation (2.1), we could prescribe boundary conditions on the entire boundary $\partial\Omega$ of the computational domain Ω.

For the heat equation (3.1), seen as an equation in the space-time domain $\mathbb{R}_{\geq 0} \times \Omega$ we could prescribe boundary conditions on $\mathbb{R}_{\geq 0} \times \partial\Omega$ and initial conditions at $\{0\} \times \Omega$, but our analysis indicates that these two conditions already lead to a unique solution (at least for the discretized problem).

Now we consider the kind of boundary conditions that can be used for transport equations in the form

$$a(z)\frac{\partial u}{\partial z_1}(z) + b(z)\frac{\partial u}{\partial z_2}(z) = c(z) \qquad \text{for all } z \in \mathbb{R}^2 \tag{4.2}$$

with functions

$$a, b, c : \mathbb{R}^2 \to \mathbb{R}.$$

An elegant approach to analyzing equations of this kind is the *method of characteristics.* We introduce a curve $\gamma \in C^1(\mathbb{R}, \mathbb{R}^2)$ and consider the function $\widehat{u} \in C^1(\mathbb{R})$ given by

$$\widehat{u}(\tau) := u(\gamma(\tau)) \qquad \text{for all } \tau \in \mathbb{R}.$$

Differentiating $\widehat{u}$ using the chain rule yields

$$\widehat{u}'(\tau) = \frac{\partial u}{\partial z_1}(\gamma(\tau))\gamma_1'(\tau) + \frac{\partial u}{\partial z_2}(\gamma(\tau))\gamma_2'(\tau) \qquad \text{for all } \tau \in \mathbb{R},$$

and comparing this equation with (4.2) suggests that we look for a curve γ such that

$$a(\gamma(\tau)) = \gamma_1'(\tau), \qquad b(\gamma(\tau)) = \gamma_2'(\tau) \qquad \text{for all } \tau \in \mathbb{R}.$$

This is a system of ordinary differential equations that can be used to describe the behaviour of the solution.

Definition 4.1 (Characteristic curve) *A function $\gamma \in C^1(\mathbb{R}, \mathbb{R}^2)$ is called a* characteristic curve *of (4.2) if*

$$\gamma'(\tau) = \begin{pmatrix} a(\gamma(\tau)) \\ b(\gamma(\tau)) \end{pmatrix} \qquad \textit{for all } \tau \in \mathbb{R}.$$

If γ is a characteristic curve of (4.2), and if $u \in C^1(\mathbb{R}^2)$ is a solution, the function $\widehat{u} = u \circ \gamma$ satisfies

$$\begin{aligned} \widehat{u}'(\tau) &= \frac{\partial u}{\partial z_1}(\gamma(\tau))\gamma_1'(\tau) + \frac{\partial u}{\partial z_2}(\gamma(\tau))\gamma_2'(\tau) \\ &= \frac{\partial u}{\partial z_1}(\gamma(\tau))a(\gamma(\tau)) + \frac{\partial u}{\partial z_2}(\gamma(\tau))b(\gamma(\tau)) \\ &= c(\gamma(\tau)) \qquad \text{for all } \tau \in \mathbb{R}, \end{aligned}$$

i.e., $\widehat{u}$ is a solution of the initial value problem

$$\widehat{u}(0) = u(\gamma(0)), \qquad \widehat{u}'(\tau) = c(\gamma(\tau)) \qquad \text{for all } \tau \in \mathbb{R}.$$

If this problem has a unique solution, e.g., if c is Lipschitz continuous, we see that if we know $u(\gamma(0))$, the solution u is uniquely determined along the entire curve γ.

This suggests how we may choose boundary conditions: if a characteristic curve intersects the boundary at two or more points, we may only prescribe the value of u in one of these points, since this fixes the values in all others.

For the simple transport equation, we let $z = (t, x)$ and have

$$a(z) = 1, \qquad b(z) = \alpha, \qquad c(z) = 0 \qquad \text{for all } z \in \mathbb{R}^2,$$

so the characteristic curves are given by

$$\gamma(\tau) = \begin{pmatrix} \tau \\ \xi + \alpha\tau \end{pmatrix} \qquad \text{for all } \tau \in \mathbb{R},$$

where $\xi \in \mathbb{R}$ can be chosen to choose the starting point $\gamma(\tau) = (0, \xi)$. Due to $c = 0$, we have

$$u(\gamma(\tau)) = u(0, \xi) \qquad \text{for all } \tau \in \mathbb{R},$$

i.e., u is constant along the characteristic curves.

For $(t, x) = z = \gamma(\tau)$, we have $t = \tau$ and $x = \xi + \alpha t$, so we can write this equation in the form

$$u(t, \xi + \alpha t) = u(0, \xi), \qquad \text{for all } t, \xi \in \mathbb{R},$$

and with $\xi := x - \alpha t$ we get

$$u(t, x) = u(0, x - \alpha t) \qquad \text{for all } t, x \in \mathbb{R}.$$

This means that the solution of the linear transport equation is uniquely determined by the values $u(0, \cdot)$ of u at time $t = 0$.

4.3. One-dimensional wave equation

Another problem that can be tackled with the method of characteristics is the *one-dimensional wave equation*

$$\frac{\partial^2 u}{\partial t^2}(t,x) - c^2 \frac{\partial^2 u}{\partial x^2}(t,x) = 0 \qquad \text{for all } t,x \in \mathbb{R} \tag{4.3}$$

with $c \in \mathbb{R} \setminus \{0\}$. We introduce

$$\gamma\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad \begin{pmatrix} \tau \\ \xi \end{pmatrix} \mapsto \begin{pmatrix} (\tau - \xi)/(2c) \\ (\tau + \xi)/2 \end{pmatrix},$$

and investigate $\widehat{u} := u \circ \gamma$. We have

$$\begin{aligned}
\frac{\partial \widehat{u}}{\partial \tau}(\tau,\xi) &= \frac{\partial u}{\partial t}(\gamma(\tau,\xi))\frac{\partial \gamma_1}{\partial \tau}(\tau,\xi) + \frac{\partial u}{\partial x}(\gamma(\tau,\xi))\frac{\partial \gamma_2}{\partial \tau}(\tau,\xi) \\
&= \frac{1}{2c}\frac{\partial u}{\partial t}(\gamma(\tau,\xi)) + \frac{1}{2}\frac{\partial u}{\partial x}(\gamma(\tau,\xi)), \\
\frac{\partial^2 \widehat{u}}{\partial \tau \partial \xi}(\tau,\xi) &= \frac{1}{2c}\frac{\partial^2 u}{\partial t^2}(\gamma(\tau,\xi))\frac{\partial \gamma_1}{\partial \xi}(\tau,\xi) + \frac{1}{2c}\frac{\partial^2 u}{\partial t \partial x}(\gamma(\tau,\xi))\frac{\partial \gamma_2}{\partial \xi}(\tau,\xi) \\
&\quad + \frac{1}{2}\frac{\partial^2 u}{\partial x \partial t}(\gamma(\tau,\xi))\frac{\partial \gamma_1}{\partial \xi}(\tau,\xi) + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(\gamma(\tau,\xi))\frac{\partial \gamma_2}{\partial \xi}(\tau,\xi) \\
&= -\frac{1}{4c^2}\frac{\partial^2 u}{\partial t^2}(\gamma(\tau,\xi)) + \frac{1}{4c}\frac{\partial^2 u}{\partial t \partial x}(\gamma(\tau,\xi)) - \frac{1}{4c}\frac{\partial^2 u}{\partial t \partial x}(\gamma(\tau,\xi)) + \frac{1}{4}\frac{\partial^2 u}{\partial x^2}(\gamma(\tau,\xi)) \\
&= -\frac{1}{4c^2}\Big(\frac{\partial^2 u}{\partial t^2}(\gamma(\tau,\xi)) - c^2\frac{\partial^2 u}{\partial x^2}(\gamma(\tau,\xi))\Big) = 0 \qquad \text{for all } \tau,\xi \in \mathbb{R}.
\end{aligned}$$

This means that $\partial\widehat{u}/\partial\tau$ is constant with respect to ξ, i.e., that there is a function v_1 such that

$$\frac{\partial \widehat{u}}{\partial \tau}(\tau,\xi) = v_1(\tau) \qquad \text{for all } \tau,\xi \in \mathbb{R},$$

and that $\partial\widehat{u}/\partial\xi$ is constant with respect to τ, i.e., that there is a function v_2 such that

$$\frac{\partial \widehat{u}}{\partial \xi}(\tau,\xi) = v_2(\xi) \qquad \text{for all } \tau,\xi \in \mathbb{R}.$$

Let u_1 and u_2 be antiderivatives of v_1 and v_2, and define

$$\widetilde{u}\colon \mathbb{R}^2 \to \mathbb{R}, \qquad (\tau,\xi) \mapsto u_1(\tau) + u_2(\xi).$$

We find

$$\begin{aligned}
\frac{\partial \widetilde{u}}{\partial \tau}(\tau,\xi) &= u_1'(\tau) = v_1(\tau) = \frac{\partial \widehat{u}}{\partial \tau}(\tau,\xi), \\
\frac{\partial \widetilde{u}}{\partial \xi}(\tau,\xi) &= u_2'(\xi) = v_2(\xi) = \frac{\partial \widehat{u}}{\partial \xi}(\tau,\xi),
\end{aligned}$$

so $\widehat{u}$ and $\widetilde{u}$ can differ only by a constant.

We can add this constant to u_1 or u_2 without changing the relevant equations and obtain

$$\widehat{u}(\tau, \xi) = u_1(\tau) + u_2(\xi) \qquad \text{for all } \tau, \xi \in \mathbb{R}.$$

Let now $t, x \in \mathbb{R}$. We observe

$$\gamma(x + ct, x - ct) = \begin{pmatrix} (x + ct - x + ct)/(2c) \\ (x + ct + x - ct)/2 \end{pmatrix} = \begin{pmatrix} t \\ x \end{pmatrix}$$

and obtain the final result

$$u(t, x) = u_1(x + ct) + u_2(x - ct) \qquad \text{for all } t, x \in \mathbb{R}.$$

4.4. Conservation laws

Hyperbolic partial differential equations are frequently connected to *conservation laws*.

In the case of the transport equation (4.1), the amount $m_{a,b}$ of fluid is conserved: we have

$$\begin{aligned} \frac{\partial}{\partial t} m_{a,b}(t) &= \frac{\partial}{\partial t} \int_a^b u(t, x)\, dx = \int_a^b \frac{\partial}{\partial t} u(t, x)\, dx \\ &= -\int_a^b \frac{\partial}{\partial x} f(t, x)\, dx = f(t, a) - f(t, b) \qquad \text{for all } t \in \mathbb{R}, \end{aligned}$$

i.e., if the flux function were zero, the amount of fluid would be constant.

In the case of the wave equation (4.3), the *energy* of the system is conserved. The energy is defined as the sum of the *kinetic energy*

$$E_{\text{kin}}(t) := \frac{1}{2} \int_{-\infty}^{\infty} \Big(\frac{\partial u}{\partial t}(t, x) \Big)^2 dx \qquad \text{for all } t \in \mathbb{R}$$

and the *potential energy*

$$E_{\text{pot}}(t) := \frac{c^2}{2} \int_{-\infty}^{\infty} \Big(\frac{\partial u}{\partial x}(t, x) \Big)^2 dx \qquad \text{for all } t \in \mathbb{R},$$

assuming that both integrals exist and are bounded. If we also assume

$$\lim_{x \to \infty} \frac{\partial u}{\partial t}(t, x) = 0, \qquad \lim_{x \to -\infty} \frac{\partial u}{\partial t}(t, x) = 0 \qquad \text{for all } t \in \mathbb{R},$$

we can use the product rule and partial integration to find

$$E'_{\text{kin}}(t) = \int_{-\infty}^{\infty} \frac{\partial u}{\partial t}(t, x) \frac{\partial^2 u}{\partial t^2}(t, x)\, dx = \int_{-\infty}^{\infty} \frac{\partial u}{\partial t}(t, x) c^2 \frac{\partial^2 u}{\partial x^2}(t, x)\, dx$$

$$= -c^2 \int_{-\infty}^{\infty} \frac{\partial^2 u}{\partial t \partial x}(t,x) \frac{\partial u}{\partial x}(t,x)\,dx = -E'_{\mathrm{pot}}(t) \qquad \text{for all } t \in \mathbb{R},$$

and therefore the total energy

$$E(t) := E_{\mathrm{kin}}(t) + E_{\mathrm{pot}}(t) \qquad \text{for all } t \in \mathbb{R}$$

satisfies

$$E'(t) = E'_{\mathrm{kin}}(t) + E'_{\mathrm{pot}}(t) = 0 \qquad \text{for all } t \in \mathbb{R},$$

i.e., the total energy is constant.

Convervation laws play an important role in many applications, and numerical algorithms for solving the corresponding partial differential equations should at least try to ensure that the quantities that are conserved in the continuous equation are also conserved in the discretized equation, at least approximately.

4.5. Higher-dimensional wave equation

We consider the two-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2}(t,x) - c^2 \Delta_x u(t,x) = 0 \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \Omega$$

in a domain $\Omega \subseteq \mathbb{R}^2$ with a parameter $c \in \mathbb{R}_{>0}$.

In order to obtain a unique solution, we impose Dirichlet boundary conditions

$$u(t,x) = 0 \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \partial\Omega.$$

We can eliminate the second time derivatives by introducing the velocity function

$$v(t,x) := \frac{\partial u}{\partial t}(t,x) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \bar{\Omega}$$

and writing the wave equation in the form

$$\frac{\partial u}{\partial t}(t,x) = v(t,x), \qquad \frac{\partial v}{\partial t}(t,x) = c^2 \Delta_x u(t,x) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \Omega, \tag{4.4a}$$

$$u(t,x) = 0, \qquad v(t,x) = 0 \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \partial\Omega. \tag{4.4b}$$

As in the case of the heat equation (3.1), these equations can be considered as an ordinary differential equation for

$$y(t) := \begin{pmatrix} u(t,\cdot) \\ v(t,\cdot) \end{pmatrix} \in C_0^\infty(\Omega) \times C_0^\infty(\Omega) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},$$

where

$$C_0^\infty(\Omega) := \{u \in C(\bar{\Omega}) \ :\ u|_\Omega \in C^\infty(\Omega),\ u|_{\partial\Omega} = 0\}$$

is the space of infinitely differentiable functions with homogeneous Dirichlet boundary conditions. The equations (4.4) correspond to the ordinary differential equation

$$y'(t) = \begin{pmatrix} y_2(t) \\ c^2 \Delta y_1(t) \end{pmatrix} \qquad \text{for all } t \in \mathbb{R}_{\geq 0},$$

so we can expect that we have to introduce initial conditions

$$y(0) = y_0 = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}$$

with $u_0, v_0 \in C_0^\infty(\Omega)$ at the time $t = 0$ to ensure that the initial value problem can have at most one solution.

As in the one-dimensional case, the wave equation (4.4) conserves the total energy. For the kinetic energy, we can essentially use the same definition as in the one-dimensional setting. For the potential energy, we have to introduce two differential operators.

Definition 4.2 (Gradient and divergence) *Let $d \in \mathbb{N}$, let $\Omega \subseteq \mathbb{R}^d$ be a domain, let $\varphi \in C^1(\Omega)$. The mapping*

$$\nabla\varphi \colon \Omega \to \mathbb{R}^d, \qquad x \mapsto \begin{pmatrix} \frac{\partial \varphi}{\partial x_1}(x) \\ \vdots \\ \frac{\partial \varphi}{\partial x_d}(x) \end{pmatrix},$$

is called the gradient *of φ. Let $u \in C^1(\Omega, \mathbb{R}^d)$. The mapping*

$$\nabla \cdot u \colon \Omega \to \mathbb{R}, \qquad x \mapsto \frac{\partial u_1}{\partial x_1}(x) + \ldots + \frac{\partial u_d}{\partial x_d}(x),$$

is called the divergence *of u.*

Reminder 4.3 (Gauß integral theorem) *Let $d \in \mathbb{N}$, let $\Omega \subseteq \mathbb{R}^d$ be a Lipschitz domain, let*

$$n : \partial\Omega \to \mathbb{R}^d$$

denote the mapping that assigns all boundary points $x \in \partial\Omega$ the unit exterior normal vector $n(x)$.

We have

$$\int_{\partial\Omega} \langle n(x), u(x) \rangle_2 \, dx = \int_\Omega \nabla \cdot u(x) \, dx \qquad \text{for all } u \in C^1(\Omega, \mathbb{R}^d).$$

Applying this result to $u := \varphi v$ for $\varphi \in C^1(\Omega)$ and $v \in C^1(\Omega, \mathbb{R}^d)$ using the product rule yields the equation

$$\int_{\partial\Omega} \varphi(x) \langle n(x), v(x) \rangle_2 \, dx = \int_\Omega \varphi(x) \nabla \cdot v(x) \, dx + \int_\Omega \langle \nabla\varphi(x), v(x) \rangle_2 \, dx$$

corresponding to multi-dimensional partial integration.

For the multi-dimensional wave equation (4.4), we define the kinetic energy by

$$E_{\text{kin}}(t) := \frac{1}{2}\int_\Omega \left(\frac{\partial u}{\partial t}(t,x)\right)^2 dx = \frac{1}{2}\int_\Omega v(x)^2\,dx \qquad \text{for all } t \in \mathbb{R}_{\geq 0}$$

and the potential energy by

$$E_{\text{pot}}(t) := \frac{c^2}{2}\int_\Omega \|\nabla u(t,x)\|_2^2\,dx = \frac{c^2}{2}\int_\Omega \langle \nabla u(t,x), \nabla u(t,x)\rangle_2\,dx \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

Corollary 4.4 (Energy conservation) *Let $u, v \in C^1(\mathbb{R}_{\geq 0}, C_0^\infty(\Omega))$ solve the wave equation (4.4). We have*

$$E'_{kin}(t) + E'_{pot}(t) = 0 \qquad \textit{for all } t \in \mathbb{R}_{\geq 0},$$

i.e., the total energy is constant.

Proof. Let $t \in \mathbb{R}_{\geq 0}$. Using the product rule and multi-dimensional partial integration, we find

$$\begin{aligned}
E'_{\text{kin}}(t) &= \int_\Omega \frac{\partial u}{\partial t}(t,x)\frac{\partial^2 u}{\partial t^2}(t,x)\,dx = \int_\Omega \frac{\partial u}{\partial t}(t,x)c^2\Delta_x u(t,x)\,dx \\
&= \int_\Omega \frac{\partial u}{\partial t}(t,x)c^2\nabla\cdot(\nabla u)(t,x)\,dx \\
&= c^2\int_{\partial\Omega} \frac{\partial u}{\partial t}(t,x)\langle n(x), \nabla u(t,x)\rangle_2\,dx - c^2\int_\Omega \langle \nabla\frac{\partial u}{\partial t}(t,x), \nabla u(t,x)\rangle_2\,dx \\
&= -c^2\int_\Omega \langle \frac{\partial}{\partial t}\nabla u(t,x), \nabla u(t,x)\rangle_2\,dx = -E'_{\text{pot}}(t).
\end{aligned}$$

■

4.6. Method of lines

As in the case of parabolic equations, we can employ the method of lines to approximate the solution of the wave equation (4.4): we replace $u(t,\cdot)$ by a grid function $u_h(t) \in G_0(\bar\Omega_h)$ and $v(t,\cdot)$ by a grid function $v_h(t) \in G_0(\bar\Omega_h)$, while the Laplace operator Δ is approximated by the finite difference operator Δ_h. This results in the following system of ordinary differential equations:

$$u'_h(t) = v_h(t), \qquad v'_h(t) = c^2\Delta_h u_h(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}, \tag{4.5a}$$

$$u_h(0) = u_{0,h}, \qquad v_h(0) = v_{0,h}. \tag{4.5b}$$

This system can be solve by time-stepping methods.

We introduce the Hilbert space

$$\mathcal{V} := G_0(\bar\Omega_h) \times G_0(\bar\Omega_h),$$

for the moment with the inner product

$$\langle (u_h, v_h), (x_h, y_h) \rangle := \langle u_h, x_h \rangle_{\Omega_h} + \langle v_h, y_h \rangle_{\Omega_h} \qquad \text{for all } (u_h, v_h), (x_h, y_h) \in \mathcal{V},$$

and let

$$\begin{aligned} y_0 &:= \begin{pmatrix} u_{0,h} \\ v_{0,h} \end{pmatrix}, \\ y(t) &:= \begin{pmatrix} u_h(t) \\ v_h(t) \end{pmatrix} && \text{for all } t \in \mathbb{R}_{\geq 0}, \\ f(t, (u_h, v_h)) &:= \begin{pmatrix} v_h \\ c^2 \Delta_h u_h \end{pmatrix} && \text{for all } t \in \mathbb{R}_{\geq 0},\ (u_h, v_h) \in \mathcal{V}. \end{aligned}$$

This allows us to write (4.5) in the usual form

$$y(0) = y_0, \qquad y'(t) = f(t, y(t)) \qquad \text{for all } t \in \mathbb{R}_{\geq 0}. \tag{4.6}$$

The explicit Euler method takes the form

$$\begin{aligned} \widetilde{u}_h(t_0) &= u_{0,h}, & \widetilde{v}_h(t_0) &= v_{0,h}, \\ \widetilde{u}_h(t_{i+1}) &= \widetilde{u}_h(t_i) + \delta \widetilde{v}_h(t_i), & \widetilde{v}_h(t_{i+1}) &= \widetilde{v}_h(t_i) + \delta c^2 \Delta_h \widetilde{u}_h(t_i) && \text{for all } i \in \mathbb{N}_0, \end{aligned}$$

for the implicit Euler method we find

$$\begin{aligned} \widetilde{u}_h(t_0) &= u_{0,h}, & \widetilde{v}_h(t_0) &= v_{0,h}, \\ \widetilde{u}_h(t_{i+1}) &= \widetilde{u}_h(t_i) + \delta \widetilde{v}_h(t_{i+1}), & \widetilde{v}_h(t_{i+1}) &= \widetilde{v}_h(t_i) + \delta c^2 \Delta_h \widetilde{u}_h(t_{i+1}) && \text{for all } i \in \mathbb{N}_0, \end{aligned}$$

and substituting the variables yields

$$\begin{aligned} \widetilde{u}_h(t_{i+1}) &= \widetilde{u}_h(t_i) + \delta(\widetilde{v}_h(t_i) + \delta c^2 \Delta_h \widetilde{u}_h(t_{i+1})), \\ (I - \delta^2 c^2 \Delta_h) \widetilde{u}_h(t_{i+1}) &= \widetilde{u}_h(t_i) + \delta \widetilde{v}_h(t_i), \\ \widetilde{v}_h(t_{i+1}) &= \widetilde{v}_h(t_i) + \delta c^2 \Delta_h (\widetilde{u}_h(t_i) + \delta \widetilde{v}_h(t_{i+1})), \\ (I - \delta^2 c^2 \Delta_h) \widetilde{v}_h(t_{i+1}) &= \widetilde{v}_h(t_i) + \delta c^2 \Delta_h \widetilde{u}_h(t_i) && \text{for all } i \in \mathbb{N}_0, \end{aligned}$$

so performing one time step requires us to solve two linear systems.

For the Crank-Nicolson method, we obtain

$$\begin{aligned} \widetilde{u}_h(t_0) &= u_{0,h}, \qquad \widetilde{v}_h(t_0) = v_{0,h}, \\ \widetilde{u}_h(t_{i+1}) &= \widetilde{u}_h(t_i) + \frac{\delta}{2}(\widetilde{v}_h(t_i) + \widetilde{v}_h(t_{i+1})), \\ \widetilde{v}_h(t_{i+1}) &= \widetilde{v}_h(t_i) + \frac{\delta}{2} c^2 \Delta_h (\widetilde{u}_h(t_i) + \widetilde{u}_h(t_{i+1})) && \text{for all } i \in \mathbb{N}_0, \end{aligned}$$

and once more substitution yields

$$\widetilde{u}_h(t_{i+1}) = \widetilde{u}_h(t_i) + \frac{\delta}{2}\Big(\widetilde{v}_h(t_i) + \widetilde{v}_h(t_i) + \frac{\delta}{2} c^2 \Delta_h (\widetilde{u}_h(t_i) + \widetilde{u}_h(t_{i+1}))\Big),$$

$$\begin{aligned}\Big(I - \frac{\delta^2}{4} c^2 \Delta_h\Big) \widetilde{u}_h(t_{i+1}) &= \Big(I + \frac{\delta^2}{4} c^2 \Delta_h\Big) \widetilde{u}_h(t_i) + \delta \widetilde{v}_h(t_i), \\ \widetilde{v}_h(t_{i+1}) &= \widetilde{v}_h(t_i) + \frac{\delta}{2} c^2 \Delta_h \Big(\widetilde{u}_h(t_i) + \widetilde{u}_h(t_i) + \frac{\delta}{2} (\widetilde{v}_h(t_i) + \widetilde{v}_h(t_{i+1})) \Big), \\ \Big(I - \frac{\delta^2}{4} c^2 \Delta_h\Big) \widetilde{v}_h(t_{i+1}) &= \Big(I + \frac{\delta^2}{4} c^2 \Delta_h\Big) \widetilde{v}_h(t_i) + \delta c^2 \Delta_h \widetilde{u}_h(t_i),\end{aligned}$$

so we can perform one time step by solving two linear systems.

In order to prove convergence, we have to establish that our time-stepping algorithms are consistent and stable. Since the original wave equation conserves the total energy, we should also consider whether the discrete solution shares this property.

4.7. Discrete conservation of energy

Since the potential energy of a solution (u, v) of the original equation (4.4) can be written in the form

$$\begin{aligned}E_{\mathrm{pot}}(t) &= \frac{c^2}{2} \int_\Omega \langle \nabla u(t,x), \nabla u(t,x) \rangle_2 \, dx \\ &= \frac{c^2}{2} \int_{\partial\Omega} u(t,x) \langle n(x), \nabla u(t,x) \rangle_2 \, dx - \frac{c^2}{2} \int_\Omega u(t,x) \nabla \cdot \nabla u(t,x) \, dx \\ &= -\frac{c^2}{2} \int_\Omega u(t,x) \Delta_x u(t,x) \, dx \qquad \text{for all } t \in \mathbb{R}_{\geq 0}\end{aligned}$$

by partial integration (cf. Reminder 4.3) due to $u(t,x)|_{\partial\Omega} = 0$, a straightforward definition of the potential energy of the discrete problem is given by

$$E_{\mathrm{pot},h}(t) := -\frac{c^2}{2} \langle u_h(t), \Delta_h u_h(t) \rangle_{\Omega_h} \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

For the kinetic energy, we choose

$$E_{\mathrm{kin},h}(t) := \frac{1}{2} \langle v_h(t), v_h(t) \rangle_{\Omega_h} \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

The product rule immediately yields

$$\begin{aligned}E'_{\mathrm{kin},h}(t) &= \langle v_h(t), v'_h(t) \rangle_{\Omega_h} = \langle v_h(t), c^2 \Delta_h u_h(t) \rangle_{\Omega_h} = \langle u'_h(t), c^2 \Delta_h u_h(t) \rangle_{\Omega_h} \\ &= -E'_{\mathrm{pot},h}(t) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\end{aligned}$$

where the last step makes use of the fact that Δ_h is a self-adjoint operator. We can see that the method of lines preserves the *discrete energy*

$$E_h(t) := E_{\mathrm{kin},h}(t) + E_{\mathrm{pot},h}(t) = \frac{1}{2} \|v_h(t)\|^2_{\Omega_h} - \frac{c^2}{2} \langle u_h(t), \Delta_h u_h(t) \rangle_{\Omega_h} \quad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

Obviously, we prefer time-stepping schemes that share this property.

In order to generalize the following results, we introduce the operator

$$\mathcal{L}_h := -c^2\Delta_h$$

and write our system in the form

$$\begin{pmatrix} u_h(0) \\ v_h(0) \end{pmatrix} = \begin{pmatrix} u_{0,h} \\ v_{0,h} \end{pmatrix}, \qquad \begin{pmatrix} u_h'(t) \\ v_h'(t) \end{pmatrix} = \begin{pmatrix} 0 & I \\ -\mathcal{L}_h & 0 \end{pmatrix} \begin{pmatrix} u_h(t) \\ v_h(t) \end{pmatrix} \qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

In the following, we only assume that $\mathcal{L}_h$ is *positive definite*, i.e., that

$$\langle v_h, \mathcal{L}_h v_h \rangle_{\Omega_h} > 0 \qquad\qquad \text{for all } v_h \in G_0(\bar{\Omega}_h) \setminus \{0\}.$$

For our model problem (4.5), this property is guaranteed by Lemma 3.2.

The energy of a state $(u_h(t), v_h(t))$ can be written as $E_h(t) = \frac{1}{2}\Phi_h(u_h(t), v_h(t))$, where

$$\Phi_h \colon G_0(\bar{\Omega}_h) \times G_0(\bar{\Omega}_h) \to \mathbb{R}_{\geq 0}, \qquad (x_h, y_h) \mapsto \|y_h\|_{\Omega_h}^2 + \langle x_h, \mathcal{L}_h x_h \rangle_{\Omega_h},$$

is called the *discrete energy functional.*

Lemma 4.5 (Explicit Euler method) *Let $\delta \in \mathbb{R}_{\geq 0}$. Let $u_h, v_h \in G_0(\bar{\Omega}_h)$, and let*

$$\widetilde{u}_h := u_h + \delta v_h, \qquad\qquad \widetilde{v}_h := v_h - \delta \mathcal{L}_h u_h$$

denote the approximations constructed in one step of the explicit Euler method. We have

$$\Phi_h(\widetilde{u}_h, \widetilde{v}_h) - \Phi_h(u_h, v_h) = \delta^2 \|\mathcal{L}_h u_h\|_{\Omega_h}^2 + \delta^2 \langle v_h, \mathcal{L}_h v_h \rangle_{\Omega_h} \geq 0.$$

Proof. By the third binomial equation, we have

$$\begin{aligned} \|\widetilde{v}_h\|_{\Omega_h}^2 - \|v_h\|_{\Omega_h}^2 &= \langle \widetilde{v}_h - v_h, \widetilde{v}_h + v_h \rangle_{\Omega_h} = -\delta \langle \mathcal{L}_h u_h, \widetilde{v}_h + v_h \rangle_{\Omega_h}, \\ \langle \widetilde{u}_h, \mathcal{L}_h \widetilde{u}_h \rangle_{\Omega_h} - \langle u_h, \mathcal{L}_h u_h \rangle_{\Omega_h} &= \langle \widetilde{u}_h - u_h, \mathcal{L}_h(\widetilde{u}_h + u_h) \rangle_{\Omega_h} = \delta \langle v_h, \mathcal{L}_h(\widetilde{u}_h + u_h) \rangle_{\Omega_h}, \end{aligned}$$

so we find

$$\begin{aligned} \Phi_h(\widetilde{u}_h, \widetilde{v}_h) - \Phi_h(u_h, v_h) &= \delta \langle v_h, \mathcal{L}_h(\widetilde{u}_h + u_h) \rangle_{\Omega_h} - \delta \langle \widetilde{v}_h + v_h, \mathcal{L}_h u_h \rangle_{\Omega_h} \\ &= \delta \langle v_h, \mathcal{L}_h \widetilde{u}_h \rangle_{\Omega_h} - \delta \langle \widetilde{v}_h, \mathcal{L}_h u_h \rangle_{\Omega_h} \\ &= \delta \langle v_h, \mathcal{L}_h u_h + \delta \mathcal{L}_h v_h \rangle_{\Omega_h} - \delta \langle v_h - \delta \mathcal{L}_h u_h, \mathcal{L}_h u_h \rangle_{\Omega_h} \\ &= \delta^2 \langle v_h, \mathcal{L}_h v_h \rangle_{\Omega_h} + \delta^2 \langle \mathcal{L}_h u_h, \mathcal{L}_h u_h \rangle_{\Omega_h} \geq 0. \end{aligned}$$

■

Due to our assumption, we know that $\langle v_h, \mathcal{L}_h v_h \rangle_{\Omega_h} \geq 0$ holds for all $v_h \in G_0(\bar{\Omega}_h)$ with $\langle v_h, \Delta_h v_h \rangle_{\Omega_h} = 0$ if and only if $v_h = 0$, so we have to conclude that the explicit Euler scheme increases the total discrete energy unless we encounter the *very* special case $u_h = v_h = 0$.

Exercise 4.6 (Implicit Euler method) *Let $\delta \in \mathbb{R}_{\geq 0}$. Let $u_h, v_h \in G_0(\bar{\Omega}_h)$, and let*

$$\widetilde{u}_h = u_h + \delta \widetilde{v}_h, \qquad\qquad \widetilde{v}_h = v_h - \delta \mathcal{L}_h \widetilde{u}_h$$

denote the approximations constructed in one step of the implicit Euler method. Prove

$$\Phi_h(\widetilde{u}_h, \widetilde{v}_h) - \Phi_h(u_h, v_h) = -\delta^2 \|\mathcal{L}_h \widetilde{u}_h\|_{\Omega_h}^2 - \delta^2 \langle \widetilde{v}_h, \mathcal{L}_h \widetilde{v}_h \rangle_{\Omega_h} \leq 0.$$

The implicit Euler method decreases the total discrete energy, and this is also not a desirable property.

Lemma 4.7 (Crank-Nicolson method) *Let $\delta \in \mathbb{R}_{\geq 0}$. Let $u_h, v_h \in G_0(\bar{\Omega}_h)$, and let*

$$\widetilde{u}_h := u_h + \frac{\delta}{2}(\widetilde{v}_h + v_h), \qquad\qquad \widetilde{v}_h := v_h - \frac{\delta}{2}\mathcal{L}_h(\widetilde{u}_h + u_h)$$

denote the approximations constructed in one step of the Crank-Nicolson method. We have

$$\Phi_h(\widetilde{u}_h, \widetilde{v}_h) = \Phi_h(u_h, v_h).$$

Proof. By the third binomial equation, we have

$$\begin{aligned}
\|\widetilde{v}_h\|_{\Omega_h}^2 - \|v_h\|_{\Omega_h}^2 &= \langle \widetilde{v}_h - v_h, \widetilde{v}_h + v_h \rangle_{\Omega_h} = -\frac{\delta}{2} \langle \mathcal{L}_h(\widetilde{u}_h + u_h), \widetilde{v}_h + v_h \rangle_{\Omega_h}, \\
\langle \widetilde{u}_h, \mathcal{L}_h \widetilde{u}_h \rangle_{\Omega_h} - \langle u_h, \mathcal{L}_h u_h \rangle_{\Omega_h} &= \langle \widetilde{u}_h - u_h, \mathcal{L}_h(\widetilde{u}_h + u_h) \rangle_{\Omega_h} \\
&= \frac{\delta}{2} \langle \widetilde{v}_h + v_h, \mathcal{L}_h(\widetilde{u}_h + u_h) \rangle_{\Omega_h},
\end{aligned}$$

so we find

$$\begin{aligned}
\Phi_h(\widetilde{u}_h, \widetilde{v}_h) - \Phi_h(u_h, v_h) = \frac{\delta}{2}\big(&- \langle \widetilde{v}_h + v_h, \mathcal{L}_h(\widetilde{u}_h + u_h) \rangle_{\Omega_h} \\
&+ \langle \widetilde{v}_h + v_h, \mathcal{L}_h(\widetilde{u}_h + u_h) \rangle_{\Omega_h} \big) = 0.
\end{aligned}$$

∎

This is a very encouraging result: the Crank-Nicolson method conserves the total discrete energy, just like the original wave equation conserves the total energy.

4.8. Consistency and stability

In order to prove convergence of a time-stepping method used to approximate the solution of (4.5), we should try to establish consistency and stability of the method for the given problem.

We would like to re-use the previous results for the parabolic case, and these results rely on the property

$$\langle f(t,x) - f(t,y), x - y \rangle \leq 0 \qquad\qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x, y \in \mathcal{V}.$$

In our case, we have $\mathcal{V} = G_0(\bar{\Omega}_h) \times G_0(\bar{\Omega}_h)$, and the obvious candidate

$$\langle \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \rangle := \langle x_1, y_1 \rangle_{\Omega_h} + \langle x_2, y_2 \rangle_{\Omega_h} \qquad \text{for all } x, y \in \mathcal{V}$$

for an inner product for $\mathcal{V}$ would lead to

$$\begin{aligned} \langle f(t,x) - f(t,y), x - y \rangle = \langle x_2 - y_2, x_1 - y_1 \rangle_{\Omega_h} & \\ - \langle \mathcal{L}_h(x_1 - y_1), x_2 - y_2 \rangle_{\Omega_h} & \qquad \text{for all } x, y \in \mathcal{V}, \end{aligned}$$

and it is not clear at all why this term should be non-positive.

A very elegant approach relies on the *energy inner product*

$$\langle \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \rangle_A := \langle x_1, \mathcal{L}_h y_1 \rangle_{\Omega_h} + \langle x_2, y_2 \rangle_{\Omega_h} \qquad \text{for all } x, y \in \mathcal{V}.$$

This inner product gets its name from the fact that

$$\Phi_h(x) = \langle x, x \rangle_A \qquad \text{for all } x \in \mathcal{V},$$

i.e., the *energy norm* corresponding to the energy inner product, defined by

$$\|x\|_A := \sqrt{\langle x, x \rangle_A} \qquad \text{for all } x \in \mathcal{V},$$

is just the square of the energy functional.

For the energy inner product, we find

$$\begin{aligned} \langle f(t,x) - f(t,y), x - y \rangle_A = \langle x_2 - y_2, \mathcal{L}_h(x_1 - y_1) \rangle_{\Omega_h} & \\ + \langle -\mathcal{L}(x_1 - y_1), x_2 - y_2 \rangle_{\Omega_h} = 0 & \quad \text{for all } t \in \mathbb{R}_{\geq 0},\ x, y \in \mathcal{V}, \end{aligned}$$

and consistency of our time-stepping methods is guaranteed by Lemma 3.8, Lemma 3.9, and Lemma 3.10.

Stability is guaranteed by Lemma 3.15 for the implicit Euler method and by Exercise 3.19 for the Crank-Nicolson method.

For the explicit Euler method, we can take a look at the eigenvalues.

Lemma 4.8 (Explicit Euler, wave equation) *Let Ψ denote the time-step function of the explicit Euler method for our model problem (4.6), and let*

$$C_\Psi := \sqrt{1 + \delta^2 c^2 \lambda_{max}} \leq 1 + \delta c \sqrt{\lambda_{max}}.$$

We have

$$\|\Psi(t,\delta,x) - \Psi(t,\delta,y)\|_A \leq C_\Psi \|x - y\|_A \qquad \textit{for all } t \in \mathbb{R}_{\geq 0},\ \delta \in \mathbb{R}_{\geq 0},\ x, y \in \mathcal{V}.$$

Proof. Let $t \in \mathbb{R}_{\geq 0}$, $\delta \in \mathbb{R}_{\geq 0}$, and $x, y \in \mathcal{V}$. We let

$$\widetilde{x} := \Psi(t,\delta,x) = \begin{pmatrix} x_1 + \delta x_2 \\ x_2 - \delta \mathcal{L}_h x_1 \end{pmatrix},$$
$$\widetilde{y} := \Psi(t,\delta,y) = \begin{pmatrix} y_1 + \delta y_2 \\ y_2 - \delta \mathcal{L}_h y_1 \end{pmatrix}.$$

Due to Lemma 3.2, we can find $(\alpha_\nu)_{\nu\in[1:N]^2}$ and $(\beta_\nu)_{\nu\in[1:N]^2}$ such that

$$x_1 - y_1 = \sum_{\nu\in[1:N]^2} \alpha_\nu e_{h,\nu}, \qquad x_2 - y_2 = \sum_{\nu\in[1:N]^2} \beta_\nu e_{h,\nu}.$$

We obtain

$$\begin{aligned}
\widetilde{x}_1 - \widetilde{y}_1 &= (x_1 - y_1) + \delta(x_2 - y_2) = \sum_{\nu\in[1:N]^2} (\alpha_\nu + \delta\beta_\nu) e_{h,\nu}, \\
\widetilde{x}_2 - \widetilde{y}_2 &= (x_2 - y_2) - \delta\mathcal{L}_h(x_1 - y_1) = \sum_{\nu\in[1:N]^2} (\beta_\nu - \delta c^2 \lambda_{h,\nu}\alpha_\nu) e_{h,\nu},
\end{aligned}$$

$$\begin{aligned}
\langle \widetilde{x}_1 - \widetilde{y}_1, \mathcal{L}_h(\widetilde{x}_1 - \widetilde{y}_1)\rangle_{\Omega_h} &= \sum_{\nu\in[1:N]^2} c^2\lambda_{h,\nu}(\alpha_\nu + \delta\beta_\nu)^2, \\
\|\widetilde{x}_2 - \widetilde{y}_2\|_{\Omega_h}^2 &= \sum_{\nu\in[1:N]^2} (\beta_\nu - \delta c^2\lambda_{h,\nu}\alpha_\nu)^2, \\
\|\widetilde{x} - \widetilde{y}\|_A^2 &= \sum_{\nu\in[1:N]^2} c^2\lambda_{h,\nu}(\alpha_\nu + \delta\beta_\nu)^2 + (\beta_\nu - \delta c^2\lambda_{h,\nu}\alpha_\nu)^2 \\
&= \sum_{\nu\in[1:N]^2} (c^2\lambda_{h,\nu}\alpha_\nu^2 + 2c^2\delta\lambda_{h,\nu}\alpha_\nu\beta_\nu + c^2\delta^2\lambda_{h,\nu}\beta_\nu^2 \\
&\qquad + \beta_\nu^2 - 2c^2\delta\lambda_{h,\nu}\alpha_\nu\beta_\nu + c^4\delta^2\lambda_{h,\nu}^2\alpha_\nu^2) \\
&= \sum_{\nu\in[1:N]^2} (1 + \delta^2c^2\lambda_{h,\nu})c^2\lambda_{h,\nu}\alpha_\nu^2 + (1 + c^2\delta^2\lambda_{h,\nu})\beta_\nu^2 \\
&\leq C_\Psi^2 \sum_{\nu\in[1:N]^2} c^2\lambda_{h,\nu}\alpha_\nu^2 + \beta_\nu^2 \\
&= C_\Psi^2(\|x_2 - y_2\|_{\Omega_h}^2 + \langle x_1 - y_1, \mathcal{L}_h(x_1 - y_1)\rangle_{\Omega_h}) \\
&= C_\Psi^2 \|x - y\|_A^2.
\end{aligned}$$

This is the first estimate. We conclude by observing

$$\sqrt{1 + \delta^2c^2\lambda_{\max}} \leq \sqrt{1 + 2\delta c\sqrt{\lambda_{\max}} + \delta^2c^2\lambda_{\max}} = \sqrt{(1 + \delta c\sqrt{\lambda_{\max}})^2} = 1 + \delta c\sqrt{\lambda_{\max}}.$$

■

We can see that this estimate cannot be improved, since we have equality for $x - y = (e_{h,(N,N)}, e_{h,(N,N)})$. This means that we can only expect a stable method if we ensure

$$\delta^2 \lesssim \frac{1}{c^2\lambda_{\max}} \approx \frac{h^2}{8c^2}, \qquad \delta \lesssim h.$$

This is similar to the Courant-Friedrichs-Lewy condition (3.11) for the heat equation: explicit time-stepping schemes for the wave equation also require that the time steps become smaller as the grid is refined.

In a way, the wave equation is less demanding than the heat equation: while the heat equation requires $\delta \in \mathcal{O}(h^2)$, we only need $\delta \in \mathcal{O}(h)$ for the wave equation.

As in the case of the heat equation, both implicit methods are unconditionally stable and therefore require no bound for the time steps.

4.9. Finite volume discretization

Conservation laws are frequently expressed in terms of integrals, and this gives rise to an important class of discretization techniques: finite volume methods split the computational domain into subsets and formulate conditions that have to be satisfied in each of these subsets.

A simple example is Darcy's model of groundwater flow, described by two quantities.

The *flux* $f\colon \Omega \to \mathbb{R}^2$ corresponds to the flow of water in the domain $\Omega \subseteq \mathbb{R}^2$. Roughly speaking, the inner product $\langle q, f(x)\rangle$ describes the amount of water flowing in direction $q \in \mathbb{R}^2$ in point $x \in \Omega$.

The *pressure* $p\colon \Omega \to \mathbb{R}$ corresponds to the force exerted, e.g., by gravitation.

Darcy's law states

$$f(x) + k\nabla p(x) = 0 \qquad\qquad \text{for all } x \in \Omega, \tag{4.7}$$

i.e., groundwater flows from high-pressure into low-pressure regions. The *permeability* $k \in \mathbb{R}_{>0}$ describes how rapidly the water can flow in response to the pressure.

In order to obtain a reasonable model, we have to add a second set of equations describing the *conservation of mass*, i.e., that water is not created or destroyed. This property is described by the equation

$$\nabla \cdot f(x) = 0 \qquad\qquad \text{for all } x \in \Omega. \tag{4.8}$$

Given a subdomain $\omega \subseteq \Omega$ with exterior normal vectors $n\colon \partial\omega \to \mathbb{R}^2$, the Gauß theorem 4.3 yields

$$0 = \int_\omega \nabla \cdot f(x)\,dx = \int_{\partial\omega} \langle n(x), f(x)\rangle\,dx,$$

i.e., the total flows into and out of ω are balanced and therefore the total amount of water is conserved.

The idea of the finite volume method is to split the domain into a finite number of subdomains and formulate equations that have to hold for each of these domains. We once again consider only the unit square $\Omega = (0,1) \times (0,1)$, choose $n \in \mathbb{N}$, let $h := 1/n$, and define sub-squares

$$\omega_i := [(i_1 - 1)h, i_1 h] \times [(i_2 - 1)h, i_2 h] \qquad\qquad \text{for all } i \in \mathcal{I} := [1:n] \times [1:n].$$

Applying the Gauß theorem to these squares yields

$$0 = \int_{\omega_i} \nabla \cdot f(x)\,dx = \int_{\partial\omega_i} \langle n(x), f(x)\rangle\,dx \qquad \text{for all } i \in \mathcal{I}.$$

The boundaries of the squares consist of edges

$$e_{x,i} := [(i_1-1)h, i_1 h] \times \{i_2 h\} \qquad \text{for all } i \in \bar{\mathcal{I}}_x := [1:n] \times [0:n]$$

in x direction and edges

$$e_{y,i} := \{i_1 h\} \times [(i_2-1)h, i_2 h] \qquad \text{for all } i \in \bar{\mathcal{I}}_y := [0:n] \times [1:n]$$

in y direction. For these edges we fix the unit normal vectors

$$n_x := \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad n_y := \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

For $i \in \mathcal{I}$, the boundary of ω_i consists of the edges $e_{x,i}$, e_{x,i_1,i_2-1}, $e_{y,i}$, and e_{y,i_1-1,i_2}. With respect to this squarem the vector n_x is an exterior normal vector on $e_{x,i}$ and an interior normal vector on e_{x,i_1,i_2-1}, while n_y is an exterior normal vector on $e_{y,i}$ and an interior normal vector on e_{y,i_1-1,i_2}, so that the boundary integral takes the form

$$\begin{aligned} 0 &= \int_{\partial\omega_i} \langle n(x), f(x)\,dx \\ &= \int_{e_{x,i}} f_2(x)\,dx - \int_{e_{x,i_1,i_2-1}} f_2(x)\,dx + \int_{e_{y,i}} f_1(x)\,dx - \int_{e_{y,i_1-1,i_2}} f_1(x)\,dx. \end{aligned}$$

We use these edge integrals as the first set of degrees of freedom in the discrete system, i.e., we let

$$\begin{aligned} f_{x,i} &:= \int_{e_{x,i}} f_2(x)\,dx && \text{for all } i \in \bar{\mathcal{I}}_x, \\ f_{y,i} &:= \int_{e_{y,i}} f_1(x)\,dx && \text{for all } i \in \bar{\mathcal{I}}_y, \end{aligned}$$

and observe that the *exact* conservation of mass in each square corresponds to the equations

$$f_{x,i} - f_{x,i_1,i_2-1} + f_{y,i} - f_{y,i_1-1,i_2} = 0 \qquad \text{for all } i \in \mathcal{I}. \tag{4.9}$$

Now we have to consider Darcy's law. Let $i \in [1:n] \times [1:n-1]$. Multiplying Darcy's law (4.7) by the normal vector and integrating along the edge $e_{x,i}$ yields

$$\begin{aligned} 0 = \int_{e_{x,i}} \langle n_x, f(x) + k\nabla p(x)\rangle\,dx &= \int_{e_{x,i}} \langle n_x, f(x)\rangle\,dx + k\int_{e_{x,i}} \frac{\partial p}{\partial x_2}(x)\,dx \\ = f_{x,i} + k\int_{(i_1-1)h}^{i_1 h} \frac{\partial p}{\partial x_2}(s, i_2 h)\,ds. \end{aligned}$$

Since we cannot handle equations of this kind directly, we have to employ a numerical approximation. For the derivative, we can rely on the central difference quotient (cf. Lemma 2.1), i.e.,

$$\frac{\partial p}{\partial x_2}(s, i_2 h) \approx \frac{p(s,(i_2-1/2)h) - p(s,(i_2+1/2)h)}{h} \qquad \text{for all } s \in [(i_1-1)h, i_1 h].$$

The integral, on the other hand, can be approximated by a quadrature rule.

Lemma 4.9 (Midpoint rule) *Let $h \in \mathbb{R}_{>0}$ and $g \in C^2[-h,h]$. We can find $\eta \in (-h,h)$ with*

$$2hg(0) = \int_{-h}^{h} g(s)\,ds - \frac{h^3}{3} g''(\eta).$$

Proof. We define

$$\hat{g}\colon [-1,1] \to \mathbb{R}, \qquad\qquad s \mapsto g(sh),$$

and apply a change of variables to obtain

$$\int_{-h}^{h} g(s)\,ds = h \int_{-1}^{1} \hat{g}(s)\,ds.$$

We introduce $\varphi(s) := (s-1)^2/2$ and observe $\varphi'(s) = s-1$ and $\varphi''(s) = 1$. Partial integration yields

$$\begin{aligned}\int_0^1 \hat{g}(s)\,ds &= \int_0^1 \varphi''(s)\hat{g}(s)\,ds = \big[\varphi'(s)\hat{g}(s)\big]_{s=0}^1 - \int_0^1 \varphi'(s)\hat{g}'(s)\,ds \\ &= \hat{g}(0) - \big[\varphi(s)\hat{g}'(s)\big]_{s=0}^1 + \int_0^1 \varphi(s)\hat{g}''(s)\,ds \\ &= \hat{g}(0) + \hat{g}'(0)/2 + \int_0^1 \varphi(s)\hat{g}''(s)\,ds.\end{aligned}$$

Due to $\varphi(s) \geq 0$, we can apply the mean value theorem to find $\eta_+ \in (0,1)$ with

$$\int_0^1 \varphi(s)\hat{g}''(s)\,ds = \hat{g}''(\eta_+) \int_0^1 \varphi(s)\,ds = \frac{1}{6}\hat{g}''(\eta_+).$$

Reflecting $\hat{g}$ by zero gives the complementary result

$$\int_{-1}^{0} \hat{g}(s)\,ds = \int_0^1 \hat{g}(-s)\,ds = \hat{g}(0) - \hat{g}'(0)/2 + \frac{1}{6}\hat{g}''(\eta_-)$$

with $\eta_- \in (-1,0)$. Adding both equations leads to

$$\int_{-1}^{1} \hat{g}(s)\,ds = 2\hat{g}(0) + \frac{1}{3}\frac{\hat{g}''(\eta_+) + \hat{g}''(\eta_-)}{2},$$

and the intermediate value theorem yields $\hat\eta \in [\eta_-, \eta_+] \subseteq (-1,1)$ with

$$\hat g''(\eta) = \frac{\hat g''(\eta_+) + \hat g''(\eta_-)}{2},$$

and the chain rule leads to the desired result

$$\int_{-h}^{h} g(s)\,ds = h\int_{-1}^{1} \hat g(s)\,ds = 2h\hat g(0) + \frac{h}{3}\hat g''(\hat\eta) = 2hg(0) + \frac{h^3}{3}g''(\eta)$$

if we choose $\eta = h\hat\eta \in (-h,h)$. ∎

We approximate Darcy's law by

$$\begin{aligned} 0 &= f_{x,i} + k\int_{(i_1-1)h}^{i_1 h} \frac{\partial p}{\partial x_2}(s, i_2 h)\,ds \\ &\approx f_{x,i} + k\int_{(i_1-1)h}^{i_1 h} \frac{p(s,(i_2+\frac{1}{2})h) - p(s,(i_2-\frac{1}{2})h)}{h}\,ds \\ &\approx f_{x,i} + k(p((i_1-\tfrac{1}{2})h,(i_2+\tfrac{1}{2})h) - p((i_1-\tfrac{1}{2})h,(i_2-\tfrac{1}{2})h) \end{aligned}$$

and conclude that we only need the values of p in the midpoints of the squares. This leads us to introduce the second set of degrees of freedom as

$$p_i := p((i_1-\tfrac{1}{2})h,(i_2-\tfrac{1}{2})h) \qquad \text{for all } i \in \mathcal{I}.$$

Performing the same approximation steps for the y edges as well and dividing by the constant k leads us to the following approximation of Darcy's law:

$$0 \approx \tfrac{1}{k}f_{x,i} + p_{i_1,i_2+1} - p_i \qquad \text{for all } i \in \mathcal{I}_x := [1:n]\times[1:n-1], \tag{4.10a}$$

$$0 \approx \tfrac{1}{k}f_{y,i} + p_{i_1+1,i_2} - p_i \qquad \text{for all } i \in \mathcal{I}_y := [1:n-1]\times[1:n]. \tag{4.10b}$$

Together with the conservation equations (4.9) and conditions for boundary edges $f_{x,i}$ with $i \in \partial\mathcal{I}_x := \bar{\mathcal{I}}_x \setminus \mathcal{I}_x = [1:n]\times\{0,n\}$ and $f_{y,i}$ with $i \in \partial\mathcal{I}_y := \bar{\mathcal{I}}_y \setminus \mathcal{I}_y = \{0,n\}\times[1:n]$ yields a linear system that can be solved as long as the boundary conditions ensure that the inflow for Ω equals the outflow, since this is obviously necessary in order to have an equilibrium state.

The solution is not unique, since the pressure is only determined up to a global constant. This problem can be solved in various ways, e.g., by including an equation that forces the mean pressure to be zero or by using a suitable iterative solver like Uzawa's method.

5. Variational problems

While finite difference methods work quite well in a variety of applications, they are not very flexible when it comes to irregular geometries or solutions with limited differentiability.

Variational techniques can handle these situations far better: variational formulations of partial differential equations can lead to weak solutions where no classical solutions exist, and the Galerkin method offers a straightforward discretization scheme that preserves many of the original problem's properties.

5.1. Variational formulation

We consider the Poisson equation

$$-\Delta u(x) = f(x) \qquad \text{for all } x \in \Omega, \tag{5.1a}$$

$$u(x) = 0 \qquad \text{for all } x \in \partial\Omega \tag{5.1b}$$

in a bounded domain $\Omega \subseteq \mathbb{R}^d$ with a right-hand side function $f \in C(\Omega)$ and a solution $u \in C(\bar{\Omega})$ with $u|_\Omega \in C^2(\Omega)$.

Unfortunately, there are domains Ω and right-hand sides g such that no twice differentiable solution exists, and these are not particularly pathological examples but appear in real applications.

Since our differentiability requirements is "too strong", we consider weaker formulations. We construct these formulations in a way that ensures that a solution of the original problem is still a solution of the weaker problem, but that the weaker problem may have solutions where the original problem does not.

In a first step, we replace point-wise equality by averaged equality: we multiply (5.1a) by *test functions* $v \in C(\Omega)$ and integrate. This leads to the following weaker formulation:

Find $u \in C(\bar{\Omega})$ with $u|_\Omega \in C^2(\Omega)$ such that

$$-\int_\Omega v(x)\Delta u(x)\,dx = \int_\Omega v(x) f(x)\,dx \qquad \text{for all } v \in C(\Omega), \tag{5.2a}$$

$$u(x) = 0 \qquad \text{for all } x \in \partial\Omega. \tag{5.2b}$$

Obviously a solution of (5.1) is also a solution of (5.2).

Next we get rid of the requirement that u has to be twice differentiable in Ω: if v is continuously differentiable, we can apply partial integration and shift one derivative from u to v. By introducing

$$C_0^1(\Omega) := \{u \in C(\bar{\Omega}) \ : \ u|_\Omega \in C^1(\Omega),\ u|_{\partial\Omega} = 0\},$$

we can also incorporate the boundary conditions.

Let now $v \in C_0^1(\Omega)$. Due to Reminder 4.3, partial integration yields

$$\begin{aligned} -\int_\Omega v(x)\Delta u(x)\,dx &= -\int_\Omega v(x)\nabla \cdot \nabla u(x)\,dx \\ &= \int_\Omega \langle \nabla v(x), \nabla u(x)\rangle_2\,dx - \int_{\partial\Omega} v(x)\langle n(x), \nabla u(x)\rangle_2\,dx. \end{aligned}$$

By definition, we have $v|_{\partial\Omega} = 0$, so the boundary integral vanishes and we conclude

$$-\int_\Omega v(x)\Delta u(x)\,dx = \int_\Omega \langle \nabla v(x), \nabla u(x)\rangle_2\,dx.$$

The right-hand side requires u only to be differentiable, not twice differentiable, and since $C_0^1(\Omega)$ already includes the boundary conditions, we find the following weaker formulation:

Find $u \in C_0^1(\Omega)$ such that

$$\int_\Omega \langle \nabla v(x), \nabla u(x)\rangle_2\,dx = \int_\Omega v(x)f(x)\,dx \qquad \text{for all } v \in C_0^1(\Omega). \tag{5.3}$$

Once again, our construction ensures that a solution of the original problem (5.1) is also a solution of (5.3).

This is called a *variational formulation* of the original equation, since the equation has to hold for varying test functions $v \in C_0^1(\Omega)$.

Unfortunately, the requirement that u is once continuously differentiable is still too strong. We have to generalize what it means for a function to be differentiable.

5.2. Sobolev spaces

A closer look at (5.3) suggests that we actually do not need $\nabla u(x)$ to be continuous, it only has to be integrable. This suggests that we could weaken the definition of differentiability in the same way we have weakened the problem formulation: by multiplying by a test function an integrating.

Reminder 5.1 ($L^2(\Omega)$ and $L^2(\Omega, \mathbb{R}^d)$) *We denote the space of real-valued square integrable functions by*

$$L^2(\Omega) := \left\{ u : \Omega \to \mathbb{R} \ : \ u \textit{ is Lebesgue-measurable}, \ \int_\Omega u(x)^2\,dx < \infty \right\}$$

and the space of vector-valued square integrable functions by

$$L^2(\Omega, \mathbb{R}^d) := \left\{ u : \Omega \to \mathbb{R}^d \ : \ \|u\|_2 \textit{ is Lebesgue-measurable}, \ \int_\Omega \|u(x)\|_2^2\,dx < \infty \right\}.$$

Both are Hilbert spaces with the inner products

$$\langle v, u\rangle_{L^2} := \int_\Omega v(x)u(x)\,dx \qquad \text{for all } v, u \in L^2(\Omega),$$
$$\langle v, u\rangle_{L^2} := \int_\Omega \langle v(x), u(x)\rangle_2\,dx \qquad \text{for all } v, u \in L^2(\Omega, \mathbb{R}^d)$$

and the corresponding norms

$$\|u\|_{L^2} := \sqrt{\langle u, u\rangle_{L^2}} \qquad \text{for all } u \in L^2(\Omega) \text{ or } u \in L^2(\Omega, \mathbb{R}^d).$$

Hölder's inequality yields the Cauchy-Schwarz inequality

$$|\langle v, u\rangle_{L^2}| \leq \|v\|_{L^2}\|u\|_{L^2} \qquad \text{for all } u, v \in L^2(\Omega) \text{ or } u \in L^2(\Omega, \mathbb{R}^d). \tag{5.4}$$

As usual in this context, we treat functions that differ only on a null set as equal.

Partial integration allows us to shift derivatives between factors in an integral, and we plan to move all derivatives to the test function. This means that we should require the test function to be infinitely differentiable, and it means that we should also ensure that no boundary integrals appear during the partial integration.

Definition 5.2 (Support) *Let $u : \Omega \to \mathbb{R}$ or $u : \Omega \to \mathbb{R}^d$. The* support *of u is defined by*

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

This definition implies $u|_{\Omega\setminus\operatorname{supp}(u)} = 0$, so in order to ensure that u and (possibly its derivatives) vanish on the boundary of Ω, we have to keep $\operatorname{supp}(u)$ and $\partial\Omega$ disjoint.

Lemma 5.3 (Compact support) *Let $K \subseteq \mathbb{R}^d$ be a compact set, and let $\|\cdot\|$ denote a norm for $\mathbb{R}^d$.*

If $K \subseteq \Omega$, there is a $\delta \in \mathbb{R}_{>0}$ such that

$$\|x - y\| > \delta \qquad \text{for all } x \in K,\ y \in \partial\Omega.$$

Proof. Let $K \subseteq \Omega$.

If $K = \emptyset$, our claim is trivially satisfied since there is no $x \in K$.

Let now $K \neq \emptyset$. We denote open balls in $\mathbb{R}^d$ by

$$B(x, r) := \{y \in \mathbb{R}^d \ : \ \|y - x\| < r\}, \qquad \text{for all } x \in \mathbb{R}^d,\ r \in \mathbb{R}_{>0}.$$

Since Ω is an open set and $K \subseteq \Omega$, we can find an $\epsilon_x \in \mathbb{R}_{>0}$ for each $x \in K$ such that

$$B(x, 3\epsilon_x) \subseteq \Omega \qquad \text{for all } x \in K.$$

Then

$$\mathcal{C} := \{B(x, \epsilon_x) \ : \ x \in K\}$$

is an open cover of K. Since K is compact and non-empty, there is a finite and non-empty subset $A \subseteq K$ such that

$$K \subseteq \bigcup \{B(\widehat{x}, \epsilon_{\widehat{x}}) \ : \ \widehat{x} \in A\}.$$

We define $\delta := \min\{\epsilon_{\widehat{x}} \ : \ \widehat{x} \in A\}$.

Let now $x \in K$ and $y \in \partial\Omega$. We have seen that we can find $\widehat{x} \in A$ such that $x \in B(\widehat{x}, \epsilon_{\widehat{x}})$.

Since y is a boundary point, each open ball centered at y intersects the complement of Ω, i.e., we can find $z \in B(y, \epsilon_{\widehat{x}}) \cap (\mathbb{R}^d \setminus \Omega)$. Due to $B(\widehat{x}, 3\epsilon_{\widehat{x}}) \subseteq \Omega$ this means $z \notin B(\widehat{x}, 3\epsilon_{\widehat{x}})$, and the triangle inequality yields

$$\|\widehat{x} - y\| \geq \|\widehat{x} - z\| - \|z - y\| > 3\epsilon_{\widehat{x}} - \epsilon_{\widehat{x}} = 2\epsilon_{\widehat{x}}.$$

We can apply the triangle inequality again to obtain

$$\|x - y\| \geq \|\widehat{x} - y\| - \|x - \widehat{x}\| > 2\epsilon_{\widehat{x}} - \epsilon_{\widehat{x}} = \epsilon_{\widehat{x}} \geq \delta.$$

■

Applying this result to $K = \operatorname{supp}(u)$, we conclude that if a function has compact support, the support has to have a positive distance to the boundary, and therefore the function must be zero in an open neighbourhood of the boundary. In particular, not only the function vanishes in this neighbourhood, but also all of its derivatives.

This leads us to the definition

$$C_0^\infty(\Omega) := \{u \in C^\infty(\mathbb{R}^d) \ : \ \operatorname{supp}(u) \text{ is compact and } \operatorname{supp}(u) \subseteq \Omega\}.$$

Let us consider the j-th partial derivative for $j \in [1 : d]$. Let $u \in C^1(\Omega)$ and $\varphi \in C_0^\infty(\Omega)$, and let $\widehat{\varphi} \in C^\infty(\Omega, \mathbb{R}^d)$ be given by

$$\widehat{\varphi}_k = \begin{cases} \varphi & \text{if } k = j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } k \in [1 : d].$$

Applying partial integration (cf. Reminder 4.3), we find

$$\begin{aligned} 0 &= \int_{\partial\Omega} u(x) \langle n(x), \widehat{\varphi}(x) \rangle_2 \\ &= \int_\Omega u(x) \nabla \cdot \widehat{\varphi}(x)\, dx + \int_\Omega \langle \nabla u(x), \widehat{\varphi}(x) \rangle_2 \, dx \\ &= \int_\Omega u(x) \frac{\partial \varphi}{\partial x_j}(x)\, dx + \int_\Omega \frac{\partial u}{\partial x_j}(x) \varphi(x)\, dx \end{aligned}$$

and conclude

$$\int_\Omega \frac{\partial u}{\partial x_j}(x) \varphi(x)\, dx = -\int_\Omega u(x) \frac{\partial \varphi}{\partial x_j}(x)\, dx. \tag{5.5}$$

This equation suggests a generalization of the derivative: if we can find a square-integrable function $v \in L^2(\Omega)$ such that

$$\int_\Omega v(x)\varphi(x)\,dx = -\int_\Omega u(x)\frac{\partial\varphi}{\partial x_j}(x)\,dx \qquad \text{for all } \varphi \in C_0^\infty(\Omega),$$

we can use v as a "weak" j-th partial derivative of u.

In order to make handling higher derivatives easier, we introduce the set $\mathbb{N}_0^d$ of *multi-indices* and write

$$|\nu| := \nu_1 + \ldots + \nu_d \qquad \text{for all } \nu \in \mathbb{N}_0^d,$$
$$\partial_\nu u(x) := \frac{\partial^{\nu_1}}{\partial x_1^{\nu_1}} \cdots \frac{\partial^{\nu_d}}{\partial x_d^{\nu_d}} u(x) \qquad \text{for all } \nu \in \mathbb{N}_0^d,\ u \in C^{|\nu|}(\Omega),\ x \in \Omega.$$

Using (5.5) and our definition of the L^2-inner product, a straightforward induction yields

$$\langle \partial_\nu u, \varphi\rangle_{L^2} = (-1)^{|\nu|}\langle u, \partial_\nu\varphi\rangle_{L^2} \qquad \text{for all } \nu \in \mathbb{N}_0^d,\ u \in C^{|\nu|}(\Omega),\ \varphi \in C_0^\infty(\Omega).$$

Definition 5.4 (Weak derivatives) *Let $u \in L^2(\Omega)$ and $\nu \in \mathbb{N}_0^d$. If $v \in L^2(\Omega)$ satisfies*

$$\langle v, \varphi\rangle_{L^2} = (-1)^{|\nu|}\langle u, \partial_\nu\varphi\rangle_{L^2} \qquad \textit{for all } \varphi \in C_0^\infty(\Omega), \tag{5.6}$$

we call v a ν-th weak derivative *of u.*

We have already seen that for classically differentiable functions the derivative is also a weak derivative, so the weak derivative is a generalization. In order to be able to work with it (almost) as if it were a proper derivative, we have to ensure that it is uniquely determined by our definition.

Reminder 5.5 (Smooth approximation) *For $u \in L^2(\Omega)$ and $\epsilon \in \mathbb{R}_{>0}$, there is a function $\widetilde{u} \in C_0^\infty(\Omega)$ such that*

$$\|u - \widetilde{u}\|_{L^2} \le \epsilon.$$

Lemma 5.6 (Uniqueness of weak derivatives) *Let $u \in L^2(\Omega)$ and $\nu \in \mathbb{N}_0^d$.*

Let $v, w \in L^2(\Omega)$ be weak ν-th derivatives of u. Then we have $\|v - w\|_{L^2} = 0$, i.e., $v = w$.

Proof. Due to Reminder 5.5, we can find $\varphi \in C_0^\infty(\Omega)$ such that

$$\|(v - w) - \varphi\|_{L^2} \le \epsilon.$$

By the definition of the inner product, we get

$$\|v - w\|_{L^2}^2 = \langle v - w, v - w\rangle_{L^2} = \langle v - w, \varphi\rangle_{L^2} + \langle v - w, (v - w) - \varphi\rangle_{L^2}.$$

Since v and w are both weak derivatives of u, we have

$$\langle v, \varphi\rangle_{L^2} = (-1)^{|\nu|}\langle u, \partial_\nu\varphi\rangle_{L^2} = \langle w, \varphi\rangle_{L^2},$$

and therefore $\langle v - w, \varphi \rangle_{L^2} = 0$ and

$$\|v - w\|_{L^2}^2 = \langle v - w, (v - w) - \varphi \rangle_{L^2}.$$

Using the Cauchy-Schwarz inequality (5.4) yields

$$\begin{aligned} \|v - w\|_{L^2}^2 &\leq \|v - w\|_{L^2} \|(v - w) - \varphi\|_{L^2}, \\ \|v - w\|_{L^2} &\leq \|(v - w) - \varphi\|_{L^2} \leq \epsilon. \end{aligned}$$

Since we have proven this estimate for arbitrary $\epsilon \in \mathbb{R}_{>0}$, we conclude $\|v - w\|_{L^2} = 0$. ∎

Remark 5.7 (Orthogonality) *The proof of Lemma 5.6 uses a fairly common approach: in order to bound $v - w$, we require that $v - w$ can be approximated in a subspace, in this case $C_0^\infty(\Omega)$, and that it is orthogonal on this subspace, i.e., $\langle v - w, \varphi \rangle_{L^2} = 0$ for all $\varphi \in C_0^\infty(\Omega)$. Combining both properties yields an estimate for $v - w$.*

Definition 5.8 (Sobolev space) *Let $u \in L^2(\Omega)$ and $\nu \in \mathbb{N}_0^d$. If u has a ν-th weak derivative $v \in L^2(\Omega)$, it is unique by Lemma 5.6, and we denote it by $\partial_\nu u := v$.*

Let $m \in \mathbb{N}_0$. The space

$$H^m(\Omega) := \{u \in L^2(\Omega) \ : \ \textit{weak derivatives } \partial_\nu u \textit{ exist for all } \nu \in \mathbb{N}_0^d,\ |\nu| \leq m\}$$

is called the Sobolev space of m times weakly differentiable functions.

We equip this space with the norm

$$\|u\|_{H^m} := \Big(\sum_{\substack{\nu \in \mathbb{N}_0^d \\ |\nu| \leq m}} \|\partial_\nu u\|_{L^2}^2 \Big)^{1/2} \qquad \textit{for all } u \in H^m(\Omega),$$

the semi-norm

$$|u|_{H^m} := \Big(\sum_{\substack{\nu \in \mathbb{N}_0^d \\ |\nu| = m}} \|\partial_\nu u\|_{L^2}^2 \Big)^{1/2} \qquad \textit{for all } u \in H^m(\Omega),$$

and the corresponding inner product

$$\langle v, u \rangle_{H^m} := \sum_{\substack{\nu \in \mathbb{N}_0^d \\ |\nu| \leq m}} \langle \partial_\nu v, \partial_\nu u \rangle_{L^2} \qquad \textit{for all } u, v \in H^m(\Omega).$$

The Cauchy-Schwarz inequality (5.4) carries over to this norm and this inner product.

Exercise 5.9 (Completeness) *Let $m \in \mathbb{N}_0$. Prove that $H^m(\Omega)$ is a complete space, i.e., a Hilbert space.*

Hint: the fact that $L^2(\Omega)$ is a Hilbert space can be used to construct limits for Cauchy sequences. The Cauchy-Schwarz inequality (5.4) can be combined with (5.6) to prove that the limit of the ν-th weak derivatives of a sequence of functions is the ν-th weak derivative of the limit of the sequence.

Definition 5.10 (Weak gradient) *The* weak gradient *of u is given by*

$$\nabla u := \begin{pmatrix} \partial_{(1,0,\ldots,0)} u \\ \vdots \\ \partial_{(0,\ldots,0,1)} u \end{pmatrix} \qquad \text{for all } u \in H^1(\Omega).$$

For $u \in C^1(\Omega)$, it coincides with the gradient introduced in Definition 4.2.

Using the weak gradient, we can generalize the variational formulation (5.3), but there is a minor obstacle: due to Reminder 5.5, we can approximate an *arbitrary* function in $L^2(\Omega)$ by functions that are zero in a neighbourhood of the boundary and therefore vanish on the boundary. This implies that we cannot define the restriction $u|_{\partial\Omega}$ in the usual way for functions $u \in L^2(\Omega)$.

For continuous functions we can introduce the *trace operator*

$$\gamma\colon C(\bar\Omega) \to C(\partial\Omega), \qquad u \mapsto u|_{\partial\Omega} \tag{5.7}$$

that maps functions in $C(\bar\Omega)$ to their boundary values in $\partial\Omega$.

Theorem 5.11 (Trace operator) *Let $\Omega = (0,1)^2$. The trace operator γ satisfies*

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq 2\|u\|_{L^2} + 2\sqrt{\|u\|_{L^2}\|\nabla u\|_{L^2}} \leq 3\|u\|_{H^1} \qquad \text{for all } u \in C^1(\bar\Omega).$$

Proof. Let $u \in C^1(\bar\Omega)$. Let $y \in (0,1)$ and define

$$f_0\colon [0,1] \to \mathbb{R}, \qquad x \mapsto (1-x)\,u(x,y)^2,$$
$$f_1\colon [0,1] \to \mathbb{R}, \qquad x \mapsto x\,u(x,y)^2.$$

We have

$$f_0'(x) = -u(x,y)^2 + 2(1-x)u(x,y)\frac{\partial u}{\partial x}(x,y),$$
$$f_1'(x) = u(x,y)^2 + 2xu(x,y)\frac{\partial u}{\partial x}(x,y) \qquad \text{for all } x \in [0,1]$$

and

$$f_0(0) = u(0,y)^2, \qquad f_0(1) = 0, \qquad f_1(0) = 0, \qquad f_1(1) = u(1,y)^2.$$

Using the fundamental theorem of calculus, we find

$$\begin{aligned} u(0,y)^2 &= f_0(0) = f_0(1) - \int_0^1 f_0'(x)\,dx = \int_0^1 u(x,y)^2 - 2(1-x)u(x,y)\frac{\partial u}{\partial x}(x,y)\,dx, \\ &\leq \int_0^1 u(x,y)^2\,dx + 2\int_0^1 (1-x)|u(x,y)| \left|\frac{\partial u}{\partial x}(x,y)\right|\,dx, \\ u(1,y)^2 &= f_1(1) = f_1(0) + \int_0^1 f_1'(x)\,dx = \int_0^1 u(x,y)^2 + 2xu(x,y)\frac{\partial u}{\partial x}(x,y)\,dx \end{aligned}$$

$$\leq \int_0^1 u(x,y)^2\,dx + 2\int_0^1 x|u(x,y)|\left|\frac{\partial u}{\partial x}(x,y)\right|\,dx.$$

Adding both estimates and applying the Cauchy-Schwarz inequality (5.4) yields

$$\begin{aligned}
u(0,y)^2 + u(1,y)^2 &\leq 2\int_0^1 u(x,y)^2\,dx + 2\int_0^1 |u(x,y)|\left|\frac{\partial u}{\partial x}(x,y)\right|\,dx \\
&\leq 2\int_0^1 u(x,y)^2\,dx + 2\left(\int_0^1 u(x,y)^2\,dx\right)^{1/2}\left(\int_0^1 \frac{\partial u}{\partial x}(x,y)^2\,dx\right)^{1/2}.
\end{aligned}$$

Integrating both sides and applying the Cauchy-Schwarz inequality again gives us

$$\begin{aligned}
\int_0^1 u(0,y)^2 + u(1,y)^2\,dy &\leq 2\int_0^1\int_0^1 u(x,y)^2\,dx\,dy \\
&\quad + 2\int_0^1 \left(\int_0^1 u(x,y)^2\,dx\right)^{1/2}\left(\int_0^1 \frac{\partial u}{\partial x}(x,y)^2\,dx\right)^{1/2} dy \\
&\leq 2\|u\|_{L^2}^2 \\
&\quad + 2\left(\int_0^1\int_0^1 u(x,y)^2\,dx\,dy\right)^{1/2}\left(\int_0^1\int_0^1 \frac{\partial u}{\partial x}(x,y)^2\,dx\,dy\right)^{1/2} \\
&= 2\|u\|_{L^2}^2 + 2\|u\|_{L^2}\left\|\frac{\partial u}{\partial x}\right\|_{L^2}.
\end{aligned}$$

Applying the same arguments with x and y exchanged results in

$$\int_0^1 u(x,0)^2 + u(x,1)^2\,dx \leq 2\|u\|_{L^2}^2 + 2\|u\|_{L^2}\left\|\frac{\partial u}{\partial y}\right\|_{L^2},$$

and adding both estimates and using $a + b \leq 2\sqrt{a^2+b^2}$ leads to

$$\|\gamma(u)\|_{L^2(\partial\Omega)}^2 \leq 4\|u\|_{L^2}^2 + 2\|u\|_{L^2}\left(\left\|\frac{\partial u}{\partial x}\right\|_{L^2} + \left\|\frac{\partial u}{\partial y}\right\|_{L^2}\right) \leq 4\|u\|_{L^2}^2 + 4\|u\|_{L^2}\|\nabla u\|_{L^2}.$$

Due to $\sqrt{a+b} \leq \sqrt{a + 2\sqrt{ab} + b} = \sqrt{a} + \sqrt{b}$, we find

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq 2\|u\|_{L^2} + 2\sqrt{\|u\|_{L^2}\|\nabla u\|_{L^2}},$$

and $2ab \leq a^2 + b^2$ gives us the final estimate

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq 2\|u\|_{L^2} + \|u\|_{L^2} + \|\nabla u\|_{L^2} \leq 3\|u\|_{H^1}.$$

■

This result can be generalized: for any Lipschitz domain Ω, there is a constant C_γ such that

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq C_\gamma\|u\|_{H^1} \qquad \text{for all } u \in C^1(\bar\Omega).$$

In order to obtain a similar result for $u \in H^1(\Omega)$, we make use of the following extension of the approximation result of Reminder 5.5.

Theorem 5.12 (Meyers-Serrin) *Let $u \in H^m(\Omega)$ and $\epsilon \in \mathbb{R}_{>0}$. There is a function $\widetilde{u} \in C^\infty(\Omega)$ such that $\|u - \widetilde{u}\|_{H^m} \leq \epsilon$.*

Proof. cf. [8] ∎

Applied to $m = 1$, this result means that for any $u \in H^1(\Omega)$, we can find a sequence $(u_n)_{n=1}^\infty$ in $C^\infty(\Omega)$ such that

$$\lim_{n\to\infty} \|u - u_n\|_{H^1} = 0.$$

Due to Theorem 5.11, we have that $(\gamma(u_n))_{n=1}^\infty$ is a Cauchy sequence in $L^2(\partial\Omega)$, and since $L^2(\partial\Omega)$ is complete, it has to be convergent. We define

$$\gamma(u) := \lim_{n\to\infty} \gamma(u_n)$$

and thus obtain the extension

$$\gamma : H^1(\Omega) \to L^2(\partial\Omega)$$

of the trace operator satisfying

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq C_\gamma \|u\|_{H^1} \qquad \text{for all } u \in H^1(\Omega).$$

We could now use

$$H_0^1(\Omega) := \{u \in H^1(\Omega) \ : \ \gamma(u) = 0\}$$

to define the weak counterpart of $C_0^1(\Omega)$. This definition would immediately imply $C_0^1(\Omega) \subseteq H_0^1(\Omega)$, but it would make the proofs of some results, particularly Friedrichs' inequality (cf. Lemma 5.25) a little complicated.

Therefore we use a more general approach: since the space of infinitely differentiable functions is a dense subset of $H^m(\Omega)$, we can define $H_0^m(\Omega)$ as the closure of $C_0^\infty(\Omega)$ with respect to the H^m-norm. For $m = 1$, this is equivalent to the definition given above, but this statement will not be proven here.

Definition 5.13 (Homogeneous boundary conditions) *Let $m \in \mathbb{N}_0$. The space*

$$H_0^m(\Omega) := \{u \in H^m(\Omega) \ : \ \textit{for all } \epsilon \in \mathbb{R}_{>0} \textit{ there is a } \varphi \in C_0^\infty(\Omega) \textit{ with } \|u - \varphi\|_{H^m} \leq \epsilon\}$$

is called the Sobolev space of m times weakly differentiable functions with Dirichlet boundary conditions.

Exercise 5.14 (Completeness) *Let $m \in \mathbb{N}_0$. Prove that $H_0^m(\Omega)$ is a complete space, i.e., a Hilbert space.*

Now we are ready to introduce the final variational formulation of our model problem, the Poisson equation (5.1): we replace $C_0^1(\Omega)$ by $H_0^1(\Omega)$ and the gradient by the weak gradient.

Find $u \in H_0^1(\Omega)$ such that

$$\langle \nabla v, \nabla u \rangle_{L^2} = \langle v, f \rangle_{L^2} \qquad \text{for all } v \in H_0^1(\Omega). \tag{5.8}$$

Proving that a solution of (5.3) is also a solution of (5.8) requires two steps: first we have to demonstrate that $u \in C_0^1(\Omega)$ implies $u \in H_0^1(\Omega)$, and second we have to show that testing with function $v \in H_0^1(\Omega)$ instead of $C_0^1(\Omega)$ will not change the validity of the equation.

We have already completed the first step: we have $C^1(\Omega) \subseteq H^1(\Omega)$ due to partial integration, and $u \in C_0^1(\Omega)$ implies $\gamma(u) = 0$ and therefore $u \in H_0^1(\Omega)$.

The second step is a simple consequence of the Cauchy-Schwarz inequality: assume that (5.3) holds, and let $v \in H_0^1(\Omega)$. For each $\epsilon \in \mathbb{R}_{>0}$, we can find $\widetilde{v} \in C_0^\infty(\Omega)$ with $\|v - \widetilde{v}\|_{H^1} \leq \epsilon$ by Definition 5.13). Due to $\widetilde{v} \in C_0^\infty(\Omega) \subseteq C_0^1(\Omega)$, we have

$$\langle \nabla \widetilde{v}, \nabla u \rangle_{L^2} = \langle \widetilde{v}, f \rangle_{L^2}$$

and obtain

$$\begin{aligned} |\langle \nabla v, \nabla u \rangle_{L^2} - \langle v, f \rangle_{L^2}| &= |\langle \nabla v, \nabla u \rangle_{L^2} - \langle \nabla \widetilde{v}, \nabla u \rangle_{L^2} + \langle \widetilde{v}, f \rangle_{L^2} - \langle v, f \rangle_{L^2}| \\ &= |\langle \nabla (v - \widetilde{v}), \nabla u \rangle_{L^2} + \langle \widetilde{v} - v, f \rangle_{L^2}| \\ &\leq |\langle \nabla (v - \widetilde{v}), \nabla u \rangle_{L^2}| + |\langle \widetilde{v} - v, f \rangle_{L^2}|. \end{aligned}$$

Now we can apply the Cauchy-Schwarz inequality (5.4) to find

$$\begin{aligned} |\langle \nabla v, \nabla u \rangle_{L^2} - \langle v, f \rangle_{L^2}| &\leq \|\nabla (v - \widetilde{v})\|_{L^2} \|\nabla u\|_{L^2} + \|\widetilde{v} - v\|_{L^2} \|f\|_{L^2} \\ &\leq \epsilon \|\nabla u\|_{L^2} + \epsilon \|f\|_{L^2}. \end{aligned}$$

Since this holds for all $\epsilon \in \mathbb{R}_{>0}$, we conclude

$$\langle \nabla v, \nabla u \rangle_{L^2} = \langle v, f \rangle_{L^2},$$

i.e., the (5.8) holds for arbitrary test functions $v \in H_0^1(\Omega)$.

5.3. Solutions of variational problems

We investigate the existence and uniqueness of solutions of variational problems of the form (5.8) in a general setting: let $\mathcal{V}$ be a $\mathbb{R}$-Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ and the norm

$$\| \cdot \|_{\mathcal{V}} \colon \mathcal{V} \to \mathbb{R}_{\geq 0}, \qquad u \mapsto \sqrt{\langle u, u \rangle_{\mathcal{V}}}.$$

We write a general variational problem in the following form:

Find $u \in \mathcal{V}$ such that

$$a(v, u) = \beta(v) \qquad \text{for all } v \in \mathcal{V}. \tag{5.9}$$

Here $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is a bilinear form and $\beta : \mathcal{V} \to \mathbb{R}$ is a linear mapping.

Definition 5.15 (Dual space) *A continuous linear function $\lambda : \mathcal{V} \to \mathbb{R}$ mapping $\mathcal{V}$ into $\mathbb{R}$ is called a* functional*.*

The space of all functionals is called the dual space *of $\mathcal{V}$ and denoted by*

$$\mathcal{V}' := \{\lambda : \mathcal{V} \to \mathbb{R} \ : \ \lambda \textit{ is a functional}\}.$$

If is usually equipped with the dual norm

$$\|\cdot\|_{\mathcal{V}'} : \mathcal{V}' \to \mathbb{R}, \qquad\qquad \lambda \mapsto \sup\left\{\frac{|\lambda(v)|}{\|v\|_{\mathcal{V}}} \ : \ v \in \mathcal{V} \setminus \{0\}\right\}.$$

This is well-defined, since a linear continuous function is always bounded.

Lemma 5.16 (Right-hand side) *The right-hand side of our model problem (5.8) is given by*

$$\beta(v) = \langle v, f\rangle_{L^2} \qquad\qquad \textit{for all } v \in \mathcal{V} = H_0^1(\Omega).$$

This is a functional, i.e., we have $\beta \in \mathcal{V}'$, and the dual norm satisfies $\|\beta\|_{\mathcal{V}'} \leq \|f\|_{L^2}$.

Proof. For the model problem, we have $\mathcal{V} = H_0^1(\Omega)$.

Due to the Cauchy-Schwarz inequality (5.4), we have

$$|\beta(v)| = |\langle v, f\rangle_{L^2}| \leq \|v\|_{L^2}\|f\|_{L^2} \leq \|v\|_{H^1}\|f\|_{L^2} = \|v\|_{\mathcal{V}}\|f\|_{L^2} \quad \text{for all } v \in \mathcal{V} = H_0^1(\Omega).$$

This implies

$$\begin{aligned}\|\beta\|_{\mathcal{V}'} &= \sup\left(\frac{|\lambda(v)|}{\|v\|_{\mathcal{V}}} \ : \ v \in \mathcal{V} \setminus \{0\}\right)\\ &\leq \sup\left(\frac{\|v\|_{\mathcal{V}}\|f\|_{L^2}}{\|v\|_{\mathcal{V}}} \ : \ v \in \mathcal{V} \setminus \{0\}\right) = \|f\|_{L^2},\end{aligned}$$

so β is bounded, and therefore also continuous. ■

Definition 5.17 (Positive definite bilinear form) *A bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is called* positive definite *if*

$$a(u, u) > 0 \qquad\qquad \textit{for all } u \in \mathcal{V} \setminus \{0\}.$$

Definition 5.18 (Symmetric bilinear form) *A bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is called* symmetric *if*

$$a(v, u) = a(u, v) \qquad\qquad \textit{for all } u, v \in \mathcal{V}.$$

Lemma 5.19 (Minimization problem) *Let $\beta \in \mathcal{V}'$, and let $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a symmetric positive definite bilinear form. We define the function*

$$J \colon \mathcal{V} \to \mathbb{R}, \qquad v \mapsto a(v,v) - 2\beta(v).$$

Let $u, v \in \mathcal{V}$. We have

$$J(u) \leq J(u+tv) \qquad \text{for all } t \in \mathbb{R} \tag{5.10a}$$

if and only if

$$a(v,u) = \beta(v). \tag{5.10b}$$

In particular, $u \in \mathcal{V}$ is a solution of (5.9) if and only if it is a global minimum of J. In this case, it is the only global minimum.

Proof. If $v = 0$ holds, the equivalence is trivial. We assume $v \neq 0$.

We start by observing

$$\begin{aligned} J(u+tv) &= a(u+tv, u+tv) - 2\beta(u+tv) \\ &= a(u,u) + ta(u,v) + ta(v,u) + t^2 a(v,v) - 2\beta(u) - 2t\beta(v) \\ &= J(u) + 2t(a(v,u) - \beta(v)) + t^2 a(v,v) \qquad \text{for all } t \in \mathbb{R}. \end{aligned}$$

Now assume (5.10b) holds. It implies

$$J(u+tv) = J(u) + t^2 a(v,v) \geq J(u) \qquad \text{for all } t \in \mathbb{R},$$

and this is (5.10a).

Now assume (5.10a) holds. We choose

$$t := -\frac{a(v,u) - \beta(v)}{a(v,v)}$$

(we look for a minimum of $t \mapsto J(u+tv)$, i.e., for a zero of its derivative) and find

$$\begin{aligned} 0 \leq J(u+tv) - J(u) &= 2t(a(v,u) - \beta(v)) + t^2 a(v,v) \\ &= -2\frac{(a(v,u) - \beta(v))^2}{a(v,v)} + \frac{(a(v,u) - \beta(v))^2}{a(v,v)^2} a(v,v) \\ &= -\frac{(a(v,u) - \beta(v))^2}{a(v,v)} \leq 0. \end{aligned}$$

Due to $a(v,v) > 0$, this implies $(a(v,u) - \beta(v))^2 = 0$, and (5.10b) holds.

Let now (5.10b) hold, and let $\widetilde{u} \in \mathcal{V}$ be a solution of

$$a(v, \widetilde{u}) = \beta(v) \qquad \text{for all } v \in \mathcal{V}.$$

Then we have

$$a(v, u - \widetilde{u}) = a(v,u) - a(v,\widetilde{u}) = \beta(v) - \beta(v) = 0 \qquad \text{for all } v \in \mathcal{V},$$

and choosing $v := u - \widetilde{u}$ yields $a(v,v) = 0$. Since a is positive definite, this implies $v = 0$, i.e., $\widetilde{u} = u$. ■

Definition 5.20 (Bounded bilinear form) *A bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is called* bounded *if there is a constant $C_B \in \mathbb{R}_{\geq 0}$ such that*

$$|a(v,u)| \leq C_B \|v\|_{\mathcal{V}} \|u\|_{\mathcal{V}} \qquad \text{for all } v,u \in \mathcal{V}.$$

Definition 5.21 (Coercive bilinear form) *A bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is called* coercive *if it is bounded and there is a constant $C_K \in \mathbb{R}_{>0}$ such that*

$$C_K \|u\|_{\mathcal{V}}^2 \leq a(u,u) \qquad \text{for all } u \in \mathcal{V}.$$

Theorem 5.22 (Riesz) *Let $\beta \in \mathcal{V}'$, and let $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a symmetric and coercive bilinear form. Then there is exactly one $u \in \mathcal{V}$ such that*

$$a(v,u) = \beta(v) \qquad \text{for all } v \in \mathcal{V},$$

and we have

$$\frac{1}{C_B}\|\beta\|_{\mathcal{V}'} \leq \|u\|_{\mathcal{V}} \leq \frac{1}{C_K}\|\beta\|_{\mathcal{V}'}.$$

Proof. According to Lemma 5.19, we only have to find a global minimum of

$$J\colon \mathcal{V} \to \mathbb{R}, \qquad v \mapsto a(v,v) - 2\beta(v)$$

to find a solution of (5.9). Since a is coercive and β is bounded, we find

$$\begin{aligned}
J(v) &= a(v,v) - 2\beta(v) \geq C_K\|v\|_{\mathcal{V}}^2 - 2\|\beta\|_{\mathcal{V}'}\|v\|_{\mathcal{V}} \\
&= C_K\|v\|_{\mathcal{V}}^2 - 2\|\beta\|_{\mathcal{V}'}\|v\|_{\mathcal{V}} + \frac{\|\beta\|_{\mathcal{V}'}^2}{C_K} - \frac{\|\beta\|_{\mathcal{V}'}^2}{C_K} \\
&= \Big(\sqrt{C_K}\|v\|_{\mathcal{V}} - \sqrt{1/C_K}\|\beta\|_{\mathcal{V}'}\Big)^2 - \frac{\|\beta\|_{\mathcal{V}'}^2}{C_K} \geq -\frac{\|\beta\|_{\mathcal{V}'}^2}{C_K} \qquad \text{for all } v \in \mathcal{V}.
\end{aligned}$$

This implies that

$$\mu := \inf\{J(v) \ : \ v \in \mathcal{V}\}$$

is a real number, i.e.,

$$0 \geq \mu \geq -\|\beta\|_{\mathcal{V}'}^2/C_K > -\infty.$$

By the definition of the infimum, we can find a sequence $(u_n)_{n=1}^\infty$ such that

$$J(u_n) \leq \mu + 1/n \qquad \text{for all } n \in \mathbb{N}.$$

We will now prove that this is a Cauchy sequence. Let $n,m \in \mathbb{N}$. We have

$$\begin{aligned}
a(u_n - u_m, u_n - u_m) &= 2a(u_n,u_n) + 2a(u_m,u_m) - a(u_n+u_m, u_n+u_m) \\
&= 2J(u_n) + 4\beta(u_n) + 2J(u_m) + 4\beta(u_m) - 4a\Big(\frac{u_n+u_m}{2}, \frac{u_n+u_m}{2}\Big) \\
&= 2J(u_n) + 4\beta(u_n) + 2J(u_m) + 4\beta(u_m)
\end{aligned}$$

$$
\begin{aligned}
&\quad - 4J\Big(\frac{u_n + u_m}{2}\Big) - 8\beta\Big(\frac{u_n + u_m}{2}\Big)\\
&= 2J(u_n) + 4\beta(u_n) + 2J(u_m) + 4\beta(u_m)\\
&\quad - 4J\Big(\frac{u_n + u_m}{2}\Big) - 4\beta(u_n) - 4\beta(u_m)\\
&= 2J(u_n) + 2J(u_m) - 4J\Big(\frac{u_n + u_m}{2}\Big)\\
&\leq 2(\mu + 1/n) + 2(\mu + 1/m) - 4\mu = 2/n + 2/m.
\end{aligned}
$$

Let $\epsilon \in \mathbb{R}_{>0}$. We can find $n_0 \in \mathbb{N}$ such that $\frac{4}{C_K n_0} < \epsilon$. For all $n, m \in \mathbb{N}$ with $n, m \geq n_0$, we have just proven

$$
\|u_n - u_m\|_{\mathcal{V}}^2 \leq \frac{1}{C_K} a(u_n - u_m, u_n - u_m) \leq \frac{2/n + 2/m}{C_K} \leq \frac{4}{C_K n_0} < \epsilon,
$$

so $(u_n)_{n=1}^{\infty}$ is indeed a Cauchy sequence. Since $\mathcal{V}$ is complete, it converges to a vector $u \in \mathcal{V}$.

Let $\epsilon \in \mathbb{R}_{>0}$. We can find $n \in \mathbb{N}$ such that $1/n \leq \epsilon$ and $\|u - u_n\|_{\mathcal{V}} \leq \epsilon$. Since a and β are bounded, we find

$$
\begin{aligned}
J(u) &= a(u, u) - 2\beta(u)\\
&= a(u_n, u_n) + a(u - u_n, u_n) + a(u, u - u_n) - 2\beta(u_n) + 2\beta(u_n - u)\\
&= J(u_n) + a(u - u_n, u_n) + a(u, u - u_n) + 2\beta(u_n - u)\\
&\leq J(u_n) + C_B\|u - u_n\|_{\mathcal{V}}\|u_n\|_{\mathcal{V}} + C_B\|u\|_{\mathcal{V}}\|u - u_n\|_{\mathcal{V}} + 2\|\beta\|_{\mathcal{V}'}\|u - u_n\|_{\mathcal{V}}\\
&\leq \mu + \epsilon + C_B\epsilon\|u_n\|_{\mathcal{V}} + C_B\epsilon\|u\|_{\mathcal{V}} + 2\epsilon\|\beta\|_{\mathcal{V}'}\\
&\leq \mu + \epsilon + C_B\epsilon\|u\|_{\mathcal{V}} + C_B\epsilon\|u_n - u\|_{\mathcal{V}} + C_B\epsilon\|u\|_{\mathcal{V}} + 2\epsilon\|\beta\|_{\mathcal{V}'}\\
&\leq \mu + \epsilon(1 + 2C_B\|u\|_{\mathcal{V}} + 2\|\beta\|_{\mathcal{V}'} + C_B\epsilon).
\end{aligned}
$$

Since ϵ can be chosen arbitrarily small, we conclude $J(u) = \mu$, i.e., the function J is minimal at the point u.

Due to Lemma 5.19, u is the unique solution of (5.9).

We have

$$
C_K\|u\|_{\mathcal{V}}^2 \leq a(u, u) = \beta(u) \leq \|\beta\|_{\mathcal{V}'}\|u\|_{\mathcal{V}},
$$

and this implies $\|u\|_{\mathcal{V}} \leq \|\beta\|_{\mathcal{V}'}/C_K$.

Let $\epsilon \in \mathbb{R}_{>0}$. By definition, we can find $v \in \mathcal{V} \setminus \{0\}$ such that

$$
\|\beta\|_{\mathcal{V}'} \leq \frac{|\beta(v)|}{\|v\|_{\mathcal{V}}} + \epsilon = \frac{|a(v, u)|}{\|v\|_{\mathcal{V}}} + \epsilon \leq \frac{C_B\|v\|_{\mathcal{V}}\|u\|_{\mathcal{V}}}{\|v\|_{\mathcal{V}}} + \epsilon = C_B\|u\|_{\mathcal{V}} + \epsilon,
$$

and since ϵ can be chosen arbitrarily small, this implies $\|\beta\|_{\mathcal{V}'}/C_B \leq \|u\|_{\mathcal{V}}$. ■

Remark 5.23 (Lower bound) *In the first part of the previous proof, we rely on the estimate*

$$
J(v) \geq -\|\beta\|_{V'}^2/C_K \qquad \textit{for all } v \in V.
$$

to ensure that the infimum of J is finite.

For the solution u of the variational problem, we obtain

$$J(u) = a(u,u) - 2\beta(u) = -\beta(u) \geq -\|\beta\|_{V'}\|u\|_V \geq -\|\beta\|_{V'}^2/C_K$$

using the stability estimate provided by the Theorem 5.22, i.e., the lower bound can be sharp for a suitable choice of β and a.

Corollary 5.24 (Riesz representation theorem) *The mapping*

$$\Psi_{\mathcal{V}} \colon \mathcal{V} \to \mathcal{V}', \qquad u \mapsto \langle \cdot, u\rangle_{\mathcal{V}},$$

is bijective and satisfies

$$\|\Psi_{\mathcal{V}} u\|_{\mathcal{V}'} = \|u\|_{\mathcal{V}} \qquad \text{for all } u \in \mathcal{V},$$

i.e, it is an isometric isomorphism *between $\mathcal{V}$ and the dual space $\mathcal{V}'$.*

Proof. The mapping $\Psi_{\mathcal{V}}$ is obviously linear.

Let $u \in \mathcal{V}$. Then we have

$$\|u\|_{\mathcal{V}}^2 = \langle u, u\rangle_{\mathcal{V}} = |(\Psi_{\mathcal{V}} u)(u)| \leq \|\Psi_{\mathcal{V}} u\|_{\mathcal{V}'}\|u\|_{\mathcal{V}},$$

and this implies

$$\|u\|_{\mathcal{V}} \leq \|\Psi_{\mathcal{V}} u\|_{\mathcal{V}'} \qquad \text{for all } u \in \mathcal{V}.$$

The Cauchy-Schwarz inequality yields

$$\begin{aligned}\|\Psi_{\mathcal{V}} u\|_{\mathcal{V}'} &= \sup\left\{\frac{|(\Psi_{\mathcal{V}} u)(v)|}{\|v\|_{\mathcal{V}}} \;:\; v \in \mathcal{V} \setminus \{0\}\right\}\\ &= \sup\left\{\frac{|\langle v, u\rangle_{\mathcal{V}}|}{\|v\|_{\mathcal{V}}} \;:\; v \in \mathcal{V} \setminus \{0\}\right\}\\ &\leq \sup\left\{\frac{\|v\|_{\mathcal{V}}\|u\|_{\mathcal{V}}}{\|v\|_{\mathcal{V}}} \;:\; v \in \mathcal{V} \setminus \{0\}\right\} = \|u\|_{\mathcal{V}} \qquad \text{for all } u \in \mathcal{V}.\end{aligned}$$

We conclude that $\Psi_{\mathcal{V}}$ is injective and isometric.

Let $\beta \in \mathcal{V}'$. Applying Theorem 5.22 to

$$a(v,u) := \langle v, u\rangle_{\mathcal{V}} \qquad \text{for all } u, v \in \mathcal{V}$$

yields $u \in \mathcal{V}$ with $\Psi_{\mathcal{V}} u = \beta$, so $\Psi_{\mathcal{V}}$ is also surjective. ■

In order to apply Theorem 5.22 to our model problem, we have to establish that it it coercive.

Lemma 5.25 (Friedrichs' inequality) *Let $B := [a,b] \times \mathbb{R}^{d-1}$ be given such that $\Omega \subseteq B$, and let $R := b - a$.*

We have

$$\|u\|_{L^2} \leq R\|\nabla u\|_{L^2} \qquad \text{for all } u \in H_0^1(\Omega).$$

Proof. We first consider $u \in C_0^\infty(\Omega)$. Let $x \in \Omega$. Due to the fundamental theorem of calculus, we have

$$u(x) = u(a) + \int_a^{x_1} \frac{\partial u}{\partial x_1}(y, x_2, \ldots, x_d)\, dy, \tag{5.11}$$

and due to $u \in C_0^\infty(\Omega)$, we also have $u(a) = 0$.

Introducing

$$\begin{aligned} \widehat{x} &:= \begin{pmatrix} x_2 \\ \vdots \\ x_d \end{pmatrix} && \text{for all } x \in \Omega, \\ \widehat{\Omega} &:= \{\widehat{x} \ :\ x \in \Omega\}, \end{aligned}$$

we can write (5.11) in the shorter form

$$u(x) = \int_a^{x_1} \frac{\partial u}{\partial x_1}(y, \widehat{x})\, dy \qquad\qquad \text{for all } x \in \Omega.$$

Squaring, integrating, and the Cauchy-Schwarz inequality (5.4) yields

$$\begin{aligned} \|u\|_{L^2}^2 &= \int_\Omega u(x)^2\, dx = \int_\Omega \Big(\int_a^{x_1} \frac{\partial u}{\partial x_1}(y, \widehat{x})\, dy \Big)^2 dx \\ &\leq \int_\Omega \int_a^{x_1} 1\, dy \int_a^{x_1} \Big(\frac{\partial u}{\partial x_1}(y, \widehat{x}) \Big)^2 dy\, dx \\ &\leq R \int_\Omega \int_a^b \Big(\frac{\partial u}{\partial x_1}(y, \widehat{x}) \Big)^2 dy\, dx \\ &= R \int_a^b \int_{\widehat{\Omega}} \int_a^b \Big(\frac{\partial u}{\partial x_1}(y, \widehat{x}) \Big)^2 dy\, d\widehat{x}\, dz \\ &= R \int_a^b \int_\Omega \Big(\frac{\partial u}{\partial x_1}(x) \Big)^2 dx\, dz \\ &= R \int_a^b \left\| \frac{\partial u}{\partial x_1} \right\|_{L^2}^2 dz = R^2 \left\| \frac{\partial u}{\partial x_1} \right\|_{L^2}^2 \leq R^2 \|\nabla u\|_{L^2}^2. \end{aligned}$$

We have proven

$$\|u\|_{L^2} \leq R \|\nabla u\|_{L^2} \qquad\qquad \text{for all } u \in C_0^\infty(\Omega).$$

Let now $u \in H_0^1(\Omega)$, and let $\epsilon \in \mathbb{R}_{>0}$. Due to Definition 5.13, we can find $\widetilde{u} \in C_0^\infty(\Omega)$ such that

$$\|u - \widetilde{u}\|_{H^1} \leq \epsilon.$$

Since we have already proven the desired estimate for all functions in $C_0^\infty(\Omega)$, we can use the triangle inequality to find

$$\begin{aligned} \|u\|_{L^2} &\leq \|\widetilde{u}\|_{L^2} + \|u - \widetilde{u}\|_{L^2} \leq R \|\nabla \widetilde{u}\|_{L^2} + \epsilon \\ &\leq R \|\nabla u\|_{L^2} + R \|\nabla(\widetilde{u} - u)\|_{L^2} + \epsilon \leq R \|\nabla u\|_{L^2} + R\epsilon + \epsilon. \end{aligned}$$

This holds for all ϵ, so our proof is complete. ∎

Corollary 5.26 (Model problem) *Let R be chosen as in Lemma 5.25. The bilinear form*

$$a\colon H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, \qquad (v,u) \mapsto \langle \nabla v, \nabla u \rangle_{L^2},$$

satisfies

$$\begin{aligned} |a(v,u)| &\leq \|\nabla v\|_{L^2} \|\nabla u\|_{L^2} \leq \|v\|_{H^1} \|u\|_{H^1} && \text{for all } v,u \in H^1(\Omega),\\ a(u,u) &\geq \frac{1}{1+R^2} \|u\|_{H^1}^2 && \text{for all } u \in H_0^1(\Omega). \end{aligned}$$

i.e., it is bounded and coercive.

Proof. Let $v,u \in H^1(\Omega)$. Due to the Cauchy-Schwarz inequality (5.4), we have

$$|a(v,u)| = |\langle \nabla v, \nabla u \rangle_{L^2}| \leq \|\nabla v\|_{L^2} \|\nabla u\|_{L^2} \leq \|v\|_{H^1} \|u\|_{H^1},$$

so a is continuous with the continuity constant $C_B = 1$.

Let $u \in H_0^1(\Omega)$. By Friedrichs' Lemma 5.25, we have

$$\begin{aligned} \|u\|_{L^2}^2 &\leq R^2 \|\nabla u\|_{L^2}^2,\\ \|u\|_{H^1}^2 &= \|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2 \leq R^2 \|\nabla u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2\\ &= (1+R^2) \|\nabla u\|_{L^2}^2 = (1+R^2) a(u,u), \end{aligned}$$

and this implies

$$\frac{1}{1+R^2} \|u\|_{H^1}^2 \leq a(u,u),$$

so a is coercive with the coercivity constant $C_K = 1/(1+R^2)$. ∎

We can interprete this result as a norm equivalence: on the subspace $H_0^1(\Omega)$ with homogeneous Dirichlet boundary conditions, the semi-norm

$$|u|_{H^1} := \|\nabla u\|_{H^1} \qquad \text{for all } u \in H^1(\Omega)$$

is in fact a norm, and equivalent to the norm $\|u\|_{H^1}$. This is obviously not the case without the boundary conditions, since we can choose $u = 1$ and obtain $\|\nabla u\|_{L^2} = 0$ and $\|u\|_{H^1} = \|u\|_{L^2} = \sqrt{|\Omega|} > 0$.

Riesz' Theorem 5.22 requires the bilinear form a to be symmetric, and this property is not guaranteed for all partial differential equations we might want to investigate. We now consider a generalization of this existence result.

If a is bounded, we can define the operator

$$\mathcal{A}\colon \mathcal{V} \to \mathcal{V}', \qquad u \mapsto a(\cdot, u),$$

and write (5.9) in the compact form

$$\mathcal{A}u = \beta.$$

Due to

$$\begin{aligned}\|\mathcal{A}u\|_{\mathcal{V}'} &= \sup\Big(\frac{|a(v,u)|}{\|v\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\Big)\\ &\leq \sup\Big(\frac{C_B\|v\|_{\mathcal{V}}\|u\|_{\mathcal{V}}}{\|v\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\Big)\leq C_B\|u\|_{\mathcal{V}} \qquad \text{for all } u\in\mathcal{V},\end{aligned}$$

the operator $\mathcal{A}$ is well-defined and bounded, i.e., continuous.

Lemma 5.27 (Bounded inverse) *Let $\mathcal{A}$ be invertible. The inverse is bounded if and only if there is an $\alpha\in\mathbb{R}_{>0}$ such that*

$$\alpha\|u\|_{\mathcal{V}}\leq\|\mathcal{A}u\|_{\mathcal{V}'}=\sup\left\{\frac{|a(v,u)|}{\|v\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\right\} \qquad \textit{for all } u\in\mathcal{V}. \tag{5.12}$$

In this case, we have $\|\mathcal{A}^{-1}\|_{\mathcal{V}\leftarrow\mathcal{V}'}\leq 1/\alpha$.

Proof. We assume that $\mathcal{A}^{-1}$ is bounded, i.e.,

$$\|\mathcal{A}^{-1}\|_{\mathcal{V}'\leftarrow\mathcal{V}}<\infty.$$

By definition, this implies

$$\begin{aligned}\|\mathcal{A}^{-1}\lambda\|_{\mathcal{V}} &\leq \|\mathcal{A}^{-1}\|_{\mathcal{V}\leftarrow\mathcal{V}'}\|\lambda\|_{\mathcal{V}'}\\ &= \|\mathcal{A}^{-1}\|_{\mathcal{V}\leftarrow\mathcal{V}'}\sup\left\{\frac{|\lambda(v)|}{\|v\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\right\} \qquad \text{for all } \lambda\in\mathcal{V}'.\end{aligned}$$

Let now $u\in\mathcal{V}$ and $\lambda:=\mathcal{A}u$. We obtain

$$\begin{aligned}\|u\|_{\mathcal{V}} &\leq \|\mathcal{A}^{-1}\|_{\mathcal{V}\leftarrow\mathcal{V}'}\sup\left\{\frac{|(\mathcal{A}u)(v)|}{\|v\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\right\}\\ &= \|\mathcal{A}^{-1}\|_{\mathcal{V}\leftarrow\mathcal{V}'}\sup\left\{\frac{|a(v,u)|}{\|v\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\right\},\end{aligned}$$

and this implies

$$\frac{1}{\|\mathcal{A}^{-1}\|_{\mathcal{V}\leftarrow\mathcal{V}'}}\leq\sup\left\{\frac{|a(v,u)|}{\|v\|_{\mathcal{V}}\|u\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\right\}.$$

Assume now that (5.12) holds. Let $\lambda\in\mathcal{V}'$ and $u:=\mathcal{A}^{-1}\lambda$. We have

$$\begin{aligned}\alpha\|\mathcal{A}^{-1}\lambda\|_{\mathcal{V}}=\alpha\|u\|_{\mathcal{V}} &\leq \sup\left\{\frac{|a(v,u)|}{\|v\|_{\mathcal{V}}\|u\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\right\}\\ &= \sup\left\{\frac{|\lambda(v)|}{\|v\|_{\mathcal{V}}\|u\|_{\mathcal{V}}}\ :\ v\in\mathcal{V}\setminus\{0\}\right\}=\|\lambda\|_{\mathcal{V}'},\end{aligned}$$

and this implies

$$\|\mathcal{A}^{-1}\lambda\|_{\mathcal{V}}\leq\frac{1}{\alpha}\|\lambda\|_{\mathcal{V}'}.$$

By definition of the operator norm, this is equivalent with $\|\mathcal{A}^{-1}\|_{\mathcal{V}\leftarrow\mathcal{V}'} \leq 1/\alpha$. ∎

The condition (5.12) is frequently written in the form

$$0 < \alpha := \inf\left\{\sup\left\{\frac{|a(v,u)|}{\|v\|_{\mathcal{V}}\|u\|_{\mathcal{V}}} \ :\ v \in \mathcal{V}\setminus\{0\}\right\} \ :\ u \in \mathcal{V}\setminus\{0\}\right\}$$

and called an *inf-sup condition.*

Although this condition does not guarantee the invertibility of $\mathcal{A}$, it ensures two other important properties.

Lemma 5.28 (Closed range) *Let a be continuous and satisfy (5.12). Then $\mathcal{A}$ is injective and the* range

$$\mathcal{R}(\mathcal{A}) := \{\mathcal{A}u \ :\ u \in \mathcal{V}\}$$

is a closed subspace of the dual space $\mathcal{V}'$.

Proof. We prove injectivity by contraposition: let $u \in \mathcal{V}\setminus\{0\}$. Applying (5.12) yields

$$\|\mathcal{A}u\|_{\mathcal{V}'} = \sup\left\{\frac{|a(v,u)|}{\|v\|_{\mathcal{V}}\|u\|_{\mathcal{V}}} \ :\ v \in \mathcal{V}\setminus\{0\}\right\}\|u\|_{\mathcal{V}} \geq \alpha\|u\|_{\mathcal{V}} > 0,$$

i.e., $\mathcal{A}u \neq 0$. Therefore $\mathcal{A}u = 0$ implies $u = 0$ and $\mathcal{A}$ has to be injective.

Let now $(\lambda_n)_{n=1}^\infty$ be a convergent sequence in $\mathcal{R}(\mathcal{A})$. By definition, we can find a sequence $(u_n)_{n=1}^\infty$ with

$$\mathcal{A}u_n = \lambda_n \qquad\qquad \text{for all } n \in \mathbb{N}.$$

Let $n, m \in \mathbb{N}$. We apply (5.12) to find

$$\alpha\|u_n - u_m\|_{\mathcal{V}} \leq \|\mathcal{A}(u_n - u_m)\|_{\mathcal{V}'} = \|\mathcal{A}u_n - \mathcal{A}u_m\|_{\mathcal{V}'} = \|\lambda_n - \lambda_m\|_{\mathcal{V}'}.$$

Since $(\lambda_n)_{n=1}^\infty$ is convergent, it is also a Cauchy sequence. We have just proven that the same holds for $(u_n)_{n=1}^\infty$, so there is a $u \in \mathcal{V}$ with

$$u = \lim_{n\to\infty} u_n.$$

Since $\mathcal{A}$ is continuous, we have

$$\lim_{n\to\infty} \lambda_n = \lim_{n\to\infty} \mathcal{A}u_n = \mathcal{A}\lim_{n\to\infty} u_n = \mathcal{A}u \in \mathcal{R}(\mathcal{A}),$$

i.e., $\mathcal{R}(\mathcal{A})$ is a closed subspace. ∎

In order to ensure that $\mathcal{A}$ is surjective, we need a criterion for checking $\mathcal{R}(\mathcal{A}) = \mathcal{V}'$.

Lemma 5.29 (Orthogonal projection) *Let $\mathcal{S} \subseteq \mathcal{V}$ be a closed subspace. There is a linear mapping*

$$\Pi_{\mathcal{S}} : \mathcal{V} \to \mathcal{S}$$

such that

$$\langle v, \Pi_{\mathcal{S}}u\rangle_{\mathcal{V}} = \langle v, u\rangle_{\mathcal{V}} \qquad\qquad \textit{for all } u \in \mathcal{V},\ v \in \mathcal{S}. \tag{5.13}$$

If $\mathcal{S} \neq \mathcal{V}$, we can find $w \in \mathcal{V}$ with $w \neq 0$ and

$$\langle v, w\rangle_{\mathcal{V}} = 0 \qquad\qquad \textit{for all } v \in \mathcal{S}.$$

Proof. Let

$$\Psi_{\mathcal{V}} \colon \mathcal{V} \to \mathcal{V}', \qquad\qquad u \mapsto \langle \cdot, u \rangle_{\mathcal{V}},$$
$$\Psi_{\mathcal{S}} \colon \mathcal{S} \to \mathcal{S}', \qquad\qquad u \mapsto \langle \cdot, u \rangle_{\mathcal{V}},$$

denote the Riesz isomorphisms on $\mathcal{V}$ and $\mathcal{S}$. Since $\mathcal{S}$ is a closed subspace of the Hilbert space $\mathcal{V}$, Corollary 5.24 guarantees that both are isometric isomorphisms

Since $\mathcal{S}$ and $\mathcal{V}$ share the same norm, we have $\mathcal{V}' \subseteq \mathcal{S}'$ and can define

$$\Pi_{\mathcal{S}} := \Psi_{\mathcal{S}}^{-1} \Psi_{\mathcal{V}}.$$

Let now $u \in \mathcal{V}$ and $v \in \mathcal{S}$. We have

$$\langle v, \Pi_{\mathcal{S}} u \rangle_{\mathcal{V}} = (\Psi_{\mathcal{V}} u)(v) = \langle v, u \rangle_{\mathcal{V}}.$$

Assume $\mathcal{S} \neq \mathcal{V}$. This implies that we can find $z \in \mathcal{V} \setminus \mathcal{S}$. Let $w := z - \Pi_{\mathcal{S}} z$. Due to $z \notin \mathcal{S}$ and $\Pi_{\mathcal{S}} z \in \mathcal{S}$, we have $w \neq 0$, and (5.13) yields

$$\langle v, w \rangle_{\mathcal{V}} = \langle v, z - \Pi_{\mathcal{S}} z \rangle_{\mathcal{V}} = \langle v, z \rangle_{\mathcal{V}} - \langle v, \Pi_{\mathcal{S}} z \rangle_{\mathcal{V}} = 0 \qquad \text{for all } v \in \mathcal{S}.$$

■

Remark 5.30 (Adjoint operator) *Let $\mathcal{V}$ and $\mathcal{W}$ be $\mathbb{R}$-Hilbert spaces, and let $\mathcal{A} \colon \mathcal{V} \to \mathcal{W}$ denote a bounded operator.*

For every $w \in \mathcal{W}$, the mapping

$$\lambda_w \colon \mathcal{V} \to \mathbb{R}, \qquad\qquad v \mapsto \langle w, \mathcal{A} v \rangle_{\mathcal{W}},$$

is a functional in $\mathcal{V}'$. Using the Riesz isomorphism $\Psi_{\mathcal{V}}$, we define a new operator $\mathcal{A}^* \colon \mathcal{W} \to \mathcal{V}$ by

$$\mathcal{A}^* w := \Psi_{\mathcal{V}}^{-1} \lambda_w \qquad\qquad \text{for all } w \in W$$

and observe

$$\langle w, \mathcal{A} v \rangle_{\mathcal{W}} = \lambda_w(v) = \langle \mathcal{A}^* w, v \rangle_{\mathcal{V}} \qquad\qquad \text{for all } v \in \mathcal{V},\ w \in \mathcal{W}.$$

We call $\mathcal{A}^*$ the adjoint *of $\mathcal{A}$.*

Denote the range of $\mathcal{A}$ by $\mathcal{S} := \mathcal{R}(\mathcal{A}) \subseteq \mathcal{W}$ and the null space of $\mathcal{A}^*$ by $\mathcal{N}(\mathcal{A}^*) := \{w \in \mathcal{W} \ : \ \mathcal{A}^* w = 0\}$, and assume that the range $\mathcal{S}$ is a closed subspace.

Let $w \in \mathcal{W}$. Using the orthogonal projection $\Pi_{\mathcal{S}}$ introduced in Lemma 5.29, we can define $w_1 := \Pi_{\mathcal{S}} w$ and $w_2 := w - w_1$. By definition, we have $w_1 \in \mathcal{R}(\mathcal{A})$. Due to Lemma 5.29, we have

$$0 = \langle w - \Pi_{\mathcal{S}} w, \mathcal{A} v \rangle_{\mathcal{W}} = \langle w_2, \mathcal{A} v \rangle_{\mathcal{W}} = \langle \mathcal{A}^* w_2, v \rangle_{\mathcal{W}} \qquad\qquad \text{for all } v \in \mathcal{V},$$

and therefore $\mathcal{A}^* w_2 = 0$, i.e., $w_2 \in \mathcal{N}(\mathcal{A}^*)$.

We conclude that $\mathcal{W}$ is the orthogonal direct sum of $\mathcal{R}(\mathcal{A})$ and $\mathcal{N}(\mathcal{A}^*)$.

In particular, this means that if a vector $w \in \mathcal{W}$ is orthogonal with respect to the null space $\mathcal{N}(\mathcal{A}^*)$, it belongs to the range $\mathcal{R}(\mathcal{A})$ of $\mathcal{A}$.

Theorem 5.31 (Babuška-Lax-Milgram) *Let a be continuous and satisfy the inf-sup condition (5.12).*

If for each $w \in \mathcal{V} \setminus \{0\}$ there is a $u \in \mathcal{V}$ such that $a(w,u) \neq 0$, the operator $\mathcal{A}$ has an inverse satisfying $\|\mathcal{A}^{-1}\|_{\mathcal{V} \leftarrow \mathcal{V}'} \leq 1/\alpha$.

Proof. Due to Lemma 5.27, we know that if an inverse exists, it has to be bounded. Due to Lemma 5.28, we know that $\mathcal{R}(\mathcal{A})$ is a closed subspace and that $\mathcal{A}$ is injective.

We prove our claim by contraposition: assume that $\mathcal{A}$ does *not* have a bounded inverse. We have already seen that this can only be the case if it is not surjective, so $\mathcal{R}(\mathcal{A}) \neq \mathcal{V}'$ holds. Applying the Riesz isomorphism yields that

$$\mathcal{X} := \Psi_{\mathcal{V}}^{-1}(\mathcal{R}(\mathcal{A}))$$

is a closed proper subspace of $\mathcal{V}$. Lemma 5.29 gives us a $w \in \mathcal{V} \setminus \{0\}$ such that

$$\langle v, w \rangle_{\mathcal{V}} = 0 \qquad \text{for all } v \in \mathcal{X}.$$

Since the inner product is symmetric, we can apply the definition of $\mathcal{X}$ to find

$$0 = \langle w, \Psi_{\mathcal{V}}^{-1}\lambda \rangle_{\mathcal{V}} = \lambda(w) \qquad \text{for all } \lambda \in \mathcal{R}(\mathcal{A})$$

and by definition

$$0 = (\mathcal{A}u)(w) = a(w,u) \qquad \text{for all } u \in \mathcal{V}.$$

■

5.4. Galerkin methods

In order to solve a variational problem of the form (5.9), we have to make it finite-dimensional, i.e., *discretize* it.

A particularly elegant and general approach is the *Galerkin discretization*: we fix a finite-dimensional subspace $\mathcal{V}_n \subseteq \mathcal{V}$ and solve the following finite-dimensional variational problem:

Find $u_n \in \mathcal{V}_n$ such that

$$a(v_n, u_n) = \beta(v_n) \qquad \text{for all } v_n \in \mathcal{V}_n. \tag{5.14}$$

Let $n \in \mathbb{N}$ denote the dimension of $\mathcal{V}_n$.

The most important property of any discretization scheme is, of course, that it yields an approximation of the original problem that can be actually be solved. In the case of the Galerkin method, we can translate the discretized variational problem (5.14) into a system of linear equations that can be solved by standard algorithms.

Lemma 5.32 (Linear system) *Let $(\varphi_i)_{i=1}^n$ be a basis of $\mathcal{V}_n$, and let $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$ be defined by*

$$a_{ij} := a(\varphi_i, \varphi_j), \qquad b_i = \beta(\varphi_i) \qquad \text{for all } i, j \in [1:n].$$

Let $x \in \mathbb{R}^n$ and

$$u_n = \sum_{j=1}^n x_j \varphi_j. \tag{5.15}$$

We have $Ax = b$ if and only if u_n is a solution of (5.14).

Proof. Assume first that $Ax = b$ holds. Let $v_n \in \mathcal{V}_n$. Since $(\varphi_i)_{i=1}^n$ is a basis of $\mathcal{V}_n$, we can find coefficients $y \in \mathbb{R}^n$ such that

$$v_n = \sum_{i=1}^n y_i \varphi_i$$

and obtain

$$\begin{aligned} a(v_n, u_n) &= a\Big(\sum_{i=1}^n y_i\varphi_i, \sum_{j=1}^n x_j\varphi_j\Big) = \sum_{i=1}^n \sum_{j=1}^n y_i a(\varphi_i, \varphi_j) x_j \\ &= \sum_{i=1}^n \sum_{j=1}^n y_i a_{ij} x_j = \sum_{i=1}^n y_i (Ax)_i = \langle y, Ax\rangle_2 = \langle y, b\rangle_2 \\ &= \sum_{i=1}^n y_i b_i = \sum_{i=1}^n y_i \beta(\varphi_i) = \beta\Big(\sum_{i=1}^n y_i\varphi_i\Big) = \beta(v_n). \end{aligned}$$

Since this holds for arbitrary $v_n \in \mathcal{V}_n$, we have proven (5.14).

Let now u_n, as defined in (5.15), be a solution of (5.14), and let $i \in [1:n]$. Due to $\varphi_i \in \mathcal{V}_n$, we have

$$(Ax)_i = \sum_{j=1}^n a_{ij} x_j = \sum_{j=1}^n a(\varphi_i, \varphi_j) x_j = a\Big(\varphi_i, \sum_{j=1}^n x_j \varphi_j\Big) = a(\varphi_i, u_n) = \beta(\varphi_i) = b_i.$$

Since this holds for all $i \in [1:n]$, we have $Ax = b$. ∎

Since $\mathcal{V}_n$ is a Hilbert space, just like $\mathcal{V}$, we can apply the results of the previous section to establish existence and uniqueness of solutions of (5.14).

Corollary 5.33 (Existence and uniqueness) *If there is an $\alpha_n \in \mathbb{R}_{>0}$ such that the bilinear form a satisfies the* discrete inf-sup condition

$$\alpha_n \|u_n\|_{\mathcal{V}} \leq \sup\left\{ \frac{|a(v_n, u_n)|}{\|v_n\|_{\mathcal{V}}} \ : \ v_n \in \mathcal{V}_n \setminus \{0\} \right\} \qquad \text{for all } u_n \in \mathcal{V}_n, \tag{5.16}$$

and if for each $w_n \in \mathcal{V}_n \setminus \{0\}$ there is a $u_n \in \mathcal{V}_n$ such that $a(w_n, u_n) \neq 0$, the discrete variational problem (5.14) has a unique solution.

Proof. Since $\mathcal{V}_n$ is finite-dimensional, the linearity of β already implies $\beta \in \mathcal{V}_n'$. For the same reasons, the bilinear form a is continuous in $\mathcal{V}_n \times \mathcal{V}_n$.

Now we can simply apply Theorem 5.31 to $\mathcal{V}_n$ instead of $\mathcal{V}$. ■

Corollary 5.34 (Coercivity) *If a is coercive, the discrete variational problem (5.14) has a unique solution.*

Proof. Let a be coercive with

$$a(u,u) \geq C_K \|u\|_{\mathcal{V}}^2 \qquad \text{for all } u \in \mathcal{V}$$

for $0 < C_K \leq C_B$.

Let $u_n \in \mathcal{V}_n \setminus \{0\}$. We have

$$C_K \leq \frac{|a(u_n,u_n)|}{\|u_n\|_{\mathcal{V}}\|u_n\|_{\mathcal{V}}},$$

and this immediately implies

$$0 < C_K \leq \sup\left\{ \frac{|a(v_n,u_n)|}{\|v_n\|_{\mathcal{V}}\|u_n\|_{\mathcal{V}}} \;:\; v_n \in \mathcal{V}_n \setminus \{0\} \right\}.$$

We have proven that the discrete inf-sup condition (5.16) holds, and Corollary 5.33 yields existence and uniqueness of the solution. ■

Lemma 5.35 (Matrix properties) *Let A be the matrix defined in Lemma 5.32.*

If the bilinear form a is symmetric, the matrix A is symmetric.

If the bilinear form a is positive definite, the matrix A is positive definite.

If the bilinear form satisfies (5.16), the matrix A is invertible.

Proof. Assume that a is symmetric. We have

$$a_{ij} = a(\varphi_i,\varphi_j) = a(\varphi_j,\varphi_i) = a_{ji} \qquad \text{for all } i,j \in [1:n].$$

Assume that a is positive definite. Let $x \in \mathbb{R}^n \setminus \{0\}$. We define u_n as in (5.15) and observe $u_n \neq 0$, since $(\varphi_i)_{i=1}^n$ is a basis.

We have

$$\begin{aligned}\langle x, Ax\rangle_2 &= \sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j = \sum_{i=1}^n \sum_{j=1}^n x_i a(\varphi_i,\varphi_j) x_j \\ &= a\Big(\sum_{i=1}^n x_i\varphi_i, \sum_{j=1}^n x_j\varphi_j\Big) = a(u_n,u_n) > 0.\end{aligned}$$

Assume that (5.16) holds. Let $x \in \mathbb{R}^n \setminus \{0\}$, and define u_n as in (5.15). Due to (5.16), we find $v_n \in \mathcal{V}_n \setminus \{0\}$ such that

$$|a(v_n,u_n)| \geq C_K \|v_n\|_{\mathcal{V}}\|u_n\|_{\mathcal{V}} > 0.$$

Since $(\varphi_i)_{i=1}^n$ is a basis, we find $y \in \mathbb{R}^n \setminus \{0\}$ such that

$$v_n = \sum_{i=1}^n y_i \varphi_i$$

and obtain

$$|\langle y, Ax\rangle_2| = \Bigl|\sum_{i=1}^n \sum_{j=1}^n y_i a_{ij} x_j\Bigr| = \left|a\Bigl(\sum_{i=1}^n y_i \varphi_i, \sum_{j=1}^n x_j \varphi_j\Bigr)\right| = |a(v_n, u_n)| > 0,$$

which implies $Ax \neq 0$. Contraposition yields that $Ax = 0$ implies $x = 0$, i.e., A is injective and therefore invertible. ∎

Of course, we are also interested in finding estimates for the accuracy of the approximate solution u_n provided by (5.14). A key property of Galerkin discretization methods is the *Galerkin orthogonality.*

Lemma 5.36 (Galerkin orthogonality) *Let $u \in \mathcal{V}$ be a solution of (5.9), and let $u_n \in \mathcal{V}_n$ be a solution of (5.14). We have*

$$a(v_n, u - u_n) = 0 \qquad\qquad \textit{for all } v_n \in \mathcal{V}_n. \tag{5.17}$$

Proof. Let $v_n \in \mathcal{V}_n$. We have

$$a(v_n, u - u_n) = a(v_n, u) - a(v_n, u_n) = \beta(v_n) - \beta(v_n) = 0$$

due to (5.9) and (5.14). ∎

Galerkin orthogonality allows us to compare the discretization error $u - u_n$ to *any* approximation error $u - \widetilde{u}_n$ for $\widetilde{u}_n \in \mathcal{V}_n$. The standard result is that u_n is “almost as good” as the best possible approximation of u.

Depending on the properties of the bilinear form, we can obtain different estimates for the discretization error.

Theorem 5.37 (Discretization error) *Let a be continuous with the continuity constant $C_B \in \mathbb{R}_{\geq 0}$ and let the discrete inf-sup condition (5.16) hold with the constant $\alpha_n \in \mathbb{R}_{>0}$.*

Let $u \in \mathcal{V}$ and $u_n \in \mathcal{V}_n$ be solutions of (5.9) and (5.14), respectively. We have

$$\|u - u_n\|_{\mathcal{V}} \leq \left(1 + \frac{C_B}{\alpha_n}\right) \|u - \widetilde{u}_n\|_{\mathcal{V}} \qquad\qquad \textit{for all } \widetilde{u}_n \in \mathcal{V}_n.$$

Proof. Let $\widetilde{u}_n \in \mathcal{V}_n$. With the triangle inequality, (5.16), the Galerkin orthogonality, and the continuity, we find

$$\begin{aligned}\|u - u_n\|_{\mathcal{V}} &\leq \|u - \widetilde{u}_n\|_{\mathcal{V}} + \|\widetilde{u}_n - u_n\|_{\mathcal{V}} \\ &\leq \|u - \widetilde{u}_n\|_{\mathcal{V}} + \frac{1}{\alpha_n} \sup\left\{\frac{|a(v_n, \widetilde{u}_n - u_n)|}{\|v_n\|_{\mathcal{V}}} \ :\ v_n \in \mathcal{V}_n \setminus \{0\}\right\}\end{aligned}$$

$$\begin{aligned}
&= \|u - \widetilde{u}_n\|_{\mathcal{V}} + \frac{1}{\alpha_n} \sup \left\{ \frac{|a(v_n, \widetilde{u}_n - u)|}{\|v_n\|_{\mathcal{V}}} \ : \ v_n \in \mathcal{V}_n \setminus \{0\} \right\} \\
&\leq \|u - \widetilde{u}_n\|_{\mathcal{V}} + \frac{1}{\alpha_n} \sup \left\{ \frac{C_B \|v_n\|_{\mathcal{V}} \|u - \widetilde{u}_n\|_{\mathcal{V}}}{\|v_n\|_{\mathcal{V}}} \ : \ v_n \in \mathcal{V}_n \setminus \{0\} \right\} \\
&= \|u - \widetilde{u}_n\|_{\mathcal{V}} + \frac{C_B}{\alpha_n} \|u - \widetilde{u}_n\|_{\mathcal{V}},
\end{aligned}$$

and this is the desired estimate. ■

Lemma 5.38 (Céa's lemma, general case) *Let a be coercive with the continuity constant $C_B \in \mathbb{R}_{\geq 0}$ and the coercivity constant $C_K \in \mathbb{R}_{>0}$.*

Let $u \in \mathcal{V}$ and $u_n \in \mathcal{V}_n$ be solutions of (5.9) and (5.14), respectively. We have

$$\|u - u_n\|_{\mathcal{V}} \leq \frac{C_B}{C_K} \|u - \widetilde{u}_n\|_{\mathcal{V}} \qquad \text{for all } \widetilde{u}_n \in \mathcal{V}_n.$$

Proof. Let $\widetilde{u}_n \in \mathcal{V}_n$. Using the Galerkin orthogonality (5.17), we obtain

$$\begin{aligned}
\|u - u_n\|_{\mathcal{V}}^2 &\leq \frac{1}{C_K} a(u - u_n, u - u_n) = \frac{1}{C_K} \big(a(u - u_n, u - u_n) + a(u_n - \widetilde{u}_n, u - u_n) \big) \\
&= \frac{1}{C_K} a(u - \widetilde{u}_n, u - u_n) \leq \frac{C_B}{C_K} \|u - \widetilde{u}_n\|_{\mathcal{V}} \|u - u_n\|_{\mathcal{V}},
\end{aligned}$$

and dividing by $\|u - u_n\|_{\mathcal{V}}$ yields our estimate. ■

We can obtain an improved result if the bilinear form a is symmetric.

Lemma 5.39 (Energy norm) *Let a be symmetric and coercive with the continuity constant $C_B \in \mathbb{R}_{\geq 0}$ and the coercivity constant $C_K \in \mathbb{R}_{>0}$. The* energy norm *is defined by*

$$\|u\|_A := \sqrt{a(u, u)} \qquad \text{for all } u \in \mathcal{V}.$$

It satisfies

$$\sqrt{C_K} \|u\|_{\mathcal{V}} \leq \|u\|_A \leq \sqrt{C_B} \|u\|_{\mathcal{V}} \qquad \text{for all } u \in \mathcal{V},$$

i.e., it is equivalent to the norm $\|\cdot\|_{\mathcal{V}}$.

Proof. Since a is symmetric and coercive, it is an inner product for $\mathcal{V}$, so the energy norm is indeed a norm.

Let $u \in \mathcal{V}$. We have

$$C_K \|u\|_{\mathcal{V}}^2 \leq a(u, u) = \|u\|_A^2 = a(u, u) \leq C_B \|u\|_{\mathcal{V}}^2,$$

and taking the square roots yields the equivalence of the norms. ■

Lemma 5.40 (Céa's lemma, symmetric case) *Let a be symmetric and coercive with the continuity constant $C_B \in \mathbb{R}_{\geq 0}$ and the coercivity constant $C_K \in \mathbb{R}_{>0}$.*

Let $u \in \mathcal{V}$ and $u_n \in \mathcal{V}_n$ be solutions of (5.9) and (5.14), respectively. We have

$$\begin{aligned} \|u - u_n\|_A &\leq \|u - \widetilde{u}_n\|_A, \\ \|u - u_n\|_{\mathcal{V}} &\leq \sqrt{\frac{C_B}{C_K}} \|u - \widetilde{u}_n\|_{\mathcal{V}} \qquad \text{for all } \widetilde{u}_n \in \mathcal{V}_n. \end{aligned}$$

Proof. (cf. [3]) Since a is symmetric and coercive, it is an inner product for the Hilbert space $\mathcal{V}$, and we have the Cauchy-Schwarz inequality (2.21) at our disposal.

Let $\widetilde{u}_n \in \mathcal{V}_n$. Using the Galerkin orthogonality (5.17) and the Cauchy-Schwarz inequality, we find

$$\|u - u_n\|_A^2 = a(u - u_n, u - u_n) = a(u - \widetilde{u}_n, u - u_n) \leq \|u - \widetilde{u}_n\|_A \|u - u_n\|_A,$$

and dividing by $\|u - u_n\|_A$ yields

$$\|u - u_n\|_A \leq \|u - \widetilde{u}_n\|_A.$$

Dividing by $a(u - u_n, u - u_n)^{1/2}$ and squaring yields

$$a(u - u_n, u - u_n) \leq a(u - \widetilde{u}_n, u - \widetilde{u}_n).$$

Lemma 5.39 gives us the second estimate. ∎

6. Finite element methods

The idea of Galerkin's discretization technique is to replace the Hilbert space $\mathcal{V}$ underlying the variational problem by a finite-dimensional subspace $\mathcal{V}_n$. If we choose a basis of $\mathcal{V}_n$, the variational problem is equivalent to a linear system of equations, and this system inherits many important properties like symmetry or coercivity from the original problem.

Now we consider how we can construct finite-dimensional subspaces and choose bases in a way that allows us to handle the resulting linear systems efficiently.

6.1. Triangulations

Before we can start to construct a space of functions, we first have to find a description of the domain of these functions. Our approach is to describe polygons or polyhedra as disjoint unions of triangles or tetrahedra following certain rules.

Definition 6.1 (Simplex) *Let $d \in \mathbb{N}$ and $k \in [0 : d]$. A set $t \subseteq \mathbb{R}^d$ of cardinality $k + 1$ is called a* k-dimensional vertex set in d-dimensional space *if there is a $w \in t$ such that*

$$\{v - w \ : \ v \in t \setminus \{w\}\}$$

is linearly independent.

The set of all k-dimensional vertex sets in d-dimensional space is denoted by S_k^d.

For all $t \in S_k^d$, the sets

$$\omega_t := \Big\{ \sum_{v \in t} \alpha_v v \ : \ \sum_{v \in t} \alpha_v = 1, \ \forall v \in t \ : \ \alpha_v \in \mathbb{R}_{>0} \Big\},$$

$$\bar{\omega}_t := \Big\{ \sum_{v \in t} \alpha_v v \ : \ \sum_{v \in t} \alpha_v = 1, \ \forall v \in t \ : \ \alpha_v \in \mathbb{R}_{\geq 0} \Big\},$$

are called the corresponding open and closed simplices*.*

Two-dimensional simplices are called triangles*, three-dimensional simplices are called* tetrahedra*.*

Lemma 6.2 (Linear independence) *Let $d \in \mathbb{N}$, $k \in [0 : d]$, and $t \in S_k^d$. Let $w \in t$. Then*

$$\{v - w \ : \ v \in t \setminus \{w\}\} \tag{6.1}$$

is linearly independent.

For all $s \subseteq t$ we have $s \in S_\ell^d$ with $\ell := \#s \leq k$.

Proof. By definition we find $\widehat{w} \in t$ such that

$$\{v - \widehat{w} \ : \ v \in t \setminus \{\widehat{w}\}\} \tag{6.2}$$

is linearly independent. Let $\alpha \in \mathbb{R}^t$. We have

$$\begin{aligned} \sum_{v \in t \setminus \{\widehat{w}\}} \alpha_v (v - \widehat{w}) &= \sum_{v \in t \setminus \{\widehat{w}\}} \alpha_v (v - w) - \sum_{v \in t \setminus \{\widehat{w}\}} \alpha_v (\widehat{w} - w) \\ &= \sum_{v \in t \setminus \{w, \widehat{w}\}} \alpha_v (v - w) - \sum_{v \in t \setminus \{\widehat{w}\}} \alpha_v (\widehat{w} - w) = \sum_{v \in t \setminus \{w\}} \beta_v (v - w) \end{aligned}$$

with

$$\beta_{\widehat{w}} := - \sum_{v \in t \setminus \{\widehat{w}\}} \alpha_v, \qquad \beta_v := \alpha_v \qquad \text{for all } v \in t \setminus \{w, \widehat{w}\}.$$

We conclude that the span of (6.1) contains the span of (6.2). Since the latter is k-dimensional, (6.1) has to be linearly independent. ■

Lemma 6.3 (Barycentric coordinates) *Let $d \in \mathbb{N}$, $k \in [0 : d]$, $t \in S_k^d$, and*

$$F_t := \Big\{ \sum_{v \in t} \alpha_v v \ : \ \sum_{v \in t} \alpha_v = 1, \ \forall v \in t \ : \ \alpha_v \in \mathbb{R} \Big\}.$$

There are unique mappings $(\lambda_{t,v})_{v \in t}$ such that

$$x = \sum_{v \in t} \lambda_{t,v}(x) v, \qquad \sum_{v \in t} \lambda_{t,v}(x) = 1 \qquad \text{for all } x \in F_t.$$

For a point $x \in F_t$, the vector $(\lambda_{t,v}(x))_{v \in t}$ is called the vector of barycentric coordinates *of x with respect to t.*

For $v \in t$, $n \in \mathbb{N}$, and $x_1, \ldots, x_n \in F_t$ we have

$$\lambda_{t,v}\left(\sum_{i=1}^n \beta_i x_i \right) = \sum_{i=1}^n \beta_i \lambda_{t,v}(x_i) \qquad \text{for all } \beta_1, \ldots, \beta_n \in \mathbb{R} \text{ with } \sum_{i=1}^n \beta_i = 1. \tag{6.3}$$

Proof. Let $x \in F_t$. By definition, we can find $(\alpha_v)_{v \in t}$ with

$$\sum_{v \in t} \alpha_v v = x, \qquad \sum_{v \in t} \alpha_v = 1.$$

We would like to define $\lambda_{t,v}(x) := \alpha_v$, but this is only admissible if the coefficients are uniquely determined by x.

We fix a second family $(\beta_v)_{v \in t}$ of coefficients also satisfying

$$\sum_{v \in t} \beta_v v = x, \qquad \sum_{v \in t} \beta_v = 1,$$

and have to prove $\alpha_v = \beta_v$ for all $v \in t$.

By definition, we can find $w \in t$ such that

$$\{v - w \ : \ v \in t \setminus \{w\}\}$$

is linearly independent. We have

$$\begin{aligned} x &= \sum_{v \in t} \alpha_v v = \alpha_w w + \sum_{v \in t \setminus \{w\}} \alpha_v v = \Big(1 - \sum_{v \in t \setminus \{w\}} \alpha_v\Big) w + \sum_{v \in t \setminus \{w\}} \alpha_v v \\ &= w + \sum_{v \in t \setminus \{w\}} \alpha_v (v - w), \\ x &= w + \sum_{v \in t \setminus \{w\}} \beta_v (v - w). \end{aligned}$$

Subtracting both equations yields

$$0 = \sum_{v \in t \setminus \{w\}} (\alpha_v - \beta_v)(v - w),$$

and linear independence yields $\alpha_v = \beta_v$ for all $v \in t \setminus \{w\}$. Due to

$$\alpha_w = 1 - \sum_{v \in t \setminus \{w\}} \alpha_v = 1 - \sum_{v \in t \setminus \{w\}} \beta_v = \beta_w,$$

we have proven uniqueness, so the mappings $\lambda_{t,v}$ are well-defined.

Let $n \in \mathbb{N}$, $x_1, \ldots, x_n \in F_t$, and $\beta_1, \ldots, \beta_n \in \mathbb{R}$ with

$$\sum_{i=1}^n \beta_i = 1.$$

We have

$$\begin{aligned} x := \sum_{i=1}^n \beta_i x_i = \sum_{i=1}^n \beta_i \sum_{v \in t} \lambda_{t,v}(x_i) v = \sum_{v \in t} \left(\sum_{i=1}^n \beta_i \lambda_{t,v}(x_i) \right) v, \\ 1 = \sum_{i=1}^n \beta_i = \sum_{i=1}^n \beta_i \sum_{v \in t} \lambda_{t,v}(x_i) \sum_{v \in t} \left(\sum_{i=1}^n \beta_i \lambda_{t,v}(x_i) \right). \end{aligned}$$

Since barycentric coordinates are unique, we conclude

$$\lambda_{t,v}\left(\sum_{i=1}^n \beta_i x_i \right) = \lambda_{t,v}(x) = \sum_{i=1}^n \beta_i \lambda_{t,v}(x_i) \qquad \text{for all } v \in t.$$

This is (6.3). ∎

Lemma 6.4 (Representation of barycentric coordinates) *Let d, k, t, F_t and the barycentric coordinates be given as in Lemma 6.3. Let $k > 0$. Let $v, w \in t$ with $v \neq w$, and let $z \in F_t - w$ with*

$$\langle z, u - w\rangle_2 = 0 \qquad \text{for all } u \in t \setminus \{v, w\}.$$

Then we have $\langle z, v - w\rangle_2 \neq 0$ and

$$\lambda_{t,v}(x) = \frac{\langle z, x - w\rangle_2}{\langle z, v - w\rangle_2} \qquad \text{for all } x \in F_t.$$

Proof. We first address the existence of z with the stated properties. Due to Lemma 6.2, the set $\{u - w \ : \ u \in t \setminus \{w\}\}$ is a basis of a k-dimensional subspace $\mathcal{V} \subseteq \mathbb{R}^d$, therefore $\{u - w \ : \ u \in t \setminus \{v, w\}\}$ is a basis of a $(k-1)$-dimensional subspace $\mathcal{W} \subseteq \mathcal{V}$. We can choose $z \in \mathcal{V} \setminus \mathcal{W}$ to be orthogonal with respect to $\mathcal{W}$.

$\langle z, v - w\rangle_2 = 0$ is impossible, since this would imply that $z \in \mathcal{V} \setminus \{0\}$ is perpendicular on the entire space $\mathcal{V}$ spanned by $\mathcal{W}$ and $v - w$.

Let now $x \in F_t$. We have

$$x = \sum_{u \in t} \lambda_{t,u}(x) u, \qquad 1 = \sum_{u \in t} \lambda_{t,u}(x)$$

by definition and therefore

$$\begin{aligned} x - w &= \lambda_{t,w}(x) w - w + \sum_{u \in t \setminus \{w\}} \lambda_{t,u}(x) u \\ &= \left(1 - \sum_{u \in t \setminus \{w\}} \lambda_{t,u}(x)\right) w - w + \sum_{u \in t \setminus \{w\}} \lambda_{t,u}(x) u \\ &= \sum_{u \in t \setminus \{w\}} \lambda_{t,u}(x)(u - w). \end{aligned}$$

Due to our choice of z, we find

$$\langle z, x - w\rangle_2 = \sum_{u \in t \setminus \{w\}} \lambda_{t,u}(x) \langle z, u - w\rangle_2 = \lambda_{t,v}(x) \langle z, v - w\rangle_2.$$

Dividing by $\langle z, v - w\rangle_2$ yields the desired equation. ∎

Definition 6.5 (Triangulation) *Let $\Omega \subseteq \mathbb{R}^d$. A finite set $T \subseteq S_d^d$ is called a* triangulation *of Ω if*

$$\bar\omega_t \cap \bar\omega_s = \bar\omega_{t \cap s} \qquad \text{for all } t, s \in T, \tag{6.4a}$$

$$\bar\Omega = \bigcup_{t \in t} \bar\omega_t. \tag{6.4b}$$

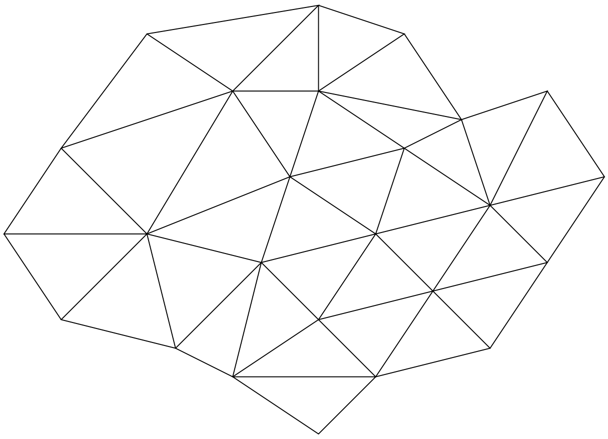

Figure 6.1.: Triangulation of a two-dimensional domain

Lemma 6.6 (Disjoint simplices) *Let $\Omega \subseteq \mathbb{R}^d$, and let T be a triangulation of Ω. If there are $t, s \in T$ with $\omega_t \cap \omega_s \neq \emptyset$, we have $t = s$.*

Proof. Let $t, s \in T$ with $\omega_t \cap \omega_s \neq \emptyset$. Let $x \in \omega_t \cap \omega_s$.

Due to (6.4a), we have

$$x \in \bar{\omega}_t \cap \bar{\omega}_s = \bar{\omega}_{t \cap s}.$$

We define

$$\alpha_v := \begin{cases} \lambda_{t\cap s,v}(x) & \text{if } v \in t \cap s, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } v \in t$$

and have

$$x = \sum_{v \in t\cap s} \alpha_v v = \sum_{v \in t} \alpha_v v.$$

Due to Lemma 6.3, the barycentric coordinates are unique, so $x \in \omega_t$ implies $\alpha_v > 0$ for all $v \in t$. The definition of α_v yields $t \cap s = t$, and with $\#t = d + 1 = \#s$ this already gives us $t = s$. ■

Lemma 6.7 (Neighbouring simplices) *Let $\Omega \subseteq \mathbb{R}^d$, and let T be a triangulation of Ω. Let $t, s, r \in T$. If $t \cap s = t \cap r$ and $\#(t \cap s) = d$, we have $s = r$.*

This means that t can share a face or an edge with at most one other element of the triangulation.

Proof. Let $t \cap s = t \cap r$ and $\#(t \cap s) = d$.

We first consider the case $d > 1$. In this case, we can find $x \in \omega_{t\cap s}$. This implies $\lambda_{t\cap s,v}(x) > 0$ for all $v \in t \cap s$.

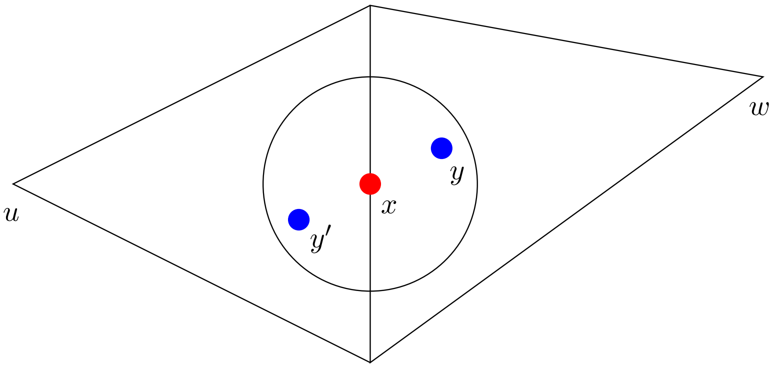

Figure 6.2.: Neighbouring simplices in the proof of Lemma 6.7

Since $\#t = d + 1$ and $\#(t \cap s) = d$, we can find $u \in t \setminus s$ such that $t = \{u\} \cup (t \cap s)$. By the same argument, we can also find $w \in s \setminus t$ such that $s = \{w\} \cup (t \cap s)$. Due to $x \in \omega_{t\cap s}$, the uniqueness of barycentric coordinates yields $\lambda_{t,u}(x) = 0 = \lambda_{s,w}(x)$ and

$$\lambda_{t,v}(x) = \lambda_{t\cap s,v}(x) > 0, \qquad \lambda_{s,v}(x) = \lambda_{t\cap s,v}(x) > 0 \qquad \text{for all } v \in t \cap s.$$

Since the barycentric coordinates are continuous due to Lemma 6.4, we can find an $\epsilon \in \mathbb{R}_{>0}$ such that

$$\|y - x\|_2 < \epsilon \Longrightarrow \forall v \in t \cap s \ : \ \lambda_{t,v}(y) > 0 \wedge \lambda_{s,v}(y) > 0 \qquad \text{for all } y \in \mathbb{R}^d. \tag{6.5}$$

Let $y \in \mathbb{R}^d$ with $\|y - x\|_2 < \epsilon$. If $\lambda_{t,u}(y) > 0$ holds, (6.5) yields $y \in \omega_t$. If $\lambda_{t,u}(y) = 0$ holds, we have $y \in \omega_{t\cap s}$. If $\lambda_{t,u}(y) < 0$ holds, we let $y' := x - (y - x)$ and use (6.3) to find that

$$\frac{y + y'}{2} = \frac{y + x - y + x}{2} = x$$

gives us

$$0 = \lambda_{t,u}(x) = \lambda_{t,u}\left(\frac{y + y'}{2}\right) = \frac{1}{2}\big(\lambda_{t,u}(y) + \lambda_{t,u}(y')\big),$$

i.e., $\lambda_{t,u}(y') > 0$, so (6.5) again gives us $y' \in \omega_t$. Due to Lemma 6.6, this implies $y' \notin \omega_s$ and therefore $\lambda_{s,w}(y') < 0$. We use (6.3) again to get

$$0 = \lambda_{s,w}(x) = \lambda_{s,w}\left(\frac{y + y'}{2}\right) = \frac{1}{2}\big(\lambda_{s,w}(y) + \lambda_{s,w}(y')\big),$$

i.e., $\lambda_{s,w}(y) > 0$ and therefore $y \in \omega_s$. To summarize, we have

$$\begin{cases} y \in \omega_t & \text{if } \lambda_{t,u}(y) > 0, \\ y \in \omega_{t\cap s} & \text{if } \lambda_{t,u}(y) = 0, \\ y \in \omega_s & \text{if } \lambda_{t,u}(y) < 0. \end{cases}$$

We have $x \in \bar{\omega}_{t\cap s} = \bar{\omega}_{t\cap r} \subseteq \bar{\omega}_r$, so any ball centered at x has to intersect ω_r. This means that we can find $y \in \omega_r$ such that $\|y - x\|_2 < \epsilon$. We cannot have $\lambda_{t,u}(y) = 0$, since this would imply $y \in \bar{\omega}_{t\cap s} \not\subseteq \omega_r$.

We cannot have $\lambda_{t,u}(y) > 0$, since this would imply $y \in \omega_t \cap \omega_r$, and Lemma 6.6 would yield $t = r$, although we have $\#(t \cap r) = d < d + 1 = \#t$.

This leaves only $\lambda_{t,u}(y) < 0$, i.e., $y \in \omega_s$. Lemma 6.6 yields $r = s$. ∎

6.2. Piecewise polynomials

Given a triangulation that describes the domain Ω, we can now investigate suitable discrete spaces $\mathcal{V}_n$ that may be used in Galerkin's method.

A simple approach would be to use polynomials. We define multidimensional monomials by

$$x^\nu := x_1^{\nu_1} \dots x_d^{\nu_d} \qquad\qquad \text{for all } \nu \in \mathbb{N}_0^d,\ x \in \mathbb{R}^d$$

and introduce the spaces

$$\Pi_m^d := \Big\{ x \mapsto \sum_{\substack{\nu \in \mathbb{N}_0^d \\ |\nu| \leq m}} \alpha_\nu x^\nu \ :\ \alpha_\nu \in \mathbb{R} \text{ for all } \nu \in \mathbb{N}_0^d,\ |\nu| \leq m \Big\} \qquad \text{for all } d \in \mathbb{N},\ m \in \mathbb{N}_0$$

of d-dimensional polynomials of m-th degree.

A first approach could be to use

$$\mathcal{V}_n = \{ p|_\Omega \ :\ p \in \Pi_m^d \}$$

for a suitable degree $m \in \mathbb{N}_0$. According to Theorem 5.37, we can only expect to be able to approximate solutions u of the variational problem that are close to polynomials, i.e., “almost” infinitely differentiable. This would be a severe limitation of the resulting discretization.

A better approach is to use *piecewise* polynomials, i.e., to fix a triangulation T and consider the space

$$\Pi_{T,m}^d := \{ u \in L^2(\Omega) \ :\ u|_{\omega_t} \in \Pi_m^d \text{ for all } t \in T \}$$

of square-integrable functions that are polynomials on each simplex of the triangulation. By definition, this is a subspace of $L^2(\Omega)$. For our variational problem, we need a subspace of $H_0^1(\Omega)$, and it is possible to prove that $\Pi_{T,m}^d$ is *not* a subspace of $H^1(\Omega)$.

This is due to the fact that an element of $\Pi_{T,m}^d$ can have “jumps” at the boundaries of the simplices ω_t. Our goal is to prove that if we can get rid of these jumps, we obtain a subspace of $H^1(\Omega)$.

Let T be a triangulation of Ω. We define the *set of faces* of a simplex

$$\mathcal{E}_t := \{ e \ :\ e \subseteq t,\ \#e = d \} \subseteq S_{d-1}^d$$

and the set of faces of the entire triangulation

$$\mathcal{E}_T := \{e \; : \; e \in \mathcal{E}_t, \; t \in T\} \subseteq S^d_{d-1}.$$

For $d = 2$, the elements of $\mathcal{E}_t$ correspond to the edges between triangles. For $d = 3$, they correspond to the triangular faces of tetrahedra. The boundary of a simplex is given by

$$\partial\omega_t = \bigcup_{e \in \mathcal{E}_t} \bar{\omega}_e \qquad \text{for all } t \in T.$$

We denote the outer unit normal vector of ω_t by

$$n_t : \partial\omega_t \to \mathbb{R}^d.$$

Due to Lemma 6.7, there can be at most two $t \in T$ with $e \subseteq t$.

A face $e \in \mathcal{E}_T$ is called a *boundary face* if there is exactly one $t \in T$ such that $e \subseteq t$. A face $e \in \mathcal{E}_T$ is a boundary face if and only if $\omega_e \subseteq \partial\Omega$ holds.

For each $e \in \mathcal{E}_t$, we fix a unit normal vector $n_e \in \mathbb{R}^d$. If e is a boundary face, we require n_e to be the outer normal vector.

If $e \in \mathcal{E}_T$ is not a boundary face, the normal vector n_e has to be an *outer* normal vector for one of the two simplices sharing e as a face, i.e., there is exactly one simplex $t \in T$ such that $e \subseteq t$ and $n_t|_{\omega_e} = n_e$. We denote this simplex by $t_{e,+}$.

There is exactly one other simplex $t \in T$ with $e \subseteq t$, and for this simplex, we have $n_t|_{\omega_e} = -n_e$. We denote this simplex by $t_{e,-}$.

Theorem 6.8 (Continuous piecewise polynomials) *Let $m \in \mathbb{N}_0$ and denote by*

$$\mathcal{P}_{T,m} := \{u \in C(\bar{\Omega}) \; : \; u \in \Pi^d_{T,m}\}$$

the continuous piecewise polynomials of degree m.

We have $\mathcal{P}_{T,m} \subseteq H^1(\Omega)$ and

$$(\partial_\nu u)|_{\omega_t} = \partial_\nu(u|_{\omega_t}) \qquad \textit{for all } t \in T, \; \nu \in \mathbb{N}^d_0, \; |\nu| = 1.$$

Proof. Let $u \in \mathcal{P}_{T,m}$ and $\nu \in [1 : d]$. Due to $u \in \Pi^d_{T,m}$, we can find polynomials $u_t \in \Pi^d_m$ such that

$$u|_{\omega_t} = u_t|_{\omega_t} \qquad \text{for all } t \in T.$$

Our candidate for a weak derivative is the function $v \in L^2(\Omega)$ given by

$$v|_{\omega_t} := \frac{\partial u_t}{\partial x_\nu}|_{\omega_t} \qquad \text{for all } t \in T.$$

We have to verify that (5.6) holds. Let $\varphi \in C^\infty_0(\Omega)$. Using partial integration (cf. Reminder 4.3), we find

$$\int_\Omega \frac{\partial\varphi}{\partial x_\nu}(x)u(x)\,dx = \sum_{t \in T} \int_{\omega_t} \frac{\partial\varphi}{\partial x_\nu}(x)u_t(x)\,dx$$

$$
\begin{aligned}
&= \sum_{t\in T} \int_{\partial\omega_t} n_{t,\nu}(x)\varphi(x)u_t(x)\,dx - \int_{\omega_t} \varphi(x)\frac{\partial u_t}{\partial x_\nu}(x)\,dx \\
&= \sum_{t\in T}\sum_{e\in\mathcal{E}_t} \int_{\omega_e} n_{t,\nu}(x)\varphi(x)u_t(x)\,dx - \int_\Omega \varphi(x)v(x)\,dx
\end{aligned}
$$

Due to $\varphi|_{\partial\Omega} = 0$, we can discard all integrals for boundary edges and get

$$
\begin{aligned}
\sum_{t\in T}\sum_{e\in\mathcal{E}_t} \int_{\omega_e} n_{t,\nu}(x)\varphi(x)u_t(x)\,dx &= \sum_{e\in\mathcal{E}_T}\sum_{\substack{t\in T\\ e\subseteq t}} \int_{\omega_e} n_{t,\nu}(x)\varphi(x)u_t(x)\,dx \\
&= \sum_{\substack{e\in\mathcal{E}_T\\ \omega_e\not\subseteq\partial\Omega}}\sum_{\substack{t\in T\\ e\subseteq t}} \int_{\omega_e} n_{t,\nu}(x)\varphi(x)u_t(x)\,dx \\
&= \sum_{\substack{e\in\mathcal{E}_t\\ \omega_e\not\subseteq\partial\Omega}} \int_{\omega_e} n_{e,\nu}\varphi(x)(u_{t_{e,+}}(x) - u_{t_{e,-}}(x))\,dx.
\end{aligned}
$$

Since u is continuous, we have

$$u_{t_{e,+}}|_{\omega_e} = u_{t_{e,-}}|_{\omega_e}$$

and conclude

$$\sum_{\substack{e\in\mathcal{E}_t\\ \omega_e\not\subseteq\partial\Omega}} \int_{\omega_e} n_{e,\nu}\varphi(x)(u_{t_{e,+}}(x) - u_{t_{e,-}}(x))\,dx = 0.$$

This implies

$$\int_\Omega \frac{\partial}{\partial x_\nu}\varphi(x)u(x)\,dx = -\int_\Omega \varphi(x)v(x)\,dx,$$

so v is indeed the weak derivative of u. ■

The space $\mathcal{P}_{T,0}$ is not of interest to us, since a continuous piecewise constant polynomial is just constant. If we take our boundary conditions into account, only the zero function would remain.

The space $\mathcal{P}_{T,1}$, on the other hand, is very useful and probably the most frequently used finite element space. Its popularity is largely due to the fact that we can construct a very convenient basis.

Theorem 6.9 (Barycentric basis) *Let $t \in S_d^d$. Then we have $F_t = \mathbb{R}^d$, and the barycentric coordinates $(\lambda_{t,v})_{v\in t}$ are a basis of Π_1^d with*

$$\lambda_{t,v}(w) = \begin{cases} 1 & \text{if } w = v, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } v, w \in t, \tag{6.6a}$$

$$p = \sum_{v\in t} p(v)\lambda_{t,v} \qquad \text{for all } p \in \Pi_1^d. \tag{6.6b}$$

Proof. Due to Definition 6.1, we find $w \in t$ such that

$$\{v - w \ : \ v \in t \setminus \{w\}\}$$

is a basis of $\mathbb{R}^d$. Let $x \in \mathbb{R}^d$, and let $(\beta_v)_{v \in t \setminus \{w\}}$ be chosen such that

$$x - w = \sum_{v \in t \setminus \{w\}} \beta_v (v - w).$$

We find

$$\begin{aligned} x &= w + \sum_{v \in t \setminus \{w\}} \beta_v (v - w) = w + \sum_{v \in t \setminus \{w\}} \beta_v v - \Bigl(\sum_{v \in t \setminus \{w\}} \beta_v \Bigr) \\ &= \underbrace{\Bigl(1 - \sum_{v \in t \setminus \{w\}} \beta_v \Bigr)}_{=:\alpha_w} w + \sum_{v \in t \setminus \{w\}} \underbrace{\beta_v}_{=:\alpha_v} v = \sum_{v \in t} \alpha_v v, \end{aligned}$$

and conclude $x \in F_t$, i.e., $F_t = \mathbb{R}^d$.

Let now $v \in t$. Lemma 6.4 yields that $\lambda_{t,v}$ is a linear polynomial.

Due to Lemma 6.3, we have (6.6a). Let $\alpha \in \mathbb{R}^d$, and let

$$p := \sum_{v \in t} \alpha_v \lambda_{t,v}.$$

Due to (6.6a), we have

$$\alpha_w = \sum_{v \in t} \alpha_v \lambda_{t,v}(w) = p(w) \qquad \text{for all } w \in t.$$

This is (6.6b), and since $p = 0$ implies $\alpha = 0$, we also obtain that $(\lambda_{t,v})_{v \in t}$ is linearly independent. Its span is a subspace of Π_1^d of dimension $\#t = d + 1$, and since $d + 1$ is also the dimension of Π_1^d, we have established that the barycentric coordinates are indeed a basis. ∎

This Theorem allows us to characterize a function $u \in \mathcal{P}_{T,1}$ entirely by its values in the vertices of the simplices: let $t \in T$, and let $\widetilde{u} \in \mathcal{P}_{T,1}$ be another function such that

$$u(v) = \widetilde{u}(v) \qquad \text{for all } v \in t.$$

Due to Theorem 6.9, we have $u|_{\omega_t} = \widetilde{u}|_{\omega_t}$. If u and $\widetilde{u}$ have identical values in *all* vertices, they have to be identical.

Now let us consider the reverse question: given values in all vertices, can we find a function $u \in \mathcal{P}_{T,1}$ that takes these values? Theorem 6.9 allows us to define a function in Π_1^d, but in order to ensure continuity, we have to extend the result.

Corollary 6.10 (Local representation) *We have*

$$p(x) = \sum_{v \in t \cap s} p(v) \lambda_{t \cap s, v}(x) \qquad \textit{for all } p \in \Pi_1^d, \ t, s \in T, \ x \in \bar{\omega}_{t \cap s}.$$

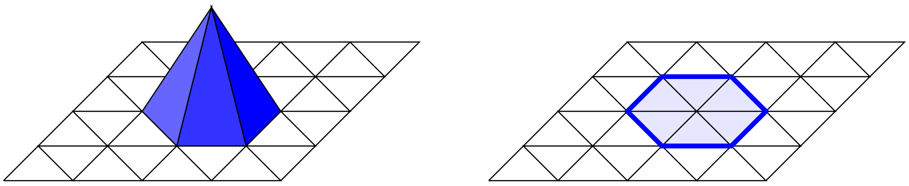

Figure 6.3.: Nodal basis function (left) and its support (right) for a two-dimensional triangulation

Proof. Let $p \in \Pi_1^d$, $t, s \in T$, and $x \in \bar{\omega}_{t \cap s}$.

Due to Theorem 6.9, we have

$$p = \sum_{v \in t} p(v) \lambda_{t,v}.$$

Due to Lemma 6.3 and $x \in \bar{\omega}_{t \cap s}$, we have

$$\lambda_{t,v}(x) = \begin{cases} \lambda_{t \cap s,v}(x) & \text{if } v \in t \cap s, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } v \in t,$$

and combining both equations yields

$$p(x) = \sum_{v \in t} p(v) \lambda_{t,v}(x) = \sum_{v \in t \cap s} p(v) \lambda_{t \cap s,v}(x).$$

■

Definition 6.11 (Nodal basis) *We denote the* set of nodes *of the triangulation T by*

$$\mathcal{N}_T := \bigcup \{t \ : \ t \in T\}.$$

For each $v \in \mathcal{N}_T$ we define $\varphi_v \in \mathcal{P}_{T,1}$ by

$$\varphi_v|_{\bar{\omega}_t} = \begin{cases} \lambda_{t,v}|_{\bar{\omega}_t} & \textit{if } v \in t, \\ 0 & \textit{otherwise} \end{cases} \qquad \textit{for all } t \in T. \tag{6.7}$$

The set $(\varphi_v)_{v \in \mathcal{N}_T}$ is called the nodal basis *of $\mathcal{P}_{T,1}$.*

Lemma 6.12 (Nodal basis) *We have $\varphi_v \in \mathcal{P}_{T,1}$ for all $v \in \mathcal{N}_T$. $(\varphi_v)_{v \in \mathcal{N}_T}$ is a basis of $\mathcal{P}_{T,1}$ satisfying*

$$\varphi_v(w) = \begin{cases} 1 & \textit{if } w = v, \\ 0 & \textit{otherwise} \end{cases} \qquad \textit{for all } v, w \in \mathcal{N}_T, \tag{6.8a}$$

$$u = \sum_{v \in \mathcal{N}_T} u(v) \varphi_v \qquad \textit{for all } u \in \mathcal{P}_{T,1}. \tag{6.8b}$$

Proof. Let $v \in \mathcal{N}_T$. By definition, we have $\varphi_v \in \Pi_1^d$.

We have to prove that φ_v is well-defined and continuous. Let $t, s \in T$ with $v \in t$ and $\bar{\omega}_t \cap \bar{\omega}_s \neq \emptyset$. Due to Definition 6.5, we have $\bar{\omega}_{t\cap s} = \bar{\omega}_t \cap \bar{\omega}_s \neq \emptyset$, and this implies $t \cap s \neq \emptyset$.

If $v \in t \cap s$, Corollary 6.10 yields

$$\lambda_{t,v}|_{\bar{\omega}_{t\cap s}} = \lambda_{t\cap s,v}|_{\bar{\omega}_{t\cap s}} = \lambda_{s,v}|_{\bar{\omega}_{t\cap s}}.$$

If $v \in t \setminus s$, the uniqueness of the barycentric coordinates (cf. Lemma 6.3) yields

$$\lambda_{t,v}|_{\bar{\omega}_{t\cap s}} = 0.$$

We conclude that φ_v is well-defined and continuous.

To prove (6.8a), let $v, w \in \mathcal{N}_T$. By definition, we find $t \in T$ with $v \in t$. If $w \not\in t$, Definition 6.11 immediately yields $\varphi_v(w) = 0$. Otherwise, we have $v, w \in t$ and (6.6a) yields (6.8a).

To prove that $\{\varphi_v \ : \ v \in \mathcal{N}_T\}$ is linearly independent, let $\alpha \in \mathbb{R}^{\mathcal{N}_T}$ and

$$u := \sum_{v\in\mathcal{N}_T} \alpha_v \varphi_v.$$

Due to (6.8a), we have

$$\alpha_w = \sum_{v\in\mathcal{N}_T} \alpha_v \varphi_v(w) = u(w) \qquad \text{for all } w \in \mathcal{N}_T. \tag{6.9}$$

In particular, $u = 0$ implies $\alpha = 0$, so the nodal basis functions are linearly independent.

Let now $u \in \mathcal{P}_{T,1}$. We define

$$\alpha_v := u(v) \qquad \text{for all } v \in \mathcal{N}_T,$$

and (6.6b) yields

$$u|_{\omega_t} = \sum_{v\in t} u(v)\lambda_{t,v}|_{\omega_t} = \sum_{v\in t} \alpha_v\varphi_v|_{\omega_t} = \sum_{v\in\mathcal{N}_T} \alpha_v\varphi_v|_{\omega_t} \qquad \text{for all } t \in T,$$

and therefore

$$u = \sum_{v\in\mathcal{N}_T} \alpha_v \varphi_v.$$

We conclude that the nodal basis spans $\mathcal{P}_{T,1}$, and (6.9) gives us (6.8b). ■

For our model problem, we require a finite-dimensional subspace of $H_0^1(\Omega)$, so we have to include our boundary condition. Since a function in $\mathcal{P}_{T,1}$ can only be non-zero on the boundary if it is non-zero in at least one vertex on the boundary, we can include the boundary condition by discarding all boundary vertices. This leads to the subspace

$$\mathcal{V}_n := \operatorname{span}\{\varphi_i \ : \ i \in \mathcal{I}\} = \{u \in \mathcal{P}_{T,1} \ : \ u|_{\partial\Omega} = 0\}, \qquad \mathcal{I} := \{i \in \mathcal{N}_T \ : \ i \not\in \partial\Omega\}.$$

6.3. Assembly of the linear system

We have already seen in Lemma 5.32 that finding the solution $u_n \in \mathcal{V}_n$ of the discretized variational problem

$$a(v_n, u_n) = \beta(v_n) \qquad\qquad \text{for all } v_n \in \mathcal{V}_n$$

is equivalent to solving the linear system

$$Ax = b$$

with $A \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ and $b \in \mathbb{R}^{\mathcal{I}}$ given by

$$a_{ij} = a(\varphi_i, \varphi_j), \qquad b_i = \beta(\varphi_i) \qquad \text{for all } i, j \in \mathcal{I}.$$

In the case of nodal basis functions, we have

$$\begin{aligned} a_{ij} &= \int_\Omega \langle \nabla\varphi_i(x), \nabla\varphi_j(x)\rangle_2 \, dx = \sum_{t\in T} \int_{\omega_t} \langle \nabla\varphi_i(x), \nabla\varphi_j(x)\rangle_2 \, dx, \\ b_i &= \int_\Omega \varphi_i(x) f(x) \, dx = \sum_{t\in T} \int_{\omega_t} \varphi_i(x) f(x) \, dx \qquad \text{for all } i, j \in \mathcal{I}. \end{aligned}$$

By our definition, $\varphi_i|_{\omega_t} \neq 0$ holds if and only if $i \in t$, so we can elimininate most of the simplices and obtain

$$\begin{aligned} a_{ij} &= \sum_{\substack{t\in T\\ i,j\in t}} \int_{\omega_t} \langle \nabla\varphi_i(x), \nabla\varphi_j(x)\rangle_2 \, dx, \\ b_i &= \sum_{\substack{t\in T\\ i\in t}} \int_{\omega_t} \varphi_i(x) f(x) \, dx \qquad\qquad \text{for all } i, j \in \mathcal{I}. \end{aligned}$$

In theory, we could evaluate the entries of A and b by these equations, but it would be challenging to obtain an efficient implementation: in order to avoid quadratic complexity for A, we would have to ensure that for each $i \in \mathcal{I}$ we can quickly find all $t \in T$ with $i \in t$, e.g., by keeping suitable lists.

A far more elegant way is to assemble the matrix and the vector *incrementally*: we start with a zero matrix and a zero vector and then add the contributions of the individual simplices associated with $t \in T$.

Definition 6.13 (Element matrix and vector) *Let $t \in T$. The matrix $A_t \in \mathbb{R}^{t\times t}$ given by*

$$a_{t,ij} := \int_{\omega_t} \langle \nabla\varphi_i(x), \nabla\varphi_j(x)\rangle_2 \, dx \qquad\qquad \textit{for all } i, j \in t$$

is called the element matrix *for t.*

The vector $b_t \in \mathbb{R}^t$ given by

$$b_{t,i} := \int_{\omega_t} \varphi_i(x) f(x)\,dx \qquad \text{for all } i \in t$$

is called the element vector *for t.*

We find

$$a_{ij} = \sum_{\substack{t\in T\\ i,j\in t}} a_{t,ij}, \qquad b_i = \sum_{\substack{t\in T\\ i\in t}} b_{t,i} \qquad \text{for all } i,j \in \mathcal{I}$$

and perform the *assembly* of the matrix A and the vector b by the following algorithm:

```
procedure assemble;
$A \leftarrow 0$;
$b \leftarrow 0$;
for $t \in T$ do begin
  Compute $A_t$ and $b_t$;
  for $i, j \in t$ do
    $a_{ij} \leftarrow a_{ij} + a_{t,ij}$;
  for $i \in t$ do
    $b_i \leftarrow b_i + b_{t,i}$
end
```

This is a very elegant approach: we compute only the entries we need (with the exception of a small number of boundary nodes), and we never touch entries of the matrix that correspond to indices not sharing the same simplex.

Unless the function f has very special properties, we may not be able to evalute $b_{t,i}$ directly. We can avoid this problem by using a quadrature formula.

Lemma 6.14 (Edge midpoint quadrature) *Let $t \in S_2^d$ with $d \geq 2$. We denote the midpoint of the edge opposite the vertex $v \in t$ by*

$$m_v := \frac{1}{2} \sum_{w \in t \setminus \{v\}} w \qquad \text{for all } v \in t.$$

The edge midpoint quadrature rule *is given by*

$$\mathcal{Q}_t \colon C(\bar\omega_t) \to \mathbb{R}, \qquad u \mapsto \frac{|\omega_t|}{3} \sum_{v\in t} u(m_v),$$

where $|\omega_t|$ denotes the Lebesgue measure of the set ω_t.

Since the weights are positive, we obtain the optimal stability estimate

$$|\mathcal{Q}_t(u)| \leq |\omega_t|\, \|u\|_{\infty,\omega_t} \qquad \text{for all } u \in C(\bar\omega_t).$$

We also find

$$\mathcal{Q}_t(\lambda_{t,v}) = \int_{\omega_t} \lambda_{t,v}(x)\,dx, \qquad \mathcal{Q}_t(\lambda_{t,v}\lambda_{t,w}) = \int_{\omega_t} \lambda_{t,v}(x)\lambda_{t,w}(x)\,dx \qquad \text{for all } v, w \in t,$$

and this implies that all quadratic polynomials are integrated exactly by $\mathcal{Q}_t$.

Proof. Left to the reader as an exercise. It may be a good idea to transform to a simple triangle like $\hat{t} = \{(0,0),(1,0),(0,1)\}$ to evaluate the exact integrals. ■

If we use the edge midpoint quadrature rule to approximate $b_{t,i}$, we can take advantage of the uniqueness of barycentric coordinates to obtain

$$\lambda_{t,w}(m_v) = \begin{cases} 1/2 & \text{if } w \neq v, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } v, w \in t,$$

i.e., evaluating the nodal basis functions $\varphi_i|_{\omega_t} = \lambda_{t,i}|_{\omega_t}$ in the edge midpoints is particularly simple.

In order to compute A_t, we require the gradients of the basis functions. A simple approach can be based on the determinant: let $i \in \mathcal{I}$ and $t \in T$ with $i \in t$. By definition, we have $\varphi_i|_{\omega_t} = \lambda_{t,i}|_{\omega_t}$. Let $t = \{v_0, \ldots, v_d\}$, where $v_0 = v$, and consider

$$\mu(x) := \frac{\det(x - v_1, v_2 - v_1, \ldots, v_d - v_1)}{\det(v_0 - v_1, v_2 - v_1, \ldots, v_d - v_1)} \qquad \text{for all } x \in \mathbb{R}^d.$$

Due to Lemma 6.2, the denominator is non-zero, so μ is well-defined, and we have $\mu(v_0) = 1$. Since the determinant is multilinear, we have $\mu(v_1) = 0$ and μ is a linear polynomial. Since the determinant is alternating, we have $\mu(v_\ell) = 0$ for all $\ell \in [2 : d]$.

This means that μ and $\lambda_{t,v}$ coincide in all vertices $v \in t$, and Theorem 6.9 yields $\mu = \lambda_{t,v}$.

For $d = 2$, μ is of the form

$$\begin{aligned} \mu(x) &= \frac{\det(x - v_1, v_2 - v_1)}{\det(v_0 - v_1, v_2 - v_1)} \\ &= \frac{(x_1 - v_{1,1})(v_{2,2} - v_{1,2}) - (x_2 - v_{1,2})(v_{2,1} - v_{1,1})}{\det(v_0 - v_1, v_2 - v_1)} \\ &= \langle x - v_1, g_v \rangle_2 \qquad \text{for all } x \in \mathbb{R}^d, \end{aligned}$$

where we use

$$g_v := \frac{1}{\det(v_0 - v_1, v_2 - v_1)} \begin{pmatrix} v_{2,2} - v_{1,2} \\ v_{1,1} - v_{2,1} \end{pmatrix}.$$

This representation immediately yields $\nabla\lambda_{t,v} = \nabla\mu = g_v$.

For $d = 3$, we have

$$\mu(x) = \frac{\det(x - v_1, v_2 - v_1, v_3 - v_1)}{\det(v_0 - v_1, v_2 - v_1, v_3 - v_1)}$$

$$= \frac{1}{\det(v_0 - v_1, v_2 - v_1, v_3 - v_1)} \left((x_1 - v_{1,1}) \det \begin{pmatrix} v_{2,2} - v_{1,2} & v_{3,2} - v_{1,2} \\ v_{2,3} - v_{1,3} & v_{3,3} - v_{1,3} \end{pmatrix} \right.$$

$$- (x_2 - v_{1,2}) \det \begin{pmatrix} v_{2,1} - v_{1,1} & v_{3,1} - v_{1,1} \\ v_{2,3} - v_{1,3} & v_{3,3} - v_{1,3} \end{pmatrix}$$

$$\left. + (x_3 - v_{1,3}) \det \begin{pmatrix} v_{2,1} - v_{1,1} & v_{3,1} - v_{1,1} \\ v_{2,2} - v_{1,2} & v_{3,2} - v_{1,2} \end{pmatrix} \right)$$

$$= \frac{\langle x - v_1, (v_2 - v_1) \times (v_3 - v_1) \rangle_2}{\det(v_0 - v_1, v_2 - v_1, v_3 - v_1)} = \langle x - v_1, g_v \rangle \qquad \text{for all } x \in \mathbb{R}^d,$$

where the cross product is defined by

$$a \times b := \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix} \qquad \text{for all } a, b \in \mathbb{R}^3$$

and the vector g_v is given by

$$g_v := \frac{(v_2 - v_1) \times (v_3 - v_1)}{\det(v_0 - v_1, v_2 - v_1, v_3 - v_1)}.$$

For $d > 3$, we can generalize this approach by using Laplace's formula and using cofactors to construct g_v.

Remark 6.15 (Cyclic evaluation) *In order to construct A_t, we require the gradients g_v for all $v \in t$. We can reduce the number of operations by taking advantage of the properties of the determinant: assume that we have already computed $\det(v_0 - v_1, v_2 - v_1, \ldots, v_{d-1} - v_1, v_d - v_1)$ and now have to compute the determinant for cyclically shifted vectors, i.e., $\det(v_1 - v_2, v_3 - v_2, \ldots, v_d - v_2, v_0 - v_2)$. Since the determinant is alternating and multilinear, we can add the first argument $v_1 - v_2$ to all other arguments without changing the result and get*

$$\det(v_1 - v_2, v_3 - v_2, \ldots, v_d - v_2, v_0 - v_2) = \det(v_1 - v_2, v_3 - v_1, \ldots, v_d - v_1, v_0 - v_1).$$

Since the determinant is linear in the first argument, we can change the sign to get

$$\det(v_1 - v_2, v_3 - v_2, \ldots, v_d - v_2, v_0 - v_2) = -\det(v_2 - v_1, v_3 - v_1, \ldots, v_d - v_1, v_0 - v_1).$$

Now we can again use the alternating property to switch the columns $d-1$ and d, then $d-2$ and $d-1$, and so on until we have performed $d-1$ switches and arrive at

$$\det(v_1 - v_2, v_3 - v_2, \ldots, v_d - v_2, v_0 - v_2) = (-1)^d \det(v_0 - v_1, v_2 - v_1, \ldots, v_d - v_1).$$

This means that every cyclic shift of the vertices only changes the sign of the determinant (and its reciprocal) by $(-1)^d$, so we only have to compute it once and just flip the sign appropriately.

6.4. Reference elements

Until now, we have only considered domains that can be split into simplices, e.g., polygons and polyhedra. In order to treat more general domains, we replace the simplices by images of a fixed *reference simplex* under suitable diffeomorphisms. This approach allows us to handle, e.g., curved domains.

Let T be a triangulation of a domain $\Omega \subseteq \mathbb{R}^d$, and let $t \in T$. We enumerate the vertices in t, i.e., we fix $v_0, \ldots, v_d \in t$ such that $t = \{v_0, \ldots, v_d\}$. Due to Lemma 6.2, we know that

$$\{v_1 - v_0, \ldots, v_d - v_0\}$$

is linearly independent, so the matrix

$$F := \begin{pmatrix} v_1 - v_0 & \ldots & v_d - v_0 \end{pmatrix} \in \mathbb{R}^{d \times d}$$

is invertible, and the mapping

$$\Phi_t \colon \mathbb{R}^d \to \mathbb{R}^d, \qquad\qquad \widehat{x} \mapsto v_0 + F\widehat{x},$$

is bijective. We define the *reference simplex* by

$$\widehat{\omega} := \Big\{ \widehat{x} \in \mathbb{R}^d \ : \ \sum_{i=1}^d \widehat{x}_i \leq 1, \ \widehat{x}_i \geq 0 \text{ for all } i \in [1:d] \Big\}$$

and observe

$$\begin{aligned}
\Phi_t(\widehat{\omega}) &= \{v_0 + F\widehat{x} \ : \ \widehat{x} \in \widehat{\omega}\} \\
&= \Big\{ v_0 + \sum_{j=1}^d (v_j - v_0)\widehat{x}_j \ : \ \sum_{i=1}^d \widehat{x}_i \leq 1, \ \widehat{x}_i \in \mathbb{R}_{\geq 0} \text{ for all } i \in [1:d] \Big\} \\
&= \left\{ \Big(1 - \sum_{i=1}^d \widehat{x}_i\Big) v_0 + \sum_{j=1}^d v_j \widehat{x}_j \ : \ 1 - \sum_{i=1}^d \widehat{x}_i \geq 0, \ \widehat{x}_i \in \mathbb{R}_{\geq 0} \text{ for all } i \in [1:d] \right\} \\
&= \Big\{ \sum_{j=0}^d v_j \widehat{x}_j \ : \ \sum_{j=0}^d \widehat{x}_j = 1, \ \widehat{x}_i \in \mathbb{R}_{\geq 0} \text{ for all } i \in [0:d] \Big\} = \bar{\omega}_t,
\end{aligned}$$

where we have introduced $\widehat{x}_0 = 1 - \sum_{j=1}^d \widehat{x}_j$ in the last step.

Remark 6.16 (Barycentric coordinates) *Since the barycentric coordinates for an element $t \in T$ are uniquely determined by*

$$\sum_{v \in t} \lambda_{t,v}(x) v = x, \qquad\qquad \sum_{v \in t} \lambda_{t,v}(x) = 1 \qquad\qquad \textit{for all } x \in \mathbb{R}^d,$$

our equation allows us to compute these coordinates by using

$$\begin{pmatrix} \lambda_{t,v_1}(x) \\ \vdots \\ \lambda_{t,v_d}(x) \end{pmatrix} = \Phi_t^{-1}(x) = F^{-1}(x - v_0), \quad \lambda_{t,v_0}(x) = 1 - \sum_{j=1}^d \lambda_{t,v_j}(x) \quad \textit{for all } x \in \mathbb{R}^d.$$

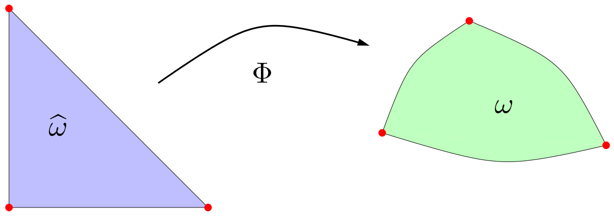

Figure 6.4.: Mapping the reference simplex $\widehat{\omega}$ to ω

Since Φ_t is a bijective diffeomorphism mapping $\widehat{\omega}$ to ω_t, we can use the transformation of variables equation to obtain

$$\int_{\omega_t} \varphi_i(x) f(x)\,dx = \int_{\Phi_t(\widehat{\omega})} \varphi_i(x) f(x)\,dx = \int_{\widehat{\omega}} |\det D\Phi_t(\widehat{x})| \varphi_i(\Phi_t(\widehat{x})) f(\Phi_t(\widehat{x}))\,d\widehat{x}$$

for $i \in t$. We can see that

$$\widehat{\varphi}_{t,i} := \varphi_i \circ \Phi_t$$

is a linear polynomial, since φ_i is a linear polynomial and Φ_t is a linear transformation.

We denote the vertices of the reference simplex by

$$\widehat{v}_0 := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \widehat{v}_1 := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \widehat{v}_2 := \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \qquad \ldots, \qquad \widehat{v}_d := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

introduce the vertex set $\hat{t} := \{\widehat{v}_0, \ldots, \widehat{v}_d\}$, and observe

$$\Phi_t(\widehat{v}_i) = v_i \qquad\qquad \text{for all } i \in [0:d].$$

Due to (6.8a), we have

$$\widehat{\varphi}_{t,v_i}(\widehat{v}_j) = \varphi_{v_i}(\Phi_t(\widehat{v}_j)) = \varphi_{v_i}(v_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i,j \in [0:d],$$

and (6.6b) yields that $\widehat{\varphi}_{t,v_i}$ is the barycentric coordinate $\lambda_{\hat{t},\widehat{v}_i}$ corresponding to the reference simplex.

This means that the nodal basis functions can also be defined by

$$\varphi_i|_{\omega_t} = \begin{cases} \lambda_{\hat{t},\widehat{v}} \circ \Phi_t^{-1} & \text{if } i = \Phi_t(\widehat{v}) \text{ for } \widehat{v} \in \hat{t}, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i \in \mathcal{N}_T,$$

and this definition can be generalized: we assume that a set $(\widehat{\varphi}_i)_{i\in\widehat{\mathcal{I}}}$ of basis functions on $\widehat{\omega}$ is given and that we have a *general* invertible diffeomorphism

$$\Phi : \widehat{\omega} \to \omega$$

with $\omega \subseteq \mathbb{R}^d$. We define *mapped basis functions* by

$$\varphi_i := \widehat{\varphi}_i \circ \Phi^{-1} \qquad \text{for all } i \in \widehat{\mathcal{I}},$$

and consider the element vector and the element matrix given by

$$\begin{aligned} b_{\omega,i} &:= \int_\omega \varphi_i(x) f(x)\,dx && \text{for all } i \in \widehat{\mathcal{I}}, \\ a_{\omega,ij} &:= \int_\omega \langle \nabla\varphi_i(x), \nabla\varphi_j(x)\rangle_2\,dx && \text{for all } i,j \in \widehat{\mathcal{I}}. \end{aligned}$$

By applying our transformation, the element vector can be easily evaluated (or at least approximated) due to

$$\begin{aligned} b_{\omega,i} := \int_\omega \varphi_i(x) f(x)\,dx &= \int_{\widehat{\omega}} |\det D\Phi(\widehat{x})|\, \varphi_i(\Phi(\widehat{x}))\, f(\Phi(\widehat{x}))\,d\widehat{x} \\ &= \int_{\widehat{\omega}} |\det D\Phi(\widehat{x})|\, \widehat{\varphi}_i(\widehat{x})\, f(\Phi(\widehat{x}))\,d\widehat{x} \qquad \text{for all } i \in \widehat{\mathcal{I}} \end{aligned}$$

if we have a suitable quadrature rule for the reference simplex $\widehat{\omega}$ at our disposal and can efficiently evaluate the Jacobi matrix

$$D\Phi(\widehat{x}) = \begin{pmatrix} \frac{\partial\Phi}{\partial\widehat{x}_1}(\widehat{x}) & \dots & \frac{\partial\Phi}{\partial\widehat{x}_d}(\widehat{x}) \end{pmatrix}$$

for all quadrature points.

Evaluating the element matrix is a little more challenging, since it requires the gradients of the mapped basis functions. Due to the chain rule, we have

$$\begin{aligned} \nabla\varphi_i(x) &= D\varphi_i(x)^* = D(\widehat{\varphi}_i \circ \Phi^{-1})(x)^* = (D\widehat{\varphi}_i(\Phi^{-1}(x)) D(\Phi^{-1})(x))^* \\ &= (D\widehat{\varphi}_i(\Phi^{-1}(x)) D\Phi(x)^{-1})^* = (D\Phi(\Phi^{-1}(x))^{-1})^* (D\widehat{\varphi}_i(\Phi^{-1}(x)) \\ &= (D\Phi(\Phi^{-1}(x))^{-1})^* \nabla\widehat{\varphi}_i(\Phi^{-1}(x)) \qquad \text{for all } x \in \omega,\ i \in \widehat{\mathcal{I}}. \end{aligned}$$

If we use the transformation Φ again, we only have to be able to evaluate $\nabla\varphi_i(\Phi(\widehat{x}))$, which is given by

$$\nabla\varphi_i(\Phi(\widehat{x})) = (D\Phi(\widehat{x})^{-1})^* \nabla\widehat{\varphi}_i(\widehat{x}) \qquad \text{for all } \widehat{x} \in \widehat{\omega},\ i \in \widehat{\mathcal{I}}.$$

We can see that we only need the gradient of the basis function $\widehat{\varphi}_i$ on the reference simplex and the transposed matrix of the inverse of $D\Phi(\widehat{x})$, and the latter can be easily computed if we have $D\Phi(\widehat{x})$ at our disposal.

The entries of the element matrix are given by

$$\begin{aligned} a_{\omega,ij} &= \int_\omega \langle \nabla\varphi_i(x), \nabla\varphi_j(x)\rangle_2\,dx \\ &= \int_{\widehat{\omega}} |\det D\Phi(\widehat{x})|\, \langle \nabla\varphi_i(\Phi(\widehat{x})), \nabla\varphi_j(\Phi(\widehat{x}))\rangle_2\,d\widehat{x} \\ &= \int_{\widehat{\omega}} |\det D\Phi(\widehat{x})|\, \langle (D\Phi(\widehat{x})^{-1})^* \nabla\widehat{\varphi}_i(\widehat{x}), (D\Phi(\widehat{x})^{-1})^* \nabla\widehat{\varphi}_j(\widehat{x})\rangle_2\,d\widehat{x} \quad \text{for all } i,j \in \widehat{\mathcal{I}}. \end{aligned}$$

By using a suitable quadrature rule, we only have to be able to evaluate $D\Phi(\widehat{x})$ and $\nabla\widehat{\varphi}_i(\widehat{x})$ in all quadrature points to find an approximation of $a_{\omega,ij}$.

6.5. Averaged Taylor expansion

We have already seen that the error $\|u - u_n\|_V$ introduced by the Galerin discretization can be bounded by $\|u - v_n\|_V$ for all $v_n \in V_n$, so in order to obtain a practical bound, we only have to find *some* element of V_n that approximates u reasonably well.

In the one-dimensional setting, this can be accomplished by using a simple interpolant: we define the interpolation operator

$$\mathfrak{I}_1 \colon C^2[-h,h] \to \Pi_1, \qquad u \mapsto \left(x \mapsto \frac{h-x}{2h} u(-h) + \frac{h+x}{2h} u(h) \right),$$

and observe that it is *stable*, i.e.,

$$\|\mathfrak{I}_1[u]\|_{\infty,[-h,h]} \leq \|u\|_{\infty,[-h,h]} \qquad \text{for all } u \in C^2[-h,h], \tag{6.10}$$

and that it is a *projection*, i.e.,

$$\mathfrak{I}_1[p] = p \qquad \text{for all } p \in \Pi_1. \tag{6.11}$$

In order to obtain an estimate for the interpolation error, we employ the linear Taylor polynomial

$$\mathfrak{T}_1 \colon C^2[-h,h] \to \Pi_1, \qquad u \mapsto \Big(x \mapsto u(0) + x u'(0) \Big),$$

and note that Taylor's theorem allows us to find an $\eta \in [-h,h]$ for each $x \in [-h,h]$ such that

$$u(x) = u(0) + x u'(0) + x^2 \frac{u''(\eta)}{2} = \mathfrak{T}_1[u](x) + x^2 \frac{u''(\eta)}{2}$$

holds, and this implies

$$\|u - \mathfrak{T}_1[u]\|_{\infty,[-h,h]} \leq \frac{h^2}{2} \|u''\|_{\infty,[-h,h]} \qquad \text{for all } u \in C^2[-h,h]. \tag{6.12}$$

Using $\mathfrak{T}_1[u] \in \Pi_1$, (6.11), and (6.10), we find

$$\begin{aligned} \|u - \mathfrak{I}_1[u]\|_{\infty,[-h,h]} &= \|u - \mathfrak{T}_1[u] - \mathfrak{I}_1[u - \mathfrak{T}_1[u]]\|_{\infty,[-h,h]} \\ &\leq \|u - \mathfrak{T}_1[u]\|_{\infty,[-h,h]} + \underbrace{\|\mathfrak{I}_1[u - \mathfrak{T}_1[u]]\|_{\infty,[-h,h]}}_{\leq \|u - \mathfrak{T}_1[u]\|_{\infty,[-h,h]}} \\ &\leq 2\|u - \mathfrak{T}_1[u]\|_{\infty,[-h,h]} \\ &\leq h^2 \|u''\|_{\infty,[-h,h]} \qquad \text{for all } u \in C^2[-h,h]. \end{aligned}$$

In order to approximate by continuous *piecewise* linear polynomials, we interpolate in the vertices of the sub-intervals. This guarantees continuity, and the Taylor error estimate yields an error bound for each sub-interval.

Our goal is now to generalize this approach, first to the multi-dimensional setting with classically differentiable functions, then to weakly differentiable functions.

Let $t \in S_d^d$, and let

$$\mathfrak{I}_t\colon C(\bar\omega_t) \to \Pi_1^d, \qquad u \mapsto \left(x \mapsto \sum_{v\in t} \lambda_{t,v}(x) u(v) \right),$$

denote the *nodal interpolation operator.*

Lemma 6.17 (Nodal interpolation) *We have*

$$\begin{aligned} \|\mathfrak{I}_t[u]\|_{\infty,\bar\omega_t} &\leq \|u\|_{\infty,\bar\omega_t} && \text{for all } u \in C(\bar\omega_t),, && (6.13a)\\ \mathfrak{I}_t[p] &= p && \text{for all } p \in \Pi_1^d. && (6.13b)\end{aligned}$$

Proof. Since $x \in \bar\omega_t$ is equivalent to $\lambda_{t,v}(x) \geq 0$ for all $v \in t$, we have

$$\sum_{v\in t} |\lambda_{t,v}(x)| = \sum_{v\in t} \lambda_{t,v}(x) = 1 \qquad \text{for all } x \in \bar\omega_t.$$

and this implies

$$\begin{aligned} |\mathfrak{I}_t[u](x)| &= \left| \sum_{v\in t} \lambda_{t,v}(x) u(v) \right| \leq \sum_{v\in t} |\lambda_{t,v}(x)|\, |u(v)| \\ &\leq \sum_{v\in t} |\lambda_{t,v}(x)|\, \|u\|_{\infty,\bar\omega_t} = \|u\|_{\infty,\bar\omega_t} \qquad \text{for all } u \in C(\bar\omega_t),\ x \in \bar\omega_t, \end{aligned}$$

and this implies the stability estimate (6.13a). Due to (6.6b), we have

$$\mathfrak{I}_t[p] = \sum_{v\in t} p(v) \lambda_{t,v}(w) = p \qquad \text{for all } p \in \Pi_1^d.$$

This is the projection property (6.13b). ∎

We are left with the task of developing a suitable multi-dimensional counterpart of the Taylor expansion.

Let $\omega \subseteq \mathbb{R}^d$ be a bounded domain, let $x, y \in \omega$ be such that

$$(1-s)y + sx \in \omega \qquad \text{for all } s \in [0,1]$$

holds, and let $u \in C^{m+1}(\omega)$. To construct a Taylor expansion centered at y, we introduce the function

$$f\colon [0,1] \to \mathbb{R}, \qquad s \mapsto u(y + s(x-y)),$$

such that $f(0) = u(y)$ and $f(1) = u(x)$. Applying Taylor's theorem yields

$$u(x) = f(1) = \sum_{\nu=0}^{m} \frac{f^{(\nu)}(0)}{\nu!} + \int_0^1 (1-s)^m \frac{f^{(m+1)}(s)}{m!}\, ds. \tag{6.14}$$

In order to express this equation in terms of the derivatives of the function u, we once again use multi-indices with the notations

$$
\begin{aligned}
&|\nu| = \nu_1 + \nu_2 + \ldots + \nu_d, \\
&\partial_\nu = \frac{\partial^{\nu_1}}{\partial x_1^{\nu_1}} \frac{\partial^{\nu_2}}{\partial x_2^{\nu_2}} \cdots \frac{\partial^{\nu_d}}{\partial x_d^{\nu_d}}, \qquad z^\nu = z_1^{\nu_1} \cdots z_d^{\nu_d}, \\
&\nu! = \nu_1!\, \nu_2! \cdots \nu_d!, \qquad \binom{\nu}{\mu} = \frac{\nu!}{\mu!\,(\nu-\mu)!} \qquad \text{for all } \nu, \mu \in \mathbb{N}_0^d,\ \mu \leq \nu,\ z \in \mathbb{R}^d,
\end{aligned}
$$

where the relation $\nu \leq \mu$ is defined by

$$\nu \leq \mu \iff \forall i \in [1:d] \ :\ \nu_i \leq \mu_i \qquad \text{for all } \nu, \mu \in \mathbb{N}_0^d.$$

Lemma 6.18 (Multi-indices) *We have*

$$(x+y)^\nu = \sum_{\mu \leq \nu} \binom{\nu}{\mu} x^\mu y^{\nu-\mu} \qquad \textit{for all } x, y \in \mathbb{R}^d,\ \nu \in \mathbb{N}_0^d, \tag{6.15a}$$

$$|x^\nu| \leq \|x\|_2^{|\nu|} \qquad \textit{for all } x \in \mathbb{R}^d,\ \nu \in \mathbb{N}_0^d, \tag{6.15b}$$

$$\sum_{\substack{\nu \in \mathbb{N}_0^d \\ |\nu| = m}} \frac{1}{\nu!} = \frac{d^m}{m!} \qquad \textit{for all } m \in \mathbb{N}_0. \tag{6.15c}$$

Proof. We prove (6.15a) by induction for the dimension d.

Base case. If $d = 1$, (6.15a) is just the generalized binomal equation.

Induction assumption. Let $d \in \mathbb{N}$ be such that (6.15a) holds.

Induction step. Let $x, y \in \mathbb{R}^{d+1}$ and $\nu \in \mathbb{N}_0^{d+1}$. We define $\widehat{x} := (x_2, \ldots, x_{d+1})$, $\widehat{y} := (y_2, \ldots, y_{d+1})$, and $\widehat{\nu} := (\nu_2, \ldots, \nu_{d+1})$, and use the induction assumption to get

$$
\begin{aligned}
(x+y)^\nu &= (x_1 + y_1)^{\nu_1} (\widehat{x} + \widehat{y})^{\widehat{\nu}} \\
&= \left(\sum_{\mu_1 = 0}^{\nu_1} \binom{\nu_1}{\mu_1} x_1^{\mu_1} y_1^{\nu_1 - \mu_1} \right) \left(\sum_{\widehat{\mu} \leq \widehat{\nu}} \binom{\widehat{\nu}}{\widehat{\mu}} \widehat{x}^{\widehat{\mu}} \widehat{y}^{\widehat{\nu} - \widehat{\mu}} \right) \\
&= \sum_{(\mu_1, \widehat{\mu}) \leq \nu} \frac{\nu_1!\, \widehat{\nu}!}{\mu_1!\, \widehat{\mu}!\, (\nu_1 - \mu_1)!\, (\widehat{\nu} - \widehat{\mu})!} x^{(\mu_1, \widehat{\mu})} y^{(\nu_1 - \mu_1, \widehat{\nu} - \widehat{\mu})} \\
&= \sum_{\mu \leq \nu} \frac{\nu!}{\mu!\, (\nu - \mu)!} x^\mu y^{\nu - \mu} = \sum_{\mu \leq \nu} \binom{\nu}{\mu} x^\mu y^{\nu - \mu}.
\end{aligned}
$$

The estimate (6.15b) is a direct consequence of $|x_i| \leq \|x\|_2$ for all $i \in [1:d]$.

We prove (6.15c) again by induction for the dimension d.

Base case. If $d = 1$, $|\nu| = m$ implies $\nu = m$.

Induction assumption. Let $d \in \mathbb{N}$ be such that (6.15c) holds.

Induction step. We have

$$\begin{aligned}
\sum_{\substack{\nu\in\mathbb{N}_0^{d+1}\\|\nu|=m}} \frac{1}{\nu!} &= \sum_{\nu_1=0}^{m} \sum_{\substack{\widehat{\nu}\in\mathbb{N}_0^{d}\\|\widehat{\nu}|=m-\nu_1}} \frac{1}{\nu_1!}\frac{1}{\widehat{\nu}!} = \frac{1}{m!}\sum_{\nu_1=0}^{m}\frac{m!}{\nu_1!} \sum_{\substack{\widehat{\nu}\in\mathbb{N}_0^{d}\\|\widehat{\nu}|=m-\nu_1}} \frac{1}{\widehat{\nu}!}\\
&= \frac{1}{m!}\sum_{\nu_1=0}^{m}\frac{m!}{\nu_1!}\frac{d^{m-\nu_1}}{(m-\nu_1)!} = \frac{1}{m!}\sum_{\nu_1=0}^{m}\binom{m}{\nu_1}1^{\nu_1}d^{m-\nu_1}\\
&= \frac{1}{m!}(1+d)^m = \frac{(d+1)^m}{m!}.
\end{aligned}$$

■

Using multi-indices, we can now derive explicit representations of the derivatives of the auxiliary function f in terms of the partial derivatives of u.

Lemma 6.19 (Derivatives) *Let $m \in \mathbb{N}_0$ and $u \in C^m(\omega)$. We have*

$$f^{(m)}(s) = \sum_{|\nu|=m} \frac{m!}{\nu!}(x-y)^\nu \partial_\nu u(y+s(x-y)) \qquad \text{for all } s\in[0,1]. \qquad (6.16)$$

Proof. By induction.
Base case: For $m=0$, the identity is trival.
Induction assumption: Let $m \in \mathbb{N}_0$ be such that (6.16) holds for all $u \in C^m(\omega)$.
Induction step: Let $u \in C^{m+1}(\omega)$, and let $\mu_i \in \{0,1\}^d$ be defined by

$$(\mu_i)_j = \begin{cases} 1 & \text{if } j=i,\\ 0 & \text{otherwise}\end{cases} \qquad \text{for all } i,j \in [1:d].$$

Applying the chain rule to the induction assumption yields

$$\begin{aligned}
f^{(m+1)}(s) &= \sum_{|\nu|=m}\frac{m!}{\nu!}(y-x)^\nu \sum_{i=1}^{d}(y_i-x_i)\partial_{\mu_i}\partial_\nu u(y+s(x-y))\\
&= \sum_{|\nu|=m}\frac{m!}{\nu!}(y-x)^\nu \sum_{i=1}^{d}(y-x)^{\mu_i}\partial_{\nu+\mu_i} u(y+s(x-y))\\
&= \sum_{|\nu|=m}\sum_{i=1}^{d}\frac{m!}{\nu!}(y-x)^{\nu+\mu_i}\partial_{\nu+\mu_i} u(y+s(x-y))\\
&= \sum_{i=1}^{d}\sum_{|\nu|=m}\frac{m!\,(\nu_i+1)}{(\nu+\mu_i)!}(y-x)^{\nu+\mu_i}\partial_{\nu+\mu_i} u(y+s(x-y))\\
&= \sum_{i=1}^{d}\sum_{\substack{|\widehat{\nu}|=m+1\\ \hat{\nu}_i>0}}\frac{m!\,\hat{\nu}_i}{\widehat{\nu}!}(y-x)^{\widehat{\nu}}\partial_{\widehat{\nu}} u(y+s(x-y))
\end{aligned}$$

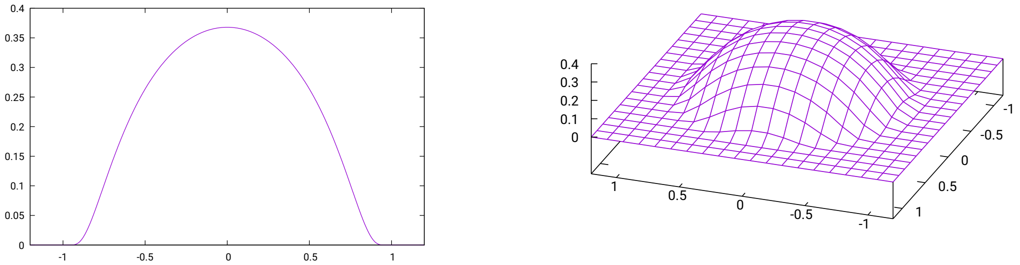

Figure 6.5.: Example for a mollifier function: $\varphi(x) = \exp\left(-\frac{1}{1-x^2}\right)$

$$= \sum_{i=1}^{d} \sum_{|\widehat{\nu}|=m+1} \frac{m!\,\hat{\nu}_i}{\widehat{\nu}!}(y-x)^{\widehat{\nu}} \partial_{\widehat{\nu}} u(y+s(x-y))$$

$$= \sum_{|\widehat{\nu}|=m+1} \frac{(m+1)!}{\widehat{\nu}!}(y-x)^{\widehat{\nu}} \partial_{\widehat{\nu}} u(y+s(x-y)) \qquad \text{for all } s \in [0,1].$$

■

Applying this result to (6.14) yields

$$u(x) = \sum_{|\nu|\leq m} \frac{\partial_\nu u(y)}{\nu!}(x-y)^\nu \tag{6.17a}$$

$$+ (m+1)\int_0^1 (1-s)^m \sum_{|\nu|=m+1} \frac{\partial_\nu u(y+s(x-y))}{\nu!}(x-y)^\nu \, ds \tag{6.17b}$$

Since we are interested in working with *weakly* differentiable functions, we cannot use the classical derivatives of u in x to define an approximation, so instead we use multiple centers of expansion and take an average. Our construction closely follows [1, Chapter 4].

Let $x_0 \in \mathbb{R}^d$ and $r \in \mathbb{R}_{>0}$ and denote the ball of radius r centered at x_0 by

$$\mathcal{B}_{x_0,r} := \{y \in \mathbb{R}^d \;:\; \|y - x_0\|_2 < r\}.$$

Definition 6.20 (Star-shaped domain) *A domain $\omega \subseteq \mathbb{R}^d$ is called* star-shaped with respect to a ball $\mathcal{B}$ *if*

$$(1-s)y + sx \in \omega \qquad \textit{for all } x \in \omega,\; y \in \mathcal{B},\; s \in [0,1].$$

Let $\omega \subseteq \mathbb{R}^d$ be star-shaped with respect to a ball $\mathcal{B} = \mathcal{B}_{x_0,r}$. Then we can apply (6.17) to all expansion points $y \in \mathcal{B}$.

In order to obtain an average, we introduce a *mollifier function*, i.e., a function $\widehat{\varphi} \in C^\infty(\mathbb{R}^d)$ such that

$$\begin{aligned} \widehat{\varphi}(x) &\geq 0 && \text{for all } x \in \mathbb{R}^d, \\ \operatorname{supp}(\widehat{\varphi}) &\subseteq \{y \in \mathbb{R}^d \ : \ \|y\|_2 < 1\}, \\ \int_{\mathbb{R}^d} \widehat{\varphi}(x)\,dx &= 1. \end{aligned}$$

An example is the function

$$\widehat{\varphi}(x) = \begin{cases} \exp\left(-\frac{1}{1-\|x\|_2^2}\right) & \text{if } \|x\|_2 < 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } x \in \mathbb{R}^d,$$

shown in Figure 6.5.

We require a function with support in $\mathcal{B}$, so we shift and scale $\widehat{\varphi}$ to define

$$\varphi\colon \mathbb{R}^d \to \mathbb{R}, \qquad x \mapsto r^{-d}\widehat{\varphi}((x-x_0)/r),$$

and observe $\operatorname{supp}(\varphi) \subseteq \mathcal{B}$ and

$$\int_{\mathbb{R}^d} \varphi(x)\,dx = \int_{\mathbb{R}^d} r^{-d}\widehat{\varphi}((x-x_0)/r)\,dx = \int_{\mathbb{R}^d} \widehat{\varphi}(\widehat{x})\,d\widehat{x} = 1.$$

Using (6.17) gives us

$$\begin{aligned} u(x) &= \int_\omega \varphi(y)u(x)\,dy = \int_{\mathcal{B}} \varphi(y)u(x)\,dy \\ &= \sum_{|\nu|\leq m} \int_{\mathcal{B}} \varphi(y)\frac{\partial_\nu u(y)}{\nu!}(x-y)^\nu\,dy && (6.18a) \\ &+ (m+1)\int_{\mathcal{B}}\int_0^1 \varphi(y)(1-s)^m \sum_{|\nu|=m+1} \frac{\partial_\nu u(y+s(x-y))}{\nu!}(x-y)^\nu\,ds\,dy, && (6.18b) \end{aligned}$$

and the right-hand side only involves integrals of $\partial_\nu u$, but no point evaluations anymore, so we can define the averaged Taylor polynomial by

$$\mathfrak{T}_m\colon H^m(\omega) \to \Pi_m^d, \qquad u \mapsto \left(x \mapsto \sum_{|\nu|\leq m} \int_{\mathcal{B}} \varphi(y)\frac{\partial_\nu u(y)}{\nu!}(x-y)^\nu\,dy\right). \tag{6.19}$$

In order to prove that $\mathfrak{T}_m[u]$ is well-defined, i.e., that the integral on the right-hand side is a polynomial of degree not higher than m, we have to show that we can split the powers $(x-y)^\nu$ into powers of x and powers of y.

Lemma 6.21 (Averaged Taylor polynomial) *Let $u \in H^m(\omega)$. We have*

$$u(x) = \sum_{|\mu|\leq m} a_\mu \frac{(x-x_0)^\mu}{\mu!} \qquad \textit{for all } x \in \mathbb{R}^d,$$

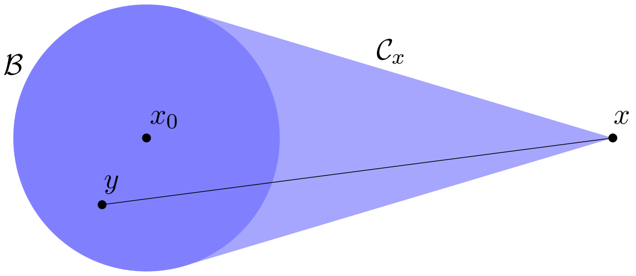

Figure 6.6.: Construction of the averaged Taylor expansion

where the coefficients are given by

$$a_\mu := \sum_{\substack{|\nu|\leq m\\ \mu\leq\nu}} \int_{\mathcal{B}} \varphi(y)\partial_\nu u(y)\frac{(x_0-y)^{\nu-\mu}}{(\nu-\mu)!}\,dy \qquad \textit{for all } \mu\in\mathbb{N}_0^d,\ |\mu|\leq m. \tag{6.20}$$

Proof. (cf. [1, Proposition (4.1.9)]) Let $x, y \in \mathbb{R}^d$ and $\nu \in \mathbb{N}_0^d$. Applying (6.15a) to (6.19) yields

$$\begin{aligned}
\mathfrak{T}_m[u](x) &= \sum_{|\nu|\leq m} \int_{\mathcal{B}} \varphi(y)\frac{\partial_\nu u(y)}{\nu!}(x-x_0+x_0-y)^\nu\,dy \\
&= \sum_{|\nu|\leq m} \int_{\mathcal{B}} \varphi(y)\frac{\partial_\nu u(y)}{\nu!}\sum_{\mu\leq\nu}\binom{\nu}{\mu}(x-x_0)^\mu(x_0-y)^{\nu-\mu}\,dy \\
&= \sum_{|\mu|\leq m}\sum_{\substack{|\nu|\leq m\\ \mu\leq\nu}} \int_{\mathcal{B}} \varphi(y)\frac{\partial_\nu u(y)}{\nu!}\frac{\nu!}{\mu!(\nu-\mu)!}(x_0-y)^{\nu-\mu}\,dy\ (x-x_0)^\mu \\
&= \sum_{|\mu|\leq m}\sum_{\substack{|\nu|\leq m\\ \mu\leq\nu}} \int_{\mathcal{B}} \varphi(y)\partial_\nu u(y)\frac{(x_0-y)^{\nu-\mu}}{(\nu-\mu)!}\,dy\frac{(x-x_0)^\mu}{\mu!} \\
&= \sum_{|\mu|\leq m} a_\mu\frac{(x-x_0)^\mu}{\mu!},
\end{aligned}$$

so the averaged Taylor polynomial is indeed a polynomial in Π_m^d. ■

A closer look at (6.20) reveals that we can define averaged Taylor polynomials even for $u \in L^2(\omega)$ by using partial integration.

Lemma 6.22 (Generalization) *We define*

$$\varphi_\mu\colon \mathbb{R}^d\to\mathbb{R}, \qquad y\mapsto\frac{(x_0-y)^\mu}{\mu!}\varphi(x) \qquad \textit{for all } \mu\in\mathbb{N}_0^d.$$

The coefficients (6.20) satisfy

$$a_\mu = \sum_{\substack{|\nu|\leq m\\ \mu\leq\nu}} (-1)^{|\nu|} \int_{\mathcal{B}} \partial_\nu \varphi_{\nu-\mu}(y) u(y)\, dy \qquad \textit{for all } \mu \in \mathbb{N}_0^d,\ |\mu| \leq m,$$

and we can extend $\mathcal{T}_m$ to a linear operator such that we have

$$\|\mathcal{T}_m[u]\|_{\infty,\omega} \leq C_{stab}\|u\|_{L^2} \qquad \textit{for all } u \in L^2(\omega)$$

with a constant $C_{stab} \in \mathbb{R}_{>0}$.

Proof. (cf. [1, Proposition (4.1.12)]) Let $u \in H^m(\omega)$, and let $\mu \in \mathbb{N}_0^d$ with $|\mu| \leq m$. Since $\varphi_{\nu-\mu}$ is in $C_0^\infty(\mathcal{B})$, we can apply partial integration due to Definition 5.4 and get

$$a_\mu = \sum_{\substack{|\nu|\leq m\\ \mu\leq\nu}} \int_{\mathcal{B}} \varphi_{\nu-\mu}(y) \partial_\nu u(y)\, dy = \sum_{\substack{|\nu|\leq m\\ \mu\leq\nu}} (-1)^{|\nu|} \int_{\mathcal{B}} \partial_\nu \varphi_{\nu-\mu}(y) u(y)\, dy,$$

so using the Cauchy-Schwarz inequality (5.4) yields

$$|a_\mu| \leq \sum_{\substack{|\nu|\leq m\\ \mu\leq\nu}} \underbrace{\|\partial_\nu \varphi_{\nu-\mu}\|_{L^2}}_{=:C_\mu} \|u\|_{L^2} \qquad \text{for all } \mu \in \mathbb{N}_0^d,\ |\mu| \leq m.$$

We define

$$\psi_\mu \colon \omega \to \mathbb{R}, \qquad x \mapsto \frac{(x-x_0)^\mu}{\mu!}, \qquad \text{for all } \mu \in \mathbb{N}_0^d,\ |\mu| \leq m.$$

Using Lemma 6.21 and the triangle inequality, we obtain

$$\|\mathfrak{T}_m[u]\|_{\infty,\omega} \leq \sum_{|\mu|\leq m} |a_\mu| \|\psi_\mu\|_{\infty,\omega} \leq \sum_{|\mu|\leq m} C_\mu \|u\|_{L^2} \|\psi_\mu\|_{\infty,\omega} \leq C_{\text{stab}} \|u\|_{L^2}$$

with

$$C_{\text{stab}} := \sum_{|\mu|\leq m} C_\mu \|\psi_\mu\|_{\infty,\omega}.$$

Since $C^\infty(\omega)$ is dense in $H^m(\omega)$ due to Theorem 5.12, $H^m(\omega)$ is dense in $L^2(\omega)$, so we can indeed extend the operator $\mathfrak{T}_m$ continuously to $L^2(\omega)$. ∎

In order to obtain an estimate for the approximation error, we have to consider the remainder (6.18b) given by

$$\mathfrak{R}_m[u](x) := (m+1) \int_{\mathcal{B}} \int_0^1 \varphi(y)(1-s)^m \sum_{|\nu|=m+1} \frac{\partial_\nu u(y+s(x-y))}{\nu!} (x-y)^\nu \, ds\, dx$$
$$\text{for all } u \in H^{m+1}(\omega),\ x \in \omega.$$

We would like to bound this quantity in terms of $\partial_\nu u$, so we have to look for a variable transformation that replaces $y + s(x-y)$ with a new variable z.

Exercise 6.23 (Polar coordinates) *Let $d \in \mathbb{N}$. We define*

$$\widehat{\Phi}_d \colon \mathbb{R}^d \to \mathbb{R}^{d+1}, \quad x \mapsto \begin{cases} \begin{pmatrix} \cos(x_1) \\ \sin(x_1) \end{pmatrix} & \text{if } d = 1, \\ \begin{pmatrix} \cos(x_1) \\ \sin(x_1)\widehat{\Phi}_{d-1}(x_2, \ldots, x_d) \end{pmatrix} & \text{otherwise} \end{cases} \qquad \text{for all } d \in \mathbb{N}$$

and

$$\Phi_d \colon \mathbb{R}^d \to \mathbb{R}^d, \qquad x \mapsto x_1 \widehat{\Phi}_{d-1}(x_2, \ldots, x_d).$$

Prove that $\widehat{\Phi}_d$ maps

$$\Omega_d := \begin{cases} [0, 2\pi) & \text{if } d = 1, \\ [0, \pi)^{d-1} \times [0, 2\pi) & \text{otherwise} \end{cases}$$

bijectively to the unit sphere $\{x \in \mathbb{R}^{d+1} \ : \ \|x\|_2 = 1\}$.

Prove that Φ_d maps $(0, r) \times \Omega_{d-1}$ bijectively to $\{x \in \mathbb{R}^d \ : \ \|x\|_2 \in (0, r)\}$, is differentiable, and satisfies $|\det D\Phi_d(x)| \le x_1^{d-1}$ for all $x \in \mathbb{R}^d$.

Lemma 6.24 (Riesz potential) *Let $p, q \in \mathbb{R}$ with $1 < p, q < \infty$ and $1/p + 1/q = 1$. Let $\Omega \subseteq \mathbb{R}^d$ be a domain and $\alpha \in \mathbb{R}_{<d}$.*

There is a constant $C_{rs} \in \mathbb{R}_{>0}$ depending only on d such that

$$\int_\Omega \frac{1}{\|x - z\|_2^\alpha} \, dx \le C_{rs} \frac{\operatorname{diam}(\Omega)^{d-\alpha}}{d - \alpha} \qquad \text{for all } z \in \Omega$$

and for all $f \in L^p(\Omega)$, the Riesz potential *of f defined by*

$$g(z) := \int_\Omega \frac{f(x)}{\|x - z\|_2^\alpha} \, dx \qquad \text{for all } z \in \Omega$$

satisfies $g \in L^p(\Omega)$ and

$$\|g\|_{L^p} \le C_{rs} \operatorname{diam}(\omega)^{d-\alpha} \|f\|_{L^p}.$$

Proof. (cf. [1, Lemma 4.3.6]) We start by considering

$$\int_\Omega \|x - z\|_2^{-\alpha} \, dx$$

for a given $z \in \Omega$. Let $r := \operatorname{diam}(\Omega)$, and define the ball (without center)

$$\mathcal{C} := \{x \in \mathbb{R}^d \ : \ 0 < \|x - z\|_2 < r\}.$$

For $d = 1$, we obtain

$$\int_\Omega |x - z|^{-\alpha} \, dx \le \int_{(-r,r)\setminus\{0\}} |y|^{-\alpha} \, dy = 2 \int_0^r y^{-\alpha} \, dy = 2 \left[\frac{y^{1-\alpha}}{1 - \alpha} \right]_{y=0}^r = 2 \frac{r^{1-\alpha}}{1 - \alpha}$$

and let $C_{\mathrm{rs}} := 2$. For $d > 1$, we use Exercise 6.23 with $\mathcal{C} = z + \Phi_d((0,r) \times \Omega_{d-1})$, so we can apply a change of variables to find

$$\begin{aligned}
\int_\Omega \|x - z\|_2^{-\alpha} \, dx \leq \int_{\mathcal{C}} \|x - z\|_2^{-\alpha} \, dx &\leq \int_{(0,r)\times\Omega_{d-1}} \widehat{x}_1^{d-1} \|\Phi_d(\widehat{x})\|_2^{-\alpha} \, d\widehat{x} \\
&= \int_{(0,r)\times\Omega_{d-1}} \widehat{x}_1^{d-1} \widehat{x}_1^{-\alpha} \, d\widehat{x} = |\Omega_{d-1}| \int_0^r y^{d-1-\alpha} \, dy.
\end{aligned}$$

We let $\beta := d - 1 - \alpha$, observe $\beta > -1$, and obtain

$$\int_0^r y^\beta \, dy = \left[\frac{y^{\beta+1}}{\beta + 1} \right]_{y=0}^r = \frac{r^{\beta+1}}{\beta + 1} = \frac{r^{d-\alpha}}{d - \alpha},$$

so we may conclude

$$\int_\Omega \|x - z\|_2^{-\alpha} \, dx \leq |\Omega_{d-1}| \frac{r^{d-\alpha}}{d - \alpha} = C_{\mathrm{rs}} \frac{r^{d-\alpha}}{d - \alpha} \qquad \text{for all } z \in \Omega \tag{6.21}$$

with $C_{\mathrm{rs}} := |\Omega_{d-1}|$.

We apply Hölder's inequality to find

$$\begin{aligned}
\|g\|_{L^p(\Omega)}^p &= \int_\Omega |g(z)|^p \, dz = \int_\Omega \left(\int_\Omega |f(x)| \|x - z\|_2^{-\alpha} \, dx \right)^p dz \\
&= \int_\Omega \left(\int_\Omega |f(x)| \|x - z\|_2^{-\alpha/p} \|x - z\|_2^{-\alpha/q} \, dx \right)^p dz \\
&\leq \int_\Omega \left[\left(\int_\Omega |f(x)|^p \|x - z\|_2^{-\alpha} \, dx \right)^{1/p} \left(\int_\Omega \|x - z\|_2^{-\alpha} \, dx \right)^{1/q} \right]^p dz \\
&\leq \left(C_{\mathrm{rs}} \frac{r^{d-\alpha}}{d - \alpha} \right)^{p/q} \int_\Omega \int_\Omega |f(x)|^p \|x - z\|_2^{-\alpha} \, dx \, dz.
\end{aligned}$$

Due to (6.21) and $v \in L^p(\Omega)$, we can apply the Fubini-Tonelli theorem to obtain

$$\begin{aligned}
\|g\|_{L^p(\Omega)}^p &\leq \left(C_{\mathrm{rs}} \frac{r^{d-\alpha}}{d - \alpha} \right)^{p/q} \int_\Omega \int_\Omega \|x - z\|_2^{-\alpha} \, dz |f(x)|^p \, dx \\
&\leq \left(C_{\mathrm{rs}} \frac{r^{d-\alpha}}{d - \alpha} \right)^{p/q+1} \int_\Omega |f(x)|^p \, dx = \left(C_{\mathrm{rs}} \frac{r^{d-\alpha}}{d - \alpha} \right)^p \|f\|_{L^p(\Omega)}^p
\end{aligned}$$

using $p/q + 1 = p(1/q + 1/p) = p$. ■

Lemma 6.25 (Error representation) *Let $u \in H^{m+1}(\omega)$ and $x \in \omega$. Let*

$$\mathcal{C}_x := \{ y + s(x - y) \ : \ y \in \mathcal{B}, \ s \in [0,1] \}.$$

We have

$$\mathfrak{R}_m[u](x) = (m + 1) \sum_{|\nu|=m+1} \int_{\mathcal{C}_x} k_\nu(x, z) \partial_\nu u(z) \, dz$$

with

$$k_\nu(x,z) = \frac{(x-z)^\nu}{\nu!} k(x,z) \qquad \textit{for all } \nu \in \mathbb{N}_0^d,\ |\nu| = m+1$$

for a function k satisfying

$$|k(x,z)| \leq C_{\mathfrak{R}} \left(1 + \frac{\|x - x_0\|_2}{r}\right)^d \|z - x\|^{-d} \qquad \textit{for all } z \in \mathcal{C}_x$$

with $C_{\mathfrak{R}}$ depending only on $\widehat{\varphi}$ and d.

Proof. (cf. [1, Proposition (4.2.8)]) Since it is more convenient to deal with a singularity at $s = 0$ instead of $s = 1$, we first apply the transformation $s \mapsto 1 - s$ and get

$$\mathfrak{R}_m[u](x) = (m+1) \int_{\mathcal{B}} \int_0^1 \varphi(y) s^m \sum_{|\nu|=m+1} \frac{(x-y)^\nu}{\nu!} \partial_\nu u(x + s(y-x))\, ds\, dx$$

due to $y + (1-s)(x-y) = y + (1-s)x - (1-s)y = x + s(y-x)$.

We want to focus on one term of the sum and fix $\nu \in \mathbb{N}_0^d$ with $|\nu| = m+1$.

We define the transformation

$$\Phi\colon \mathbb{R}^d \times (0,1] \to \mathbb{R}^d \times (0,1], \qquad (y,s) \mapsto (x + s(y-x), s).$$

In order to compute its inverse, we let $(z,s) \in \mathbb{R}^d \times (0,1]$ and find

$$\begin{aligned} z = x + s(y-x) \iff z - x = s(y-x) \iff (z-x)/s = y - x \\ \iff (z-x)/s + x = y. \end{aligned}$$

This means that $(y,s) := \Phi^{-1}(z,s) \in \mathcal{B} \times (0,1]$ holds if and only if

$$(z,s) \in \mathcal{A} := \{(z,s) \in \mathbb{R}^d \times (0,1] \ :\ \|(z-x)/s + x - x_0\|_2 < r\}.$$

In this case we have

$$z = x + s(y-x) \in \mathcal{C}_x, \tag{6.22a}$$

$$(x-y)^\nu = (x - (z-x)/s - x)^\nu = s^{-(m+1)}(x-z)^\nu, \tag{6.22b}$$

$$\begin{aligned} s &= \frac{\|z-x\|_2}{\|(z-x)/s\|_2} = \frac{\|z-x\|_2}{\|(z-x)/s - (x_0 - x) + (x_0 - x)\|_2} \\ &\geq \frac{\|z-x\|_2}{\|(z-x)/s + x - x_0\|_2 + \|x_0 - x\|_2} > \frac{\|z-x\|_2}{r + \|x_0 - x\|_2}. \end{aligned} \tag{6.22c}$$

We conclude that

$$\Phi : \mathcal{B} \times (0,1] \to \mathcal{A}$$

is a bijective differentiable mapping with

$$\det D\Phi(y,s) = \det \begin{pmatrix} sI & (y-x) \\ 0 & 1 \end{pmatrix} = s^d \qquad \text{for all } (y,s) \in \mathcal{B} \times (0,1],$$

$$\Phi^{-1}(z,s) = ((z-x)/s + x, s) \qquad\qquad \text{for all } (z,s) \in \mathcal{A},$$

so we can apply a change of variables. Using (6.22b), we find

$$\begin{aligned}
&\int_{\mathcal{B}} \int_0^1 \varphi(y) s^m \frac{(x-y)^\nu}{\nu!} \partial_\nu u(x + s(y-x)) \, ds \, dy \\
&\quad = \int_{\mathcal{B}\times(0,1]} |\det D\Phi(y,s)| \; \varphi(y) s^{m-d} \frac{(x-y)^\nu}{\nu!} \partial_\nu u(\Phi(y,s)) \, d(y,s) \\
&\quad = \int_{\mathcal{A}} \varphi((z-x)/s + x) s^{m-d} s^{-(m+1)} \frac{(x-z)^\nu}{\nu!} \partial_\nu u(z) \, d(z,s) \\
&\quad = \int_{\mathcal{A}} \varphi((z-x)/s + x) s^{-d-1} \frac{(x-z)^\nu}{\nu!} \partial_\nu u(z) \, d(z,s).
\end{aligned}$$

Due to (6.22a), we have

$$\mathcal{A} \subseteq \mathcal{C}_x \times (0,1].$$

We introduce the characteristic function

$$\begin{aligned}
1_{\mathcal{A}} \colon \mathcal{C}_x \times (0,1] &\to \mathbb{R}, \\
(z,s) &\mapsto \begin{cases} 1 & \text{if } (z,s) \in \mathcal{A}, \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}$$

and use Fubini's theorem to get

$$\begin{aligned}
&\int_{\mathcal{B}} \int_0^1 \varphi(y) s^m \frac{(x-y)^\nu}{\nu!} \partial_\nu u(x + s(y-x)) \, ds \, dy \\
&\quad = \int_{\mathcal{C}_x\times(0,1]} 1_{\mathcal{A}}(z,s) \varphi((z-x)/s + x) s^{-d-1} \frac{(x-z)^\nu}{\nu!} \partial_\nu u(z) \, d(z,s) \\
&\quad = \int_{\mathcal{C}_x} \int_0^1 1_{\mathcal{A}}(z,s) \varphi((z-x)/s + x) s^{-d-1} \frac{(x-z)^\nu}{\nu!} \partial_\nu u(z) \, ds \, dz \\
&\quad = \int_{\mathcal{C}_x} \frac{(x-z)^\nu}{\nu!} \partial_\nu u(z) \int_0^1 1_{\mathcal{A}}(z,s) \varphi((z-x)/s + x) s^{-d-1} \, ds \, dz.
\end{aligned}$$

The second term does not depend on u, so we define

$$k(x,z) := \int_0^1 1_{\mathcal{A}}(z,s) \varphi((z-x)/s + x) s^{-d-1} \, ds \qquad\qquad \text{for all } z \in \mathcal{C}_x$$

and obtain

$$\begin{aligned}
&\int_{\mathcal{B}} \int_0^1 \varphi(y) s^m \frac{\partial_\nu u(x + s(y-x))}{\nu!} (x-y)^\nu \, ds \, dy \\
&\quad = \int_{\mathcal{C}_x} k_\nu(x,z) \partial_\nu u(z) \, dz,
\end{aligned}$$

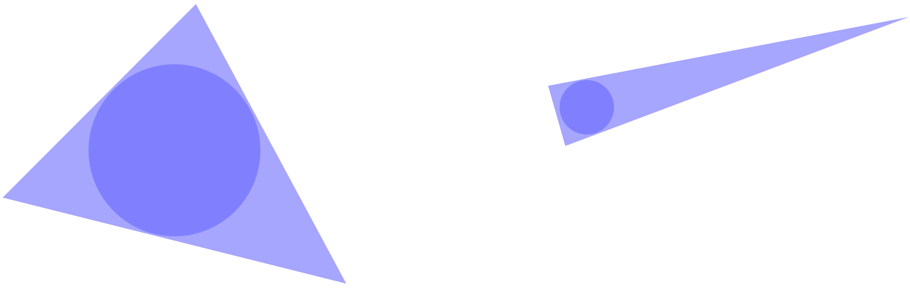

Figure 6.7.: Dependence of the chunkiness parameter on the domain's shape: γ is small in the left domain and large in the right domain

$$\mathfrak{R}_m[u](x) = (m+1) \sum_{|\nu|=m+1} \int_{\mathcal{C}_x} k_\nu(x,z) \partial_\nu u(z) \, dz.$$

To obtain a bound for $k(x,z)$, recall that (6.22c) implies that for $z \in \mathcal{C}_x$ we have

$$s > s_0 := \frac{\|z-x\|_2}{r + \|x_0 - x\|_2} > 0 \qquad \text{for all } s \in \mathbb{R} \text{ with } (z,s) \in \mathcal{A}.$$

We find

$$\begin{aligned} k(x,z) &= \int_0^1 1_{\mathcal{A}}(z,s) \varphi((z-x)/s + x) s^{-d-1} \, ds = \int_{s_0}^1 \varphi((z-x)/s + x) s^{-d-1} \, ds \\ &\leq \|\varphi\|_\infty \left[\frac{s^{-d}}{-d} \right]_{s=s_0}^1 = \|\varphi\|_\infty \left(\frac{s_0^{-d}}{d} - \frac{1}{d} \right) \leq \frac{\|\varphi\|_\infty}{d} s_0^{-d} \\ &= \frac{\|\varphi\|_\infty}{d} (r + \|x_0 - x\|_2)^d \|x - z\|_2^{-d} = \frac{r^{-d} \|\widehat{\varphi}\|_\infty}{d} (r + \|x_0 - x\|_2)^d \|x - z\|_2^{-d} \\ &= \frac{\|\widehat{\varphi}\|_\infty}{d} \left(1 + \frac{\|x_0 - x\|_2}{r} \right)^d \|x - z\|_2^{-d}. \end{aligned}$$

We complete the proof by choosing $C_{\mathfrak{R}} := \|\widehat{\varphi}\|_\infty / d$. ∎

Definition 6.26 (Chunkiness parameter) *Let $\omega \subseteq \mathbb{R}^d$ be a domain. We define*

$$r_{max} := \sup\{r \in \mathbb{R}_{>0} \ : \ \omega \textit{ is star-shaped with respect to } \mathcal{B}_{x_0,r} \textit{ for an } x_0 \in \omega\}$$

(with the convention $\sup \emptyset = -\infty$) and call

$$\gamma := \begin{cases} \frac{\operatorname{diam}(\omega)}{r_{max}} & \textit{if } r_{max} > 0, \\ \infty & \textit{otherwise} \end{cases}$$

the chunkiness parameter *of ω.*

For a convex domain ω, $r_{\max}$ is the radius of the largest ball contained in ω.

Theorem 6.27 (Bramble-Hilbert lemma) *Let $\omega \subseteq \mathbb{R}^d$ be a domain with chunkiness parameter $\gamma < \infty$, and let $m \in \mathbb{N}_0$. If $\mathcal{B}$ is chosen appropriately, there is a constant $C_{bh} \in \mathbb{R}_{>0}$ depending only on d, $\widehat{\varphi}$, and γ, such that*

$$\|u - \mathfrak{T}_m[u]\|_{L^2(\omega)} \leq \frac{C_{bh} d^{m+1}}{(m+1)!} \operatorname{diam}(\omega)^{m+1} |u|_{H^{m+1}(\omega)} \qquad \textit{for all } u \in H^{m+1}(\omega).$$

Proof. (cf. [1, Lemma 4.3.8]) Let $u \in H^{m+1}(\omega)$ and $\delta := \operatorname{diam}(\omega)$. By definition of γ, we can find a ball $\mathcal{B} \subseteq \omega$ of radius $r > \operatorname{diam}(\omega)/(2\gamma)$ such that ω is star-shaped with respect to $\mathcal{B}$. Applying Lemma 6.25 to this ball, we have

$$\begin{aligned}
\|u - \mathfrak{T}_m[u]\|_{L^2(\omega)} = \|\mathfrak{R}_m[u]\|_{L^2(\omega)} &\leq (m+1) \left\| \sum_{|\nu|=m+1} \int_\omega |k_\nu(\cdot, z)| \, |\partial_\nu u(z)| \, dz \right\|_{L^2(\omega)} \\
&\leq (m+1) \sum_{|\nu|=m+1} \left\| \int_\omega C_{\mathfrak{R}} \frac{(1+2\gamma)^d}{\nu!} \|z - \cdot\|_2^{|\nu|-d} |\partial_\nu u(z)| \, dz \right\|_{L^2(\omega)} \\
&\leq C_{\mathfrak{R}} (m+1)(1+2\gamma)^d \sum_{|\nu|=m+1} \left\| \int_\omega \|z - \cdot\|_2^{m+1-d} \frac{|\partial_\nu u(z)|}{\nu!} \, dz \right\|_{L^2(\omega)} \\
&= C_1 (m+1) \sum_{|\nu|=m+1} \left\| \int_\omega \|z - \cdot\|_2^{m+1-d} \frac{|\partial_\nu u(z)|}{\nu!} \, dz \right\|_{L^2(\omega)}
\end{aligned}$$

with $C_1 := C_{\mathfrak{R}}(1+2\gamma)^d$. Now we can use Lemma 6.24 with $\alpha = d - (m+1)$, the Cauchy-Schwarz inequality, and the equation (6.15c) of Lemma 6.18 to obtain

$$\begin{aligned}
\|u - \mathfrak{T}_m[u]\|_{L^2(\omega)} &\leq C_1 (m+1) C_{\mathrm{rs}} \frac{\delta^{m+1}}{m+1} \sum_{|\nu|=m+1} \frac{1}{\nu!} \|\partial_\nu u\|_{L^2} \\
&\leq C_1 C_{\mathrm{rs}} \delta^{m+1} \left(\sum_{|\nu|=m+1} \frac{1}{(\nu!)^2} \right)^{1/2} \left(\sum_{|\nu|=m+1} \|\partial_\nu u\|_{L^2}^2 \right)^{1/2} \\
&\leq C_1 C_{\mathrm{rs}} \delta^{m+1} \left(\sum_{|\nu|=m+1} \frac{1}{\nu!} \right) \left(\sum_{|\nu|=m+1} \|\partial_\nu u\|_{L^2}^2 \right)^{1/2} \\
&= C_1 C_{\mathrm{rs}} \delta^{m+1} \frac{d^{m+1}}{(m+1)!} \left(\sum_{|\nu|=m+1} \|\partial_\nu u\|_{L^2}^2 \right)^{1/2} \\
&= \frac{C_{\mathrm{bh}} d^{m+1}}{(m+1)!} \delta^{m+1} |u|_{H^{m+1}(\omega)},
\end{aligned}$$

where C_{rs} is the constant of Lemma 6.24 and $C_{\mathrm{bh}} := C_1 C_{\mathrm{rs}} = C_{\mathfrak{R}}(1+2\gamma)^d C_{\mathrm{rs}}$. ■

This result allows us to construct approximating polynomials on star-shaped domains, and since a triangulation consists of such domains, we can also construct piecewise polynomial approximations.

Exercise 6.28 (Approximation of derivatives) *Let ω, γ, m be as in Theorem 6.27. Let $\mu \in \mathbb{N}_0^d$ with $|\mu| \leq m$, and let $\ell := m - |\mu|$. Prove*

$$\partial_\mu \mathfrak{T}_m[u] = \mathfrak{T}_\ell[\partial_\mu u] \qquad \text{for all } u \in H^m(\omega).$$

Combine this result with Theorem 6.27 to prove

$$\|\partial_\mu u - \partial_\mu \mathfrak{T}_m[u]\|_{L^2(\omega)} \leq \frac{C_{bh} d^{\ell+1}}{(\ell+1)!} \operatorname{diam}(\omega)^{\ell+1} |u|_{H^{m+1}(\omega)} \qquad \text{for all } u \in H^{m+1}(\omega).$$

In order to obtain the *continuous* piecewise polynomial approximation we require, we can employ interpolation. Unfortunately, standard interpolation relies on the evaluation of the interpolant in the interpolation points, and functions in $L^2(\Omega)$ are only defined up to null sets.

This problem can be solved by the *Sobolev's lemma*: if a function has weak derivatives of sufficiently high order, it is continuous, so evaluation in interpolation points is possible.

Lemma 6.29 (Stability) *Let $\omega \subseteq \mathbb{R}^d$ be a domain with chunkiness parameter $\gamma < 0$, and let $m \in \mathbb{N}$.*

If $\mathcal{B}$ is chosen appropriately, there is a constant $C_{st} \in \mathbb{R}_{>0}$ depending only on $\widehat{\varphi}$, γ, d, and m with

$$\|\mathfrak{T}_m[u]\|_{\infty,\omega} \leq C_{st} \max\{\operatorname{diam}(\omega)^{-d/2}, \operatorname{diam}(\omega)^{m-d/2}\} \|u\|_{H^m(\omega)} \qquad \text{for all } u \in H^m(\omega).$$

Proof. Let $\mathcal{B} \subseteq \omega$ be a ball of radius $r > \operatorname{diam}(\omega)/(2\gamma)$ such that ω is star-shaped with respect to $\mathcal{B}$.

Let $u \in H^m(\omega)$ and $x \in \omega$.

We have

$$\mathfrak{T}_m[u](x) = \sum_{|\nu| \leq m} \int_{\mathcal{B}} \varphi(y) \frac{(x-y)^\nu}{\nu!} \partial_\nu u(y)\, dy$$

by definition. We also have $\|\varphi\|_{\infty,\omega} = r^{-d} \|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}$ by construction.

Let $\nu \in \mathbb{N}_0^d$ with $|\nu| \leq m$. Using the triangle inequality and (6.15b), we obtain

$$\begin{aligned} \left| \int_{\mathcal{B}} \varphi(y) \frac{(x-y)^\nu}{\nu!} \partial_\nu u(y)\, dy \right| &\leq \int_{\mathcal{B}} \varphi(y) \frac{\|x-y\|_2^{|\nu|}}{\nu!} |\partial_\nu u(y)|\, dy \\ &\leq \|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}^{1/2} r^{-d/2} \int_{\mathcal{B}} \sqrt{\varphi(y)} \frac{\operatorname{diam}(\omega)^{|\nu|}}{\nu!} |\partial_\nu u(y)|\, dy \\ &= \|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}^{1/2} r^{-d/2} \operatorname{diam}(\omega)^{|\nu|} \int_{\mathcal{B}} \frac{\sqrt{\varphi(y)}}{\nu!} |\partial_\nu u(y)|\, dy. \end{aligned}$$

Now we can apply the Cauchy-Schwarz inequality (5.4) to get

$$\begin{aligned}\left|\int_{\mathcal{B}} \varphi(y)\frac{(x-y)^\nu}{\nu!}\partial_\nu u(y)\,dy\right| &\leq \|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}^{1/2} r^{-d/2}\operatorname{diam}(\omega)^{|\nu|}\left(\int_{\mathcal{B}}\frac{\varphi(y)}{(\nu!)^2}\,dy\right)^{1/2}\|\partial_\nu u\|_{L^2(\omega)}\\ &= \|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}^{1/2} r^{-d/2}\operatorname{diam}(\omega)^{|\nu|}\frac{1}{\nu!}\|\partial_\nu u\|_{L^2(\omega)}.\end{aligned}$$

Let $\delta := \max\{\operatorname{diam}(\omega)^{-d/2}, \operatorname{diam}(\omega)^{m-d/2}\}$. We have

$$\begin{aligned} r^{-d/2}\operatorname{diam}(\omega)^{|\nu|} &= r^{-d/2}\operatorname{diam}(\omega)^{d/2}\operatorname{diam}(\omega)^{|\nu|-d/2}\\ &\leq r^{-d/2}(2\gamma r)^{d/2}\operatorname{diam}(\omega)^{|\nu|-d/2}\\ &= (2\gamma)^{d/2}\operatorname{diam}(\omega)^{|\nu|-d/2} \leq (2\gamma)^{d/2}\delta,\end{aligned}$$

and due to the Cauchy-Schwarz inequality, we find

$$\begin{aligned} |\mathfrak{T}_m[u](x)| &\leq \sum_{|\nu|\leq m}\|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}^{1/2}(2\gamma)^{d/2}\delta\frac{1}{\nu!}\|\partial_\nu u\|_{L^2(\omega)}\\ &\leq \|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}^{1/2}(2\gamma)^{d/2}\delta\left(\sum_{|\nu|\leq m}\frac{1}{(\nu!)^2}\right)^{1/2}\left(\sum_{|\nu|\leq m}\|\partial_\nu u\|_{L^2(\omega)}^2\right)^{1/2}\\ &\leq \|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}^{1/2}(2\gamma)^{d/2}\delta\left(\sum_{|\nu|\leq m}\frac{1}{\nu!}\right)\|u\|_{H^m(\omega)}.\end{aligned}$$

We let

$$C_{\mathrm{st}} := \|\widehat{\varphi}\|_{\infty,\mathbb{R}^d}^{1/2}(2\gamma)^{d/2}\left(\sum_{|\nu|\leq m}\frac{1}{\nu!}\right)$$

and conclude

$$|\mathfrak{T}_m[u](x)| \leq C_{\mathrm{st}}\delta\|u\|_{H^m(\omega)} \qquad \text{for all } u\in H^m(\omega),\ x\in\omega.$$

This is equivalent to the required estimate. ■

Lemma 6.30 (Maximal approximation error) *Let $\omega \subseteq \mathbb{R}^d$ be a domain with chunkiness parameter $\gamma < 0$, and let $m \in \mathbb{N}$ with $m+1 > d/2$.*

If $\mathcal{B}$ is chosen appropriately, there is a constant $C_{er} \in \mathbb{R}_{>0}$ depending only on $\widehat{\varphi}$, γ, d, and m with

$$\|u - \mathfrak{T}_m[u]\|_{\infty,\omega} \leq C_{er}\operatorname{diam}(\omega)^{m+1-d/2}|u|_{H^{m+1}(\omega)} \qquad \textit{for all } u \in H^{m+1}(\omega).$$

Proof. Let $\mathcal{B} \subseteq \omega$ be a ball of radius $r > \operatorname{diam}(\omega)/(2\gamma)$ such that ω is star-shaped with respect to $\mathcal{B}$.

Let $u \in H^{m+1}(\omega)$ and $x \in \omega$. Denote the diameter of ω by $\delta := \operatorname{diam}(\omega)$.

Lemma 6.25 yields

$$\begin{aligned}
|u(x) - \mathfrak{T}_m[u](x)| &= |\mathfrak{R}_m[u](x)| \\
&\leq (m+1) \sum_{|\nu|=m+1} \int_\omega \frac{\|x-z\|_2^{m+1}}{\nu!} C_{\mathcal{R}} (1+2\gamma)^d \|x-z\|_2^{-d} |\partial_\nu u(z)|\, dz \\
&= C_{\mathfrak{R}} (m+1)(1+2\gamma)^d \sum_{|\nu|=m+1} \int_\omega \|x-z\|_2^{m+1-d} \frac{|\partial_\nu u(z)|}{\nu!}\, dz.
\end{aligned}$$

By our assumption, we have $2(m+1)-d > 0$ and $\alpha := 2d-2(m+1) < d$, and Lemma 6.24 yields

$$\int_\omega \frac{1}{\|x-z\|_2^\alpha}\, dx \leq \frac{C_{\mathrm{rs}}}{d-\alpha} \delta^{d-\alpha},$$

i.e., $\|x - \cdot\|_2^{m+1-d} \in L^2(\omega)$.

Let $\nu \in \mathbb{N}_0^d$ with $|\nu| = m+1$. With the Cauchy-Schwarz inequality (5.4), we obtain

$$\begin{aligned}
\int_\omega \|x-z\|_2^{m+1-d} |\partial_\nu u(z)|\, dz &\leq \left(\int_\omega \|x-z\|_2^{2(m+1-d)}\, dz \right)^{1/2} \left(\int_\omega |\partial_\nu u(z)|^2\, dz \right)^{1/2} \\
&\leq \sqrt{\frac{C_{\mathrm{rs}}}{d-\alpha} \delta^{d-\alpha}} \left(\int_\omega |\partial_\nu u(z)|^2\, dz \right)^{1/2} \\
&= \sqrt{\frac{C_{\mathrm{rs}}}{2(m+1)-d}} \delta^{m+1-d/2} \|\partial_\nu u\|_{L^2(\omega)}.
\end{aligned}$$

Combining the terms yields

$$|u(x) - \mathfrak{T}_m[u](x)| \leq C_{\mathfrak{R}} (m+1)(1+2\gamma)^d \sqrt{\frac{C_{\mathrm{rs}}}{2m+2-d}} \delta^{m+1-d/2} \sum_{|\nu|=m+1} \frac{\|\partial_\nu u\|_{L^2(\omega)}}{\nu!}.$$

We let $C_1 := C_{\mathfrak{R}} (1+2\gamma)^d \sqrt{C_{\mathrm{rs}}/(2m+2-d)}$ and use the Cauchy-Schwarz inequality and (6.15c) to find

$$\begin{aligned}
|u(x) - \mathfrak{T}_m[u](x)| &\leq C_1 (m+1) \delta^{m+1-d/2} \sum_{|\nu|=m+1} \frac{\|\partial_\nu u\|_{L^2(\omega)}}{\nu!} \\
&\leq C_1 (m+1) \delta^{m+1-d/2} \left(\sum_{|\nu|=m+1} \frac{1}{(\nu!)^2} \right)^{1/2} \left(\sum_{|\nu|=m+1} \|\partial_\nu u\|_{L^2(\omega)}^2 \right)^{1/2} \\
&\leq C_1 (m+1) \delta^{m+1-d/2} \left(\sum_{|\nu|=m+1} \frac{1}{\nu!} \right) |u|_{H^{m+1}(\omega)} \\
&= C_1 (m+1) \delta^{m+1-d/2} \frac{d^{m+1}}{(m+1)!} |u|_{H^{m+1}(\omega)}
\end{aligned}$$

$$= C_1 \operatorname{diam}(\omega)^{m+1-d/2} \frac{d^{m+1}}{m!} |u|_{H^{m+1}(\omega)}$$

We complete the proof by choosing

$$C_{\text{er}} := C_1 \frac{d^{m+1}}{m!}$$

and observing that C_1 depends only on $C_{\mathfrak{R}}$, γ, d, and m, while $C_{\mathfrak{R}}$ depends only on $\widehat{\varphi}$ and d. ∎

Remark 6.31 (Maximum norm vs. L^2-norm) *The factor* $\operatorname{diam}(\omega)^{-d/2}$ *in the estimate of Lemma 6.30 might seem a little disappointing, since it means that the estimate will go to infinity if we let the diameter go to zero.*

A closer look suggests that this factor might be necessary: we can find a constant C such that $|\omega| \leq C \operatorname{diam}(\omega)^d$ holds for all domains, e.g., by choosing C as the Lebesgue measure of a ball with radius $\operatorname{diam}(\omega)$*. We have*

$$\|u\|_{L^2(\omega)} \leq |\omega|^{1/2} \|u\|_{\infty,\omega} \leq C \operatorname{diam}(\omega)^{d/2} \|u\|_{\infty,\omega} \qquad \text{for all } u \in C(\omega),$$

so if the maximum norm estimate would not involve the factor $\operatorname{diam}(\omega)^{-d/2}$*, we would get a higher power of* $\operatorname{diam}(\omega)$ *in the Bramble-Hilbert lemma. A comparison with the standard Taylor expansion suggests that this is unrealistic.*

Theorem 6.32 (Sobolev's lemma) *Let $\omega \subseteq \mathbb{R}^d$ be a domain with chunkiness parameter $\gamma < 0$, and let $m \in \mathbb{N}$ with $m > d/2$.*

There is a constant $C_{so} \in \mathbb{R}_{>0}$ depending only on $\widehat{\varphi}$, γ, d, and m such that all functions $u \in H^m(\Omega)$ are in $u \in C(\Omega)$ with

$$\|u\|_{\infty,\omega} \leq C_{so} \delta \|u\|_{H^m(\omega)},$$

where $\delta := \max\{\operatorname{diam}(\omega)^{-d/2}, \operatorname{diam}(\omega)^{m-d/2}\}$.

Proof. (cf. [1, Lemma 4.3.4]) We first prove this result for $u \in C^\infty(\omega)$. Let $\omega \subseteq \mathbb{R}^d$ be star-shaped with respect to a ball $\mathcal{B} \subseteq \omega$ of radius $r > \operatorname{diam}(\omega)/(2\gamma)$. The triangle inequality yields

$$\|u\|_{\infty,\omega} \leq \|\mathfrak{T}_{m-1}[u]\|_{\infty,\omega} - \|u_n - \mathfrak{T}_{m-1}[u]\|_{\infty,\omega}.$$

For the first term, Lemma 6.29 yields

$$\begin{aligned} \|\mathfrak{T}_{m-1}[u]\|_{\infty,\omega} &\leq C_{\text{st}} \max\{\operatorname{diam}(\omega)^{-d/2}, \operatorname{diam}(\omega)^{m-1-d/2}\} \|u\|_{H^{m-1}(\omega)} \\ &\leq C_{\text{st}} \max\{\operatorname{diam}(\omega)^{-d/2}, \operatorname{diam}(\omega)^{m-d/2}\} \|u\|_{H^m(\omega)}. \end{aligned}$$

With $\delta := \max\{\operatorname{diam}(\omega)^{-d/2}, \operatorname{diam}(\omega)^{m-d/2}\}$, we can write this estimate in the short form

$$\|\mathfrak{T}_{m-1}[u]\|_{\infty,\omega} \leq C_{\text{st}} \delta \|u\|_{H^m(\omega)}.$$

For the second term, Lemma 6.30 yields

$$\begin{aligned}\|u - \mathfrak{I}_{m-1}[u]\|_{\infty,\omega} &\leq C_{\mathrm{er}}\delta |u|_{H^m(\omega)} \\ &\leq C_{\mathrm{er}}\delta \|u\|_{H^m(\omega)}.\end{aligned}$$

Combining both estimates yields

$$\begin{aligned}\|u\|_{\infty,\omega} &\leq C_{\mathrm{st}}\delta \|u\|_{H^m(\omega)} + C_{\mathrm{er}}\delta \|u\|_{H^m(\omega)} \\ &\leq (C_{\mathrm{st}} + C_{\mathrm{er}})\delta \|u\|_{H^m(\omega)} = C_{\mathrm{so}}\delta \|u\|_{H^m(\omega)}\end{aligned}$$

with

$$C_{\mathrm{so}} := C_{\mathrm{st}} + C_{\mathrm{er}}.$$

This completes the proof for $u \in C^\infty(\omega)$.

Now let $u \in H^m(\omega)$. Due to Theorem 5.12, we can find a sequence $(u_n)_{n=1}^\infty$ in $C^\infty(\omega)$ with

$$\lim_{n\to\infty} \|u - u_n\|_{H^m(\omega)} = 0.$$

This implies that $(u_n)_{n=1}^\infty$ is a Cauchy sequence with respect to the H^m-norm. Due to our previous result, it is also a Cauchy sequence with respect to the maximum norm. Since $C(\omega)$ is a complete space if equipped with the maximum norm, we find $\widetilde{u} \in C(\omega)$ such that

$$\lim_{n\to\infty} \|\widetilde{u} - u_n\|_{\infty,\omega} = 0.$$

The domain ω is bounded, so convergence in the maximum norm implies convergence in the L^2-norm, i.e.,

$$\lim_{n\to\infty} \|\widetilde{u} - u_n\|_{L^2(\omega)} = 0.$$

By our definition, the sequence $(u_n)_{n=1}^\infty$ also converges to u in the L^2-norm, so we have $\|u - \widetilde{u}\|_{L^2(\omega)} = 0$, i.e., u and $\widetilde{u}$ are identical up to a null set.

The norm estimate for u follows immediately from the estimates for $(u_n)_{n=1}^\infty$. ■

Remark 6.33 **($C(\omega)$ and $L^2(\omega)$)** *The statement of Theorem 6.32 is a little ambiguous, since $u \in H^m(\omega)$ is only defined up to null sets, but the maximum norm takes the pointwise maximum.*

The proof of the theorem suggests the correct interpretation: in the equivalence class of functions u that differ only on a null set, there is a continuous function, and this continuous function satisfies the estimate.

In fact, this continuous function is unique, since continuous functions that are identical up to null sets already have to be identical.

6.6. Approximation error estimates

In view of Theorem 6.32, we require $d \in \{1, 2, 3\}$. Let $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{T} \subseteq S_d^d$ be a triangulation of Ω.

Corollary 6.34 (Sobolev's lemma) *Assume that there is a family $(\Omega_i)_{i=1}^{\ell}$ of subdomains such that*

$$\Omega = \bigcup_{i=1}^{\ell} \Omega_i,$$

and that the subdomains Ω_i are star-shaped with respect to suitable balls.

Then there is a constant $C_{so,\Omega} \in \mathbb{R}_{>0}$ such that

$$\|u\|_{\infty,\Omega} \leq C_{so,\Omega} \|u\|_{H^2} \qquad \text{for all } u \in H^2(\Omega).$$

Proof. For each $i \in [1 : \ell]$, the domain Ω_i has finite chunkiness by definition. Since $2 > 3/2 \geq d/2$, we can apply Theorem 6.32 to find that $u|_{\Omega_i}$ is continuous and that its maximum norm can be bounded by a constant times $\|u\|_{H^2}$.

We can complete the proof by defining $C_{\mathrm{so},\Omega}$ as the maximum of the constants for the subdomains. ■

Under the assumptions of Corollary 6.34, we can define the *nodal interpolation operator*

$$\mathfrak{I}_T \colon H^2(\Omega) \to \mathcal{P}_{T,1}, \qquad u \mapsto \sum_{v \in \mathcal{N}_T} u(v)\varphi_v,$$

since $u \in H^2(\Omega)$ ensures that u is continuous, so the pointwise evaluation is well-defined.

Lemma 6.35 (Stability) *Under the assumptions of Corollary 6.34, we have*

$$\|\mathfrak{I}_T[u]\|_{\infty,\Omega} \leq C_{so,\Omega} \|u\|_{H^2} \qquad \text{for all } u \in H^2(\Omega).$$

Proof. Let $t \in T$. Due to Lemma 6.17, we have

$$\|\mathfrak{I}_T[u]\|_{\infty,\omega_t} = \|\mathfrak{I}_t[u]\|_{\infty,\omega_t} \leq \|u\|_{\infty,\omega_t} \leq C_{\mathrm{so},\Omega} \|u\|_{H^2} \qquad \text{for all } u \in H^2(\Omega).$$

Due to (6.4b), this implies our claim. ■

In order to obtain an estimate for the interpolation error, we can apply the Bramble-Hilbert lemma (cf. Theorem 6.27) to all simplices in the triangulation. We introduce the *maximal meshwidth* by

$$h_T := \max\{\mathrm{diam}(\omega_t) \ : \ t \in T\},$$

denote the chunkiness of ω_t by γ_t, and define the *maximal chunkiness* by

$$\gamma_T := \max\{\gamma_t \ : \ t \in T\}.$$

Lemma 6.36 (Inverse estimate) *Let $t \in S_d^d$, and let $\gamma > 0$ denote the chunkiness parameter of ω_t. We have*

$$\|\nabla p\|_2 \leq \gamma \, \mathrm{diam}(\omega_t)^{-1} \|p\|_{\infty,\omega_t} \qquad \text{for all } p \in \Pi_1^d.$$

Proof. Let $p \in \Pi_1^d$. If $\nabla p = 0$, the estimate is trivial.

Assume now $\nabla p \neq 0$. Let $\epsilon \in \mathbb{R}_{>0}$. Let $\mathcal{B} \subseteq \omega_t$ be a ball of radius $r > \operatorname{diam}(\omega_t)/(\gamma+\epsilon)$ centered at x_0. We let

$$\alpha := \frac{r}{\|\nabla p\|_2}, \qquad y := x_0 + \alpha \nabla p, \qquad z := x_0 - \alpha \nabla p,$$

and have

$$\|y - x_0\|_2 = \alpha \|\nabla p\|_2 = r, \qquad \|z - x_0\|_2 = \alpha \|\nabla p\|_2 = r,$$

and since $\mathcal{B} \subseteq \omega_t$ holds, we conclude $y, z \in \bar{\omega}_t$. This implies

$$\begin{aligned} 2\|p\|_{\infty,\omega_t} &\geq |p(y) - p(z)| = |\langle \nabla p, y - z\rangle_2| = |\langle \nabla p, 2\alpha \nabla p\rangle_2| \\ &= 2\alpha \|\nabla p\|_2^2 = \frac{2r}{\|\nabla p\|_2} \|\nabla p\|_2^2 = 2r \|\nabla p\|_2, \end{aligned}$$

and we find

$$\|\nabla p\|_2 \leq \frac{1}{r} \|p\|_{\infty,\omega_t} \leq \frac{\gamma + \epsilon}{\operatorname{diam}(\omega_t)} \|p\|_{\infty,\omega_t}.$$

Since this estimate holds for all $\epsilon > 0$, we get the final result. ■

Theorem 6.37 (Interpolation error) *Under the assumptions of Corollary 6.34, we can find $C_{in} \in \mathbb{R}_{>0}$ with*

$$\|u - \mathfrak{I}_T[u]\|_{L^2} \leq C_{in} h_T^2 |u|_{H^2}, \tag{6.23a}$$

$$\|u - \mathfrak{I}_T[u]\|_{H^1} \leq C_{in} h_T |u|_{H^2} \qquad \textit{for all } u \in H^2(\Omega). \tag{6.23b}$$

Proof. Let $u \in H^2(\Omega)$, and let $t \in T$. By definition, we can find a ball $\mathcal{B} \subseteq \omega_t$ of radius $r > \operatorname{diam}(\omega_t)/(2\gamma_T)$. Let $\mathfrak{T}_1$ denote the corresponding averaged Taylor expansion operator of degree $m = 1$ and let $p := \mathfrak{T}_1[u]$.

Due to (6.13b), we have

$$\begin{aligned} \|u - \mathfrak{I}_T[u]\|_{L^2(\omega_t)} &= \|u - \mathfrak{I}_t[u]\|_{L^2(\omega_t)t} = \|u - p + \mathfrak{I}_t[p] - \mathfrak{I}_t[u]\|_{L^2(\omega_t)} \\ &\leq \|u - p\|_{L^2(\omega_t)} + \|\mathfrak{I}_t[u - p]\|_{L^2(\omega_t)}. \end{aligned}$$

We can apply the Bramble-Hilbert lemma (cf. Theorem 6.27) to the first term to get

$$\|u - p\|_{L^2(\omega_t)} = \|u - \mathfrak{T}_1[u]\|_{L^2(\omega_t)} \leq C_1 \operatorname{diam}(\omega_t)^{m+1} |u|_{H^{m+1}(\omega_t)}$$

with a constant C_1 depending only on d, $\widehat{\varphi}$, γ_T, d, and m.

For the second term, we use

$$\|\mathfrak{I}_t[u - p]\|_{L^2(\omega_t)} = \left(\int_{\omega_t} \mathfrak{I}_t[u - p](x)^2 \, dx \right)^{1/2} \leq |\omega_t|^{1/2} \|\mathfrak{I}_t[u - p]\|_{\infty,\omega_t}.$$

We can find a constant C_2 depending only on d such that

$$|\omega_t| \leq C_2 \operatorname{diam}(\omega_t)^d,$$

and combining (6.13a) with Lemma 6.30 yields

$$\begin{aligned}\|\mathfrak{I}_t[u-p]\|_{L^2(\omega_t)} &\leq C_2^{1/2} \operatorname{diam}(\omega_t)^{d/2} \|\mathfrak{I}_t[u-p]\|_{\infty,\omega_t} \leq C_2^{1/2} \operatorname{diam}(\omega_t)^{d/2} \|u-p\|_{\infty,\omega_t} \\ &\leq C_2^{1/2} \operatorname{diam}(\omega_t)^{d/2} C_{\text{er}} \operatorname{diam}(\omega_t)^{m+1-d/2} |u|_{H^{m+1}(\omega_t)} \\ &= C_2^{1/2} C_{\text{er}} \operatorname{diam}(\omega_t)^{m+1} |u|_{H^{m+1}(\omega_t)}. \qquad (6.24)\end{aligned}$$

Combining the estimates for $u-p$ and $\mathfrak{I}_t[u-p]$ gives us

$$\|u - \mathfrak{I}_T[u]\|_{L^2(\omega_t)} \leq C_3 \operatorname{diam}(\omega_t)^{m+1} |u|_{H^2(\omega_t)} \qquad (6.25)$$

with $C_3 := C_1 + C_2^{1/2} C_{\text{er}}$. This gives us a local version of (6.23a).

In order to get a similar result for (6.23b), we fix $\nu \in \mathbb{N}_0^d$ with $|\nu| \leq m$ and use Exercise 6.28 to find

$$\begin{aligned}\|\partial_\nu(u - \mathfrak{I}_t[u])\|_{L^2(\omega_t)} &= \|\partial_\nu(u-p) + \partial_\nu \mathfrak{I}_t[u-p]\|_{L^2(\omega_t)} \\ &\leq \|\partial_\nu(u-p)\|_{L^2(\omega_t)} + \|\partial_\nu \mathfrak{I}_t[u-p]\|_{L^2(\omega_t)} \\ &= \|\partial_\nu u - \mathfrak{T}_0[\partial_\nu u]\|_{L^2(\omega_t)} + \|\partial_\nu \mathfrak{I}_t[u-p]\|_{L^2(\omega_t)}.\end{aligned}$$

We can again apply the Bramble-Hilbert lemma to the first term to get

$$\|\partial_\nu(u-p)\|_{L^2(\omega_t)} \leq C_4 \operatorname{diam}(\omega_t)^m |\partial_\nu u|_{H^m(\omega_t)} \leq C_4 \operatorname{diam}(\omega_t)^m |u|_{H^{m+1}(\omega_t)}.$$

For the second term, we can use Lemma 6.36 and (6.24) to find

$$\begin{aligned}\|\partial_\nu \mathfrak{I}_t[u-p]\|_{L^2(\omega_t)} &\leq C_2^{1/2} \operatorname{diam}(\omega_t)^{d/2} \|\partial_\nu \mathfrak{I}_t[u-p]\|_{\infty,\omega_t} \\ &\leq C_2^{1/2} \gamma \operatorname{diam}(\omega_t)^{d/2-1} \|\mathfrak{I}_t[u-p]\|_{\infty,\omega_t} \\ &\leq C_2^{1/2} \gamma C_{\text{er}} \operatorname{diam}(\omega_t)^m |u|_{H^{m+1}(\omega_t)},\end{aligned}$$

and combining the two estimates yields

$$\|\partial_\nu(u - \mathfrak{I}_T[u])\|_{L^2(\omega_t)} \leq C_5 \operatorname{diam}(\omega_t)^m |u|_{H^{m+1}(\omega_t)} \qquad (6.26)$$

with $C_5 := C_4 + \gamma C_2^{1/2} C_{\text{er}}$.

Now we only have to combine the local estimates to obtain the global results. Due to Lemma 6.6, the local estimate (6.25) gives rise to

$$\begin{aligned}\|u - \mathfrak{I}_T[u]\|_{L^2(\Omega)}^2 &= \sum_{t \in T} \|u - \mathfrak{I}_T[u]\|_{L^2(\omega_t)}^2 \leq \sum_{t \in T} C_3^2 \operatorname{diam}(\omega_t)^{2(m+1)} |u|_{H^2(\omega_t)}^2 \\ &\leq C_3^2 \sum_{t \in T} h_T^{2(m+1)} |u|_{H^2(\omega_t)}^2 = C_3^2 h_T^{2(m+1)} |u|_{H^2(\Omega)}^2.\end{aligned}$$

We take the square root to get the first estimate (6.23a).

For the second estimate (6.23b), we can apply the same reasoning to (6.26). ∎

Corollary 6.38 (Discretization error) *Let $\mathcal{V} = H_0^1(\Omega)$ and $\mathcal{V}_n = \mathcal{P}_{\mathcal{T},1} \cap H_0^1(\Omega)$. Let $u \in \mathcal{V}$ be the solution of the variational problem (5.8), and let $u_n \in \mathcal{V}_n$ be the solution of the discretized variational problem (5.14).*

If $u \in H^2(\Omega)$ holds, we have

$$\|u - u_n\|_{H^1} \leq C_{in} \left(\frac{C_B}{C_K} \right)^{1/2} h_{\mathcal{T}}\, |u|_{H^2}$$

where C_{in} is the constant from Theorem 6.37, C_B is the continuity constant of the bilinear form a (in this case $C_B = 1$), and C_K is the coercivity constant (in this case $C_K \geq 1/(1 + \operatorname{diam}(\Omega)^2)$ due to Corollary 5.26).

Proof. Let $\widetilde{u}_n := \mathfrak{I}_{\mathcal{T}}[u]$. Due to $u \in H^2(\Omega)$, we can apply Theorem 6.32 to find $u \in C(\Omega)$, and $u \in H_0^1(\Omega)$ yields $u|_{\partial\Omega} = 0$. Since we use a *nodal* interpolation operator, this implies $\widetilde{u}_n \in \mathcal{V}_n$.

Now we can apply Céa's Lemma 5.40 and Theorem 6.37 to get

$$\|u - u_n\|_{H^1} \leq \sqrt{\frac{C_B}{C_K}} \|u - \widetilde{u}_n\|_{H^1} \leq \sqrt{\frac{C_B}{C_K}} C_{\mathrm{in}} h_{\mathcal{T}} |u|_{H^2}.$$

■

Definition 6.39 (H^2-regularity) *A variational problem (5.9) with $\mathcal{V} \subseteq H^1(\Omega)$ is called* H^2-regular *if for $f \in L^2(\Omega)$ and*

$$\beta(v) = \langle v, f \rangle_{L^2} \qquad \text{for all } v \in \mathcal{V}$$

the solution $u \in \mathcal{V}$ satisfies $u \in H^2(\Omega)$ and

$$\|u\|_{H^2} \leq C_R \|f\|_{L^2} \tag{6.27}$$

with a constant C_R depending only on the bilinear form a and the domain Ω.

If the variational problem (5.9) is H^2-regular, Corollary 6.38 takes the form

$$\|u - u_n\|_{H^1} \leq C_R C_{\mathrm{in}} \sqrt{\frac{C_B}{C_K}}\, h_{\mathcal{T}}\, \|f\|_{L^2},$$

i.e., we can bound the discretization error in terms of the L^2-norm of the right-hand side.

Compared to error estimates like the one provided by Theorem 2.10 that guarantee that the error falls like h^2 for a finite difference discretization, it is a little disappointing to get only $h_{\mathcal{T}}$ in the case of the finite element method.

A closer look reveals that the two estimates are not really comparable: in the finite difference case, we only get a bound for the maximum of the error, while in the finite difference case the derivatives of the error are also controlled.

Our goal is now to prove that the L^2-norm of the error does indeed converge like $h_{\mathcal{T}}^2$.

Lemma 6.40 (Aubin-Nitsche) *Let a be symmetric, bounded, and coercive with respect to the H^1-norm. Let the variational problem (5.9) be H^2-regular, and let $f \in L^2(\Omega)$.*

There is a constant C_A such that the solutions $u \in \mathcal{V}$ and $u_n \in \mathcal{V}_n$ of (5.9) and (5.14) satisfy

$$\|u - u_n\|_{L^2} \leq C_A h_T \|u - u_n\|_{H^1},$$
$$\|u - u_n\|_{L^2} \leq C_A h_T^2 \|f\|_{L^2}.$$

Proof. The proof relies on *Nitsche's trick*: we consider the functional

$$\lambda \colon \mathcal{V} \to \mathbb{R}, \qquad v \mapsto \langle v, u - u_n \rangle_{L^2}.$$

Due to the Cauchy-Schwarz inequality (5.4), it is an element of $\mathcal{V}'$ with

$$\|\lambda\|_{\mathcal{V}'} \leq \|u - u_n\|_{L^2}.$$

The Riesz theorem 5.22 yields that we can find a solution $e \in \mathcal{V}$ of the variational problem

$$a(e, v) = \lambda(v) \qquad \text{for all } v \in \mathcal{V},$$

and that this solution satisfies $\|e\|_{\mathcal{V}} \leq \|\lambda\|_{\mathcal{V}'}/C_K$. We have

$$\|u - u_n\|_{L^2}^2 = \lambda(u - u_n) = a(e, u - u_n).$$

We let $e_n := \mathfrak{I}_T[e]$ and take advantage of Galerkin orthogonality (cf. Lemma 5.36) to get

$$\|u - u_n\|_{L^2}^2 = a(e, u - u_n) = a(e - e_n, u - u_n) \leq C_B \|e - e_n\|_{\mathcal{V}} \|u - u_n\|_{\mathcal{V}}.$$

Now we can apply Theorem 6.37 and (6.27) to obtain

$$\|e - e_n\|_{\mathcal{V}} \leq C_{\text{in}}\, h_T\, \|e\|_{H^2} \leq C_{\text{in}} C_R\, h_T\, \|u - u_n\|_{L^2}$$

and thus

$$\|u - u_n\|_{L^2}^2 \leq C_B C_{\text{in}} C_R\, h_T\, \|u - u_n\|_{L^2} \|u - u_n\|_{\mathcal{V}}.$$

Dividing both sides by $\|u - u_n\|_{L^2}$ yields the first estimate.

Applying Corollary 6.38 to $\|u - u_n\|_{\mathcal{V}}$ gives us

$$\|u - u_n\|_{\mathcal{V}} \leq C_{\text{in}} C_R \frac{\sqrt{C_B}}{\sqrt{C_K}}\, h_T\, \|f\|_{L^2},$$

and combining this with the first estimate yields

$$\|u - u_n\|_{L^2} \leq C_B C_{\text{in}} C_R h_T \|u - u_n\|_{\mathcal{V}} \leq C_B C_{\text{in}}^2 C_R^2 \frac{\sqrt{C_B}}{\sqrt{C_K}}\, h_T^2\, \|f\|_{L^2}.$$

Now we choose

$$C_A := \max\left\{ C_B C_{\text{in}} C_R, C_{\text{in}}^2 C_R^2 \frac{C_B^{3/2}}{C_K^{1/2}} \right\}$$

to complete the proof. ■

Remark 6.41 (Unsymmetric case) *If the bilinear form a is not symmetric, we can still derive a version of the Aubin-Nitsche lemma if the* adjoint problem

Find $e \in \mathcal{V}$ such that

$$a(e, v) = \lambda(v) \qquad\qquad \text{for all } v \in \mathcal{V}$$

is H^2-regular.

Remark 6.42 (Residual error estimator) *If the bilinear form a is symmetric, continuous, and coercive, the energy norm satisfies*

$$\|u - u_h\|_A = a\left(\tfrac{u-u_h}{\|u-u_h\|_A}, u - u_h\right) = \sup\left\{\frac{a(v, u - u_h)}{\|v\|_A} \ : \ v \in \mathcal{V} \setminus \{0\}\right\}$$

due to the Cauchy-Schwarz inequality. The functional $a(\cdot, u - u_h)$ is called the residual *for the solution u and its Galerkin approximation u_n. If we can find a bound for the dual norm of the residual, we have a bound for the error in the energy norm.*

Let $v \in \mathcal{V}$, and let $v_h \in \mathcal{V}_n$ be a suitable approximation. We denote by $u_t = u_n|_{\omega_t}$ and $v_t = v_n|_{\omega_t}$ the element-wise polynomials. Using the Galerkin orthogonality and partial integration, we obtain

$$\begin{aligned} a(v, u - u_h) &= a(v - v_h, u - u_h) = a(v - v_h, u) - a(v - v_h, u_h) \\ &= \int_\Omega (v - v_h)(x) f(x)\, dx - a(v - v_h, u_h) \\ &= \sum_{t \in T} \int_{\omega_t} (v - v_h)(x) f(x)\, dx - \int_{\omega_t} \langle \nabla(v - v_h)(x), \nabla u_h(x)\rangle_2\, dx \\ &= \sum_{t \in T} \int_{\omega_t} (v - v_h)(x) f(x)\, dx - \int_{\partial\omega_t} (v - v_t)(x) \,\langle n(x), \nabla u_t(x)\rangle_2 \\ &\quad + \sum_{t \in T} \int_{\omega_t} (v - v_t)(x) \Delta u_t(x)\, dx. \end{aligned}$$

Since u_t is a linear polynomial, we have $\Delta u_t = 0$ for all $t \in T$, and the last term vanishes.

We can write $\partial\omega_t$ again in terms of the edges of T and find

$$\begin{aligned} a(v, u - u_h) = &- \sum_{e \in \mathcal{E}_T} \int_{\omega_e} (v - v_t)(x) \,\langle n_e(x), \nabla(u_{t_{e,+}} - u_{t_{e,-}})\rangle_2\, dx \\ &+ \sum_{t \in T} \int_{\omega_t} (v - v_t)(x) f(x)\, dx. \end{aligned}$$

With suitable local mesh width parameters h_e and h_t for all edges $e \in \mathcal{E}_T$ and all simplices $t \in T$, we can use the Cauchy-Schwarz inequality to get

$$a(v, u - u_h) \leq \sum_{e \in \mathcal{E}_T} \|v - v_t\|_{L^2(e)} \|\langle n_e, \nabla(u_{t_{e,+}} - u_{t_{e,-}})\rangle_2\|_{L^2(e)}$$

$$
\begin{aligned}
&+ \sum_{t\in T} \|v - v_t\|_{L^2(\omega_t)} \|f\|_{L^2(\omega_t)} \\
&= \sum_{e\in\mathcal{E}_T} \frac{1}{\sqrt{h_e}} \|v - v_t\|_{L^2(e)} \sqrt{h_e} \|\langle n_e, \nabla(u_{t_{e,+}} - u_{t_{e,-}})\rangle_2\|_{L^2(e)} \\
&\quad + \sum_{t\in T} \frac{1}{h_t} \|v - v_t\|_{L^2(\omega_t)} h_t \|f\|_{L^2(\omega_t)}.
\end{aligned}
$$

Using the Clément interpolation operator*, we can find* v_t *such that* $\|v - v_t\|_{L^2(e)}/\sqrt{h_e}$ *and* $\|v - v_t\|_{L^2(t)}/h_t$ *can be bounded in terms of* $\|v\|_{H^1}$*, and we can use the Cauchy-Schwarz inequality again to conclude*

$$
a(v, u - u_h) \leq \|v\|_{H^1} \Big(\sum_{e\in\mathcal{E}_T} h_e \|\langle n_e, \nabla(u_{t_{e,+}} - u_{t_{e,-}})\rangle_2\|^2_{L^2(\omega_e)} + \sum_{t\in T} h_t^2 \|f\|^2_{L^2(\omega_t)} \Big)^{1/2}.
$$

Dividing by $\|v\|_{H^1}$ *yields an upper bound for the residual norm.*

6.7. Time-dependent problems

We consider the heat equation

$$
\frac{\partial u}{\partial t}(t, x) = g(t, x) + \Delta_x u(t, x) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ x \in \Omega.
$$

For the finite element discretization, we multiply by a test function $v \in H^1_0(\Omega)$, integrate, and apply partial integration to get

$$
\begin{aligned}
\int_\Omega v(x) \frac{\partial u}{\partial t}(t, x)\, dx &= \int_\Omega v(x) g(t, x)\, dx + \int_\Omega v(x) \Delta_x u(t, x)\, dx, \\
\frac{\partial}{\partial t} \int_\Omega v(x) u(t, x)\, dx &= \int_\Omega v(x) g(t, x)\, dx - \int_\Omega \langle \nabla v(x), \nabla u(t, x)\rangle_2\, dx, \\
\frac{\partial}{\partial t} \langle v, u(t, \cdot)\rangle_{L^2} &= \langle v, g(t, \cdot)\rangle_{L^2} - a(v, u(t, \cdot)) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ v \in H^1_0(\Omega),
\end{aligned}
$$

where $a(\cdot, \cdot)$ denotes the familiar bilinear form

$$
a\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}, \qquad (v, u) \mapsto \langle \nabla v, \nabla u\rangle_{L^2},
$$

used already for Poisson's equation.

Replacing $\mathcal{V} := H^1_0(\Omega)$ by a finite element subspace $\mathcal{V}_n \subseteq \mathcal{V}$ yields

$$
\frac{\partial}{\partial t} \langle v_n, u_n(t, \cdot)\rangle_{L^2} = \langle v_n, g(t, \cdot)\rangle_{L^2} - a(v_n, u_n(t, \cdot)) \qquad \text{for all } t \in \mathbb{R}_{\geq 0},\ v_n \in \mathcal{V}_n.
$$

As before, we choose a suitable basis $(\varphi_i)_{i=1}^n$ of $\mathcal{V}_n$ and represent the solution u_n by its coefficients:

$$
u_n(t, \cdot) = \sum_{i=1}^n y_i(t) \varphi_i \qquad \text{for all } t \in \mathbb{R}_{\geq 0},
$$

where

$$y\colon \mathbb{R}_{\geq 0} \to \mathbb{R}^n, \qquad\qquad t \mapsto y(t),$$

is now a function of time.

Introducing the *stiffness matrix* $A \in \mathbb{R}^{n\times n}$, the *mass matrix* $M \in \mathbb{R}^{n\times n}$, and the *forcing vector* $b\colon \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ by

$$\begin{aligned} a_{ij} &:= a(\varphi_i, \varphi_j), & m_{ij} &:= \langle \varphi_i, \varphi_j \rangle_{L^2}, \\ b_i(t) &:= \langle \varphi_i, g(t, \cdot) \rangle_{L^2} & &\text{for all } t \in \mathbb{R}_{\geq 0},\ i, j \in [1:n], \end{aligned}$$

the Galerkin formulation is equivalent with

$$\frac{\partial}{\partial t} My(t) = b(t) - Ay(t) \qquad\qquad \text{for all } t \in \mathbb{R}_{\geq 0}.$$

Since the mass matrix is always positiv definite and symmetric, we can multiply both sides of the equation by M^{-1} and obtain an ordinary differential equation that can be treated by time-stepping schemes.

We consider the Crank-Nicolson method as an example. The starting point is the trapezoidal rule

$$\begin{aligned} My(t_{i+1}) &= My(t_i) + \int_{t_i}^{t_{i+1}} My'(s)\, ds \\ &\approx My(t_i) + \frac{\delta}{2}(b(t_i) + b(t_{i+1}) - Ay(t_i) - Ay(t_{i+1})), \\ \left(M + \frac{\delta}{2}A\right) y(t_{i+1}) &\approx My(t_i) + \frac{\delta}{2}(b(t_i) + b(t_{i+1}) - Ay(t_i)), \end{aligned}$$

and we can obtain an approximation for $y(t_{i+1})$ by solving this linear system. Since both M and A are positive definite, symmetric, and sparse, essentially the same solvers as in the case of Poisson's equation can be applied.

A. Appendix

A.1. Perron-Frobenius theory

Now we can return our attention to the invertibility of a weakly diagonally dominant matrix $\mathbf{A}$. Let us take a closer look at the matrix

$$\mathbf{M} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$$

introduced in (2.10) during the proof of invertibility for *strictly* diagonally dominant matrices given in Lemma 2.17. In our model problem, the off-diagonal elements of $\mathbf{A}$ are non-positive, therefore the off-diagonal elements of $\mathbf{M}$ have to be non-negative.

In order to establish convergence of the Neumann series for $\mathbf{M}$, and therefore the invertibility of $\mathbf{A}$, it is a good idea to investigate the eigenvalues and eigenvectors of non-negative matrices. The corresponding results go back to Oskar Perron [9] and Georg Frobenius [6].

Definition A.1 (Positive vectors and matrices) *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ *and* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$. *We write*

$$\begin{aligned}
\mathbf{x} \leq \mathbf{y} &\iff \forall i \in \mathcal{I} \ : \ x_i \leq y_i,\\
\mathbf{x} < \mathbf{y} &\iff \forall i \in \mathcal{I} \ : \ x_i < y_i,\\
\mathbf{A} \leq \mathbf{B} &\iff \forall i, j \in \mathcal{I} \ : \ a_{ij} \leq b_{ij},\\
\mathbf{A} < \mathbf{B} &\iff \forall i, j \in \mathcal{I} \ : \ a_{ij} < b_{ij},
\end{aligned}$$

and denote the cone of non-negative vectors by

$$\mathbb{R}^{\mathcal{I}}_{\geq 0} := \{\mathbf{x} \in \mathbb{R}^{\mathcal{I}} \ : \ \mathbf{x} \geq \mathbf{0}\}.$$

Our goal is to prove that — under certain conditions — a positive matrix $\mathbf{A}$ has a positive maximal eigenvalue corresponding to a positive eigenvector. Since the matrices $\mathbf{M}$ appearing in our convergence analysis are not strictly positive, but only non-negative, we will also prove a generalization of this result.

Let $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ be a matrix with $\mathbf{A} \geq \mathbf{0}$. For our proof, we follow a rather elegant approach introduced by Helmut Wielandt [12]. We consider the function

$$r \colon \mathbb{R}^{\mathcal{I}} \setminus \{\mathbf{0}\} \to \mathbb{R}, \qquad \mathbf{x} \mapsto \min\left\{ \frac{(\mathbf{A}\mathbf{x})_i}{x_i} \ : \ i \in \mathcal{I},\ x_i \neq 0 \right\}. \tag{A.1}$$

Lemma A.2 (Quotient function) *If $\mathbf{e} \in \mathbb{R}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ is an eigenvector of $\mathbf{A}$ for an eigenvalue $\lambda \in \mathbb{R}$, we have*

$$r(\mathbf{e}) = \lambda.$$

We also have

$$r(\alpha\mathbf{x}) = r(\mathbf{x}) \qquad \textit{for all } \alpha \in \mathbb{R}_{>0},\ \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}, \tag{A.2a}$$

$$r(\mathbf{x})\mathbf{x} \leq \mathbf{A}\mathbf{x} \qquad \textit{for all } \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}, \tag{A.2b}$$

$$\varrho(\mathbf{A}) \leq \sup\{r(\mathbf{x}) \ :\ \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}\}. \tag{A.2c}$$

Proof. Let $\mathbf{e} \in \mathbb{R}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ be an eigenvector of $\mathbf{A}$ for an eigenvalue $\lambda \in \mathbb{R}$. Due to $\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$, we have

$$r(\mathbf{e}) = \min\left\{\frac{(\mathbf{A}\mathbf{e})_i}{e_i} \ :\ i \in \mathcal{I},\ e_i \neq 0\right\} = \min\left\{\frac{\lambda e_i}{e_i} \ :\ i \in \mathcal{I},\ e_i \neq 0\right\} = \lambda.$$

Let now $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ and $\alpha \in \mathbb{R}_{>0}$. We have

$$r(\alpha\mathbf{x}) = \min\left\{\frac{\alpha(\mathbf{A}\mathbf{x})_i}{\alpha x_i} \ :\ i \in \mathcal{I},\ x_i \neq 0\right\} = \min\left\{\frac{(\mathbf{A}\mathbf{x})_i}{x_i} \ :\ i \in \mathcal{I},\ x_i \neq 0\right\} = r(\mathbf{x}),$$

and the definition implies

$$r(\mathbf{x}) \leq \frac{(\mathbf{A}\mathbf{x})_i}{x_i}, \qquad r(\mathbf{x})x_i \leq (\mathbf{A}\mathbf{x})_i \qquad \text{for all } i \in \mathcal{I},\ x_i \neq 0.$$

Due to $\mathbf{A} \geq \mathbf{0}$ and $\mathbf{x} \geq \mathbf{0}$, the right-hand side cannot be negative, and we find

$$r(\mathbf{x})x_i \leq (\mathbf{A}\mathbf{x})_i \qquad \text{for all } i \in \mathcal{I}.$$

This is equivalent to $r(\mathbf{x})\mathbf{x} \leq \mathbf{A}\mathbf{x}$.

Let $\lambda \in \mathbb{C}$ be an eigenvalue of $\mathbf{A}$, and let $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ be a corresponding eigenvector. We define $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ by

$$x_i := |e_i| \qquad \text{for all } i \in \mathcal{I}.$$

The triangle inequality yields

$$|\lambda|x_i = |\lambda e_i| = |(\mathbf{A}\mathbf{e})_i| = \left|\sum_{j\in\mathcal{I}} a_{ij}e_j\right| \leq \sum_{j\in\mathcal{I}} a_{ij}|e_j| = (\mathbf{A}\mathbf{x})_i,$$

so we have

$$|\lambda| \leq \frac{(\mathbf{A}\mathbf{x})_i}{x_i} \qquad \text{for all } i \in \mathcal{I},\ x_i \neq 0,$$

and conclude $|\lambda| \leq r(\mathbf{x})$. ■

We are looking for an eigenvector of $\mathbf{A}$, i.e., a vector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ satisfying the equation $r(\mathbf{x})\mathbf{x} = \mathbf{A}\mathbf{x}$. Due to (A.2b), we have $r(\mathbf{x})\mathbf{x} \leq \mathbf{A}\mathbf{x}$ for all vectors $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$. In order to reach equality, we should therefore try to find a maximum of the function r.

Lemma A.3 (Eigenvector) *Let* $\mathbf{A} > \mathbf{0}$*, and let* $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ *be such that*

$$r(\mathbf{z}) \leq r(\mathbf{x}) \qquad \text{for all } \mathbf{z} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}.$$

Then $\mathbf{x}$ *is an eigenvector of* $\mathbf{A}$ *for the eigenvalue* $r(\mathbf{x})$ *and satisfies* $\mathbf{x} > \mathbf{0}$.

Proof. Let $\lambda := r(\mathbf{x})$, and let

$$\mathbf{y} := \mathbf{A}\mathbf{x} - \lambda\mathbf{x}.$$

Due to (A.2b), we have $\mathbf{y} \geq \mathbf{0}$.

Let $\mathbf{z} := \mathbf{A}\mathbf{x}$. Due to $\mathbf{A} > \mathbf{0}$ and $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$, we have $\mathbf{z} > \mathbf{0}$.

Since λ is maximal, we obtain

$$\mathbf{A}\mathbf{z} - r(\mathbf{z})\mathbf{z} \geq \mathbf{A}\mathbf{z} - \lambda\mathbf{z} = \mathbf{A}^2\mathbf{x} - \lambda\mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{A}\mathbf{x} - \lambda\mathbf{x}) = \mathbf{A}\mathbf{y}.$$

By definition, we can find $i \in \mathcal{I}$ such that

$$r(\mathbf{z}) = \frac{(\mathbf{A}\mathbf{z})_i}{z_i}, \qquad (\mathbf{A}\mathbf{y})_i = (\mathbf{A}\mathbf{z} - r(\mathbf{z})\mathbf{z})_i = 0.$$

Due to $\mathbf{A} > \mathbf{0}$, this gives us $\mathbf{y} = \mathbf{0}$, so $\mathbf{x}$ is an eigenvector for the eigenvalue λ.

The definition (A.1) implies $\lambda = r(\mathbf{x}) > 0$, and $\lambda\mathbf{x} = \mathbf{A}\mathbf{x} = \mathbf{z} > \mathbf{0}$ yields $\mathbf{x} > \mathbf{0}$. ■

All we have to do is to prove that r has a maximum in $\mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$. Due to (A.2a), we can restrict our attention to a compact subset

$$C := \left\{\mathbf{x} \in \mathbb{R}^{\mathcal{I}} \ : \ \mathbf{x} \geq \mathbf{0},\ \sum_{i \in \mathcal{I}} x_i = 1\right\},$$

since we can scale any vector in $\mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ by $1/\sum_{i \in \mathcal{I}} x_i$ to put it into this set without changing the value of r.

Lemma A.4 (Upper semi-continuity) *The function* r *is upper semi-continuous, i.e., for each* $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ *and* $\epsilon \in \mathbb{R}_{>0}$*, we can find* $\delta \in \mathbb{R}_{>0}$ *such that*

$$r(\mathbf{y}) \leq r(\mathbf{x}) + \epsilon \qquad \text{for all } \mathbf{y} \in \mathbb{R}^{\mathcal{I}},\ \|\mathbf{x} - \mathbf{y}\|_\infty < \delta.$$

Proof. Let $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ and $\epsilon \in \mathbb{R}_{>0}$.

We define the set

$$N := \{i \in \mathcal{I} \ : \ x_i \neq 0\}$$

of indices corresponding to non-zero entries in $\mathbf{x}$. Due to $\mathbf{x} \neq \mathbf{0}$, it cannot be empty.

We let

$$\hat{\delta} := \min\{x_i/2 \ : \ i \in N\}$$

and observe that for each $\mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ with $\|\mathbf{x} - \mathbf{y}\|_\infty < \hat{\delta}$ we have

$$y_i = x_i - x_i + y_i \geq x_i - |x_i - y_i| \geq x_i - \hat{\delta} \geq x_i/2 > 0 \qquad \text{for all } i \in N. \tag{A.3}$$

This means that for each $i \in N$, the mapping

$$\{\mathbf{y} \in \mathbb{R}^{\mathcal{I}} \ : \ \|\mathbf{x} - \mathbf{y}\|_\infty < \hat{\delta}\} \to \mathbb{R}, \qquad\qquad \mathbf{y} \mapsto \frac{(\mathbf{A}\mathbf{y})_i}{y_i}$$

is continuous, so we can find $\delta \in (0, \hat{\delta})$ such that we have

$$\frac{(\mathbf{A}\mathbf{y})_i}{y_i} < \frac{(\mathbf{A}\mathbf{x})_i}{x_i} + \epsilon \qquad\qquad \text{for all } i \in N,\ \mathbf{y} \in \mathbb{R}^{\mathcal{I}} \text{ with } \|\mathbf{x} - \mathbf{y}\|_\infty < \delta.$$

Let now $\mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ with $\|\mathbf{x} - \mathbf{y}\|_\infty < \delta$. We have

$$\begin{aligned} r(\mathbf{y}) &= \min\left\{ \frac{(\mathbf{A}\mathbf{y})_i}{y_i} \ : \ i \in \mathcal{I},\ y_i \neq 0 \right\} \overset{(A.3)}{\leq} \min\left\{ \frac{(\mathbf{A}\mathbf{y})_i}{y_i} \ : \ i \in N \right\} \\ &< \min\left\{ \frac{(\mathbf{A}\mathbf{x})_i}{x_i} + \epsilon \ : \ i \in N \right\} = r(\mathbf{x}) + \epsilon. \end{aligned}$$

■

Lemma A.5 (Maximum) *There exists* $\mathbf{x} \in C$ *such that*

$$r(\mathbf{z}) \leq r(\mathbf{x}) \qquad\qquad \textit{for all } \mathbf{z} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}.$$

Proof. Let $\mathbf{e} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ denote the vector with $e_i = 1$ for all $i \in \mathcal{I}$. Due to (A.2b), we have

$$\langle \mathbf{e}, \mathbf{A}\mathbf{x} \rangle_2 - r(\mathbf{x}) \langle \mathbf{e}, \mathbf{x} \rangle_2 = \sum_{i \in \mathcal{I}} (\mathbf{A}\mathbf{x})_i - r(\mathbf{x}) x_i \geq 0 \qquad\qquad \text{for all } \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0},$$

and we obtain

$$r(\mathbf{x}) \leq \frac{\langle \mathbf{e}, \mathbf{A}\mathbf{x} \rangle_2}{\langle \mathbf{e}, \mathbf{x} \rangle_2} = \frac{\langle \mathbf{A}^* \mathbf{e}, \mathbf{x} \rangle_2}{\langle \mathbf{e}, \mathbf{x} \rangle_2} \qquad\qquad \text{for all } \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}.$$

Let $\mu \in \mathbb{R}_{>0}$ denote the maximal entry of the vector $\mathbf{A}^* \mathbf{e}$. We have

$$\langle \mathbf{A}^* \mathbf{e}, \mathbf{x} \rangle_2 = \sum_{i \in \mathcal{I}} (\mathbf{A}^* \mathbf{e})_i x_i \leq \alpha \sum_{i \in \mathcal{I}} x_i = \alpha \langle \mathbf{e}, \mathbf{x} \rangle_2 \qquad\qquad \text{for all } \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0},$$

so we can conclude

$$r(\mathbf{x}) \leq \frac{\langle \mathbf{A}^* \mathbf{e}, \mathbf{x} \rangle_2}{\langle \mathbf{e}, \mathbf{x} \rangle_2} \leq \frac{\alpha \langle \mathbf{e}, \mathbf{x} \rangle_2}{\langle \mathbf{e}, \mathbf{x} \rangle_2} = \alpha \qquad\qquad \text{for all } \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}.$$

This means that the supremum

$$\lambda := \sup\{ r(\mathbf{x}) \ : \ \mathbf{x} \in C \}$$

is bounded by α and therefore a real number.

By definition of the supremum, we can find a sequence $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ in C satisfying

$$r(\mathbf{x}^{(m)}) > \lambda - 1/m \qquad\qquad \text{for all } m \in \mathbb{N},$$

and since C is compact, this sequence has a limit point $\mathbf{x} \in C$. We only have to prove $r(\mathbf{x}) = \lambda$.

Let $\epsilon \in \mathbb{R}_{>0}$. Due to Lemma A.4, we can find $\delta \in \mathbb{R}_{>0}$ such that

$$r(\mathbf{y}) < r(\mathbf{x}) + \epsilon/2 \qquad\qquad \text{for all } \mathbf{y} \in \mathbb{R}^{\mathcal{I}} \text{ with } \|\mathbf{y} - \mathbf{x}\|_\infty < \delta.$$

Since $\mathbf{x}$ is a limit point, we can also find $m \in \mathbb{N}$ such that

$$\|\mathbf{x}^{(m)} - \mathbf{x}\| < \delta, \qquad\qquad 1/m < \epsilon/2.$$

Combining both estimates yields

$$\lambda - \epsilon/2 < \lambda - 1/m < r(\mathbf{x}^{(m)}) < r(\mathbf{x}) + \epsilon/2, \qquad\qquad \lambda - \epsilon < r(\mathbf{x}).$$

Since ϵ can be chosen arbitrarily, this implies $\lambda \leq r(\mathbf{x})$. Since λ is the supremum of r in C, we also have $r(\mathbf{x}) \leq \lambda$ and therefore $\lambda = r(\mathbf{x})$. ∎

Corollary A.6 (Spectral radius) *Let* $\mathbf{A} > \mathbf{0}$. *The spectral radius* $\varrho(\mathbf{A})$ *of* $\mathbf{A}$ *is an eigenvalue with an eigenvector* $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ *satisfying* $\mathbf{x} > \mathbf{0}$.

Proof. Lemma A.5 gives us a vector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ with

$$r(\mathbf{z}) \leq r(\mathbf{x}) \qquad\qquad \text{for all } \mathbf{z} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}.$$

Due to Lemma A.3, this vector $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ for the eigenvalue $\lambda = r(\mathbf{x})$ with $\mathbf{x} > \mathbf{0}$.

Due to (A.2c), we have $\varrho(\mathbf{A}) \leq \lambda$. Since λ is an eigenvalue, we also have $\lambda = |\lambda| \leq \varrho(\mathbf{A})$, and can conclude $\lambda = \varrho(\mathbf{A})$. ∎

Unfortunately, matrices resulting from finite difference discretization schemes tend to be *sparse*, i.e., each row and column contains only a non-zero few coefficients. This means that $\mathbf{A} > \mathbf{0}$ is not a useful requirement for our investigation of finite difference methods. We need another condition that can take its place.

Definition A.7 (Connections) *Let* $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$, *and let* $i, j \in \mathcal{I}$.

We call a tuple $(i_\ell)_{\ell=0}^{m}$ *of indices in* $\mathcal{I}$ *a* connection *from* j *to* i *if*

$$j = i_0, \qquad i = i_m, \qquad a_{i_\ell, i_{\ell-1}} \neq 0 \qquad \text{for all } \ell \in [1:m].$$

The number $m \in \mathbb{N}_0$ *is called the* length *of the connection.*

Definition A.8 (Irreducible matrices) *A matrix* $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ *is called* irreducible *if for each pair* $i, j \in \mathcal{I}$ *there is a connection of* i *and* j. *If the matrix is not irreducible, it is called* reducible.

This definition has a geometric interpretation: we can define the *graph* of a matrix $\mathbf{A}$ by using $\mathcal{V} := \mathcal{I}$ as its vertices and

$$\mathcal{E} := \{(i,j) \ : \ a_{ij} \neq 0, \ i,j \in \mathcal{I}\}$$

as its edges. The matrix is irreducible if the graph is strongly connected.

In the case of a finite difference discretization, $a_{ij} \neq 0$ holds if and only if either $i = j$ or j is a neighbour of i. A matrix resulting from a finite difference scheme therefore is irreducible if we can reach any grid point from any other grid point by moving from one neighbour to the next. This is typically ensured if the underlying domain Ω is connected.

Exercise A.9 (Reducible matrix) *Prove that a matrix $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ is reducible if and only if there are index sets $\mathcal{I}_1, \mathcal{I}_2 \subseteq \mathcal{I}$ such that*

$$\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2, \qquad \emptyset = \mathcal{I}_1 \cap \mathcal{I}_2, \qquad \mathbf{A}|_{\mathcal{I}_2\times\mathcal{I}_1} = \mathbf{0}.$$

These conditions mean that $\mathbf{A}$ can be written in the form

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

of a block upper triangular matrix with non-trivial blocks.

For irreducible matrices, we can replace Lemma A.3 by the following more general result.

Lemma A.10 (Eigenvector) *Let $\mathbf{A} \geq \mathbf{0}$ be an irreducible matrix, let $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ be such that*

$$r(\mathbf{z}) \leq r(\mathbf{x}) \qquad \textit{for all } \mathbf{z} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}.$$

Then $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ for the eigenvalue $r(\mathbf{x})$ and satisfies $\mathbf{x} > \mathbf{0}$.

Proof. Let $\lambda := r(\mathbf{x})$, and let

$$\mathbf{y} := \mathbf{A}\mathbf{x} - \lambda\mathbf{x}.$$

Due to (A.2b), we have $\mathbf{y} \geq \mathbf{0}$.

Since $\mathbf{A}$ is irreducible, we can find $n \in \mathbb{N}_0$ such that $(\mathbf{A}+\mathbf{I})^n > \mathbf{0}$. Let $\mathbf{z} := (\mathbf{A}+\mathbf{I})^n\mathbf{x}$. Due to $(\mathbf{A}+\mathbf{I})^n > \mathbf{0}$ and $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$, we have $\mathbf{z} > \mathbf{0}$.

Since λ is maximal, we obtain

$$\begin{aligned}\mathbf{A}\mathbf{z} - r(\mathbf{z})\mathbf{z} \geq \mathbf{A}\mathbf{z} - \lambda\mathbf{z} &= (\mathbf{A}+\mathbf{I})\mathbf{z} - (\lambda+1)\mathbf{z} = (\mathbf{A}+\mathbf{I})^{n+1}\mathbf{x} - (\lambda+1)(\mathbf{A}+\mathbf{I})^n\mathbf{x} \\ &= (\mathbf{A}+\mathbf{I})^n((\mathbf{A}+\mathbf{I})\mathbf{x} - (\lambda+1)\mathbf{x}) = (\mathbf{A}+\mathbf{I})^n(\mathbf{A}\mathbf{x} - \lambda\mathbf{x}) = (\mathbf{A}+\mathbf{I})^n\mathbf{y}.\end{aligned}$$

By definition (A.1), we can find $i \in \mathcal{I}$ such that

$$r(\mathbf{z}) = \frac{(\mathbf{A}\mathbf{z})_i}{z_i}, \qquad ((\mathbf{A}+\mathbf{I})^n\mathbf{y})_i = (\mathbf{A}\mathbf{z} - r(\mathbf{z})\mathbf{z})_i = 0.$$

Due to $(\mathbf{A}+\mathbf{I})^n > \mathbf{0}$, this gives us $\mathbf{y} = \mathbf{0}$, so $\mathbf{x}$ is an eigenvector for the eigenvalue λ.

The definition (A.1) implies $\lambda = r(\mathbf{x}) \geq 0$, and

$$(\lambda+1)^n \mathbf{x} = (\mathbf{A}+\mathbf{I})^n \mathbf{x} = \mathbf{z} > \mathbf{0}$$

yields $\mathbf{x} > \mathbf{0}$. ■

Theorem A.11 (Perron-Frobenius) *Let $\mathbf{A} \geq \mathbf{0}$ be irreducible. The spectral radius $\varrho(\mathbf{A})$ of $\mathbf{A}$ is an eigenvalue with an eigenvector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ satisfying $\mathbf{x} > \mathbf{0}$.*

$\varrho(\mathbf{A})$ is the maximum of the function r introduced in (A.1).

Proof. Lemma A.5 gives us a vector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ such that

$$r(\mathbf{z}) \leq r(\mathbf{x}) \qquad\qquad \text{for all } \mathbf{z} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}.$$

Due to Lemma A.10, this vector $\mathbf{x}$ is an eigenvector of the matrix $\mathbf{A}$ for the eigenvalue $\lambda = r(\mathbf{x}) \geq 0$ with $\mathbf{x} > \mathbf{0}$.

Due to (A.2c), we have $\varrho(\mathbf{A}) \leq \lambda$. Since λ is an eigenvalue, we also have $\lambda = |\lambda| \leq \varrho(\mathbf{A})$, and can conclude $\lambda = \varrho(\mathbf{A})$. ■

Exercise A.12 (Stochastic matrix) *Let $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ be a* (left) stochastic matrix, *i.e., a matrix satisfying $\mathbf{A} \geq \mathbf{0}$ and*

$$\sum_{i\in\mathcal{I}} a_{ij} = 1 \qquad\qquad \textit{for all } j \in \mathcal{I}.$$

Prove $\varrho(\mathbf{A}) = 1$ (Hint: $\lambda \in \sigma(\mathbf{A}) \iff \bar{\lambda} \in \sigma(\mathbf{A}^*)$).

Assuming that $\mathbf{A}$ is irreducible, show that there is a vector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ with $\mathbf{x} > \mathbf{0}$ and

$$\mathbf{A}\mathbf{x} = \mathbf{x}, \qquad\qquad \sum_{i\in\mathcal{I}} x_i = 1.$$

Background: *If the elements of $\mathcal{I}$ are the states of a Markov chain, a vector $\mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ with $\mathbf{y} \geq \mathbf{0}$ and $\sum_{i\in\mathcal{I}} y_i = 1$ corresponds to a probability distribution for the states. If the coefficients a_{ij} are the probabilities for switching from state j to state i, $\mathbf{A}\mathbf{y}$ gives us the probability distribution after one step of the Markov chain.*

The conditions above ensure that there is an invariant probability distribution, *i.e., a distribution that will not change as the Markov chain progresses. If $\mathbf{A}$ is irreducible, i.e., if it is possible for the Markov chain to reach any state from any other state with non-zero probability, it is even possible to prove that the invariant probability distribution is unique and that the sequence of probability distributions obtained by starting with an arbitrary distribution and stepping through the Markov chain converges to the invariant distribution.*

Theorem A.11 provides us with the tool we need to obtain a generalized stability result for finite difference discretizations.

Definition A.13 (Irreducibly diagonally dominant) *A matrix* $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}}$ *is called* irreducibly diagonally dominant *if it is irreducible, weakly diagonally dominant, and if there is an index* $k \in \mathcal{I}$ *such that*

$$\sum_{\substack{j\in\mathcal{I}\\ j\neq k}} |a_{kj}| < |a_{kk}|.$$

Lemma A.14 (Spectral radius) *Let* $\mathbf{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ *be irreducibly diagonally dominant, and let* $\widehat{\mathbf{A}} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ *be defined by*

$$\widehat{a}_{ij} = |a_{ij}| \qquad\qquad \textit{for all } i,j \in \mathcal{I}.$$

Then $\varrho(\mathbf{A}) \leq \varrho(\widehat{\mathbf{A}})$*.*

Proof. Let $\lambda \in \sigma(\mathbf{A})$, and let $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ be a corresponding eigenvalue.

We define $\widehat{\mathbf{e}} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \setminus \{\mathbf{0}\}$ by

$$\widehat{e}_i = |e_i| \qquad\qquad \text{for all } i \in \mathcal{I}.$$

Let $i \in \mathcal{I}$ with $|e_i| \neq 0$. Then we have $e_i \neq 0$ and the triangle inequality yields

$$|\lambda| = \frac{|(\mathbf{A}\mathbf{e})_i|}{|e_i|} = \frac{1}{\widehat{e}_i}\left|\sum_{j\in\mathcal{I}} a_{ij}e_j\right| \leq \frac{1}{\widehat{e}_i}\sum_{j\in\mathcal{I}} |a_{ij}||e_j| = \frac{(\widehat{\mathbf{A}}\widehat{\mathbf{e}})_i}{\widehat{e}_i}.$$

Introducing the function

$$\widehat{r}\colon \mathbb{R}^{\mathcal{I}}_{\geq 0} \to \mathbb{R}, \qquad\qquad \mathbf{x} \mapsto \min\left\{\frac{(\widehat{\mathbf{A}}\mathbf{x})_i}{x_i} \ :\ i \in \mathcal{I},\ x_i \neq 0\right\},$$

we conclude $|\lambda| \leq \widehat{r}(\widehat{\mathbf{e}})$. Theorem A.11 yields that $\varrho(\widehat{\mathbf{A}})$ is the maximum of $\widehat{r}$. ∎

Lemma A.15 (Generalized maximum principle) *Let* $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ *be irreducibly diagonally dominant, and let* $\widehat{\mathbf{M}} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ *be defined by*

$$\widehat{m}_{ij} = \begin{cases} |a_{ij}|/|a_{ii}| & \textit{if } i \neq j, \\ 0 & \textit{otherwise} \end{cases} \qquad\qquad \textit{for all } i,j \in \mathcal{I}.$$

If there is a vector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ *such that* $\mathbf{x} \leq \widehat{\mathbf{M}}\mathbf{x}$*,* $\mathbf{x}$ *has to be a constant vector.*

Proof. Let $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ be a vector satisfying $\mathbf{x} \leq \widehat{\mathbf{M}}\mathbf{x}$.

Let $\mu := \max\{x_i \ :\ i \in \mathcal{I}\}$, and let $q \in \mathcal{I}$ be an index with $\mu = x_q$.

We will prove that the existence of a connection of length $m \in \mathbb{N}_0$ from an index $i \in \mathcal{I}$ to q implies $x_i = \mu$ by induction.

Base case: If there is a connection of length $m = 0$, from $i \in \mathcal{I}$ to q, we have $i = q$ and therefore $x_i = x_q = \mu$.

Induction assumption: Let $m \in \mathbb{N}_0$ be chosen such that our claim holds for all connections of length m.

Induction step: Let $i \in \mathcal{I}$ be an index such that a connection $(i_\ell)_{\ell=0}^{m+1}$ from i to q exists. Let $k := i_1$. Obviously, $(i_\ell)_{\ell=1}^{m+1}$ is a connection of length m from k to q, and our assumption yields $x_k = \mu$. Since $\widehat{\mathbf{M}}\mathbf{x} \geq \mathbf{x}$ and since $\mathbf{A}$ is weakly diagonally dominant, we have

$$x_k \leq (\widehat{\mathbf{M}}\mathbf{x})_k = \frac{1}{|a_{kk}|} \sum_{\substack{j \in \mathcal{I} \\ j \neq k}} |a_{kj}| x_j \leq \frac{1}{|a_{kk}|} \sum_{\substack{j \in \mathcal{I} \\ j \neq k}} |a_{kj}| \mu \leq \mu = x_k$$

and conclude $x_j = \mu$ for all $j \in \mathcal{I}$ with $a_{kj} \neq 0$. Due to $a_{ki} = a_{i_1,i_0} \neq 0$, this implies $x_i = \mu$. ■

Lemma A.16 (Irreducibly diagonally dominant) *Let $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ be irreducibly diagonally dominant. Then $\varrho(\mathbf{M}) < 1$ holds for $\mathbf{M} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$ and $\mathbf{A}$ is invertible.*

Proof. Since $\mathbf{A}$ is irreducible, each row has to contain at least one non-zero element. Since $\mathbf{A}$ is weakly diagonally dominant, this implies that all diagonal elements are non-zero, so the diagonal matrix $\mathbf{D} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ given by

$$d_{ij} := \begin{cases} a_{ii} & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in \mathcal{I}$$

is invertible. We consider the matrix

$$\mathbf{M} := \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$$

and its non-negative version $\widehat{\mathbf{M}} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ given by

$$\widehat{m}_{ij} := |m_{ij}| = \begin{cases} |a_{ij}|/|a_{ii}| & \text{if } i \neq j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in \mathcal{I}.$$

Since $\mathbf{A}$ is irreducible and since $a_{ij} \neq 0$ implies $m_{ij} \neq 0$ for all $i, j \in \mathcal{I}$ with $i \neq j$, the matrices $\mathbf{M}$ and $\widehat{\mathbf{M}}$ are irreducible.

Due to Theorem A.11, we can find an eigenvector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ of $\widehat{\mathbf{M}}$ for the eigenvalue $\varrho(\widehat{\mathbf{M}})$ with $\mathbf{x} > \mathbf{0}$. We have to prove $\varrho(\widehat{\mathbf{M}}) < 1$.

Let $\mathbf{y} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ be a vector satisfying $\mathbf{y} \leq \widehat{\mathbf{M}}\mathbf{y}$. Due to Lemma A.15, this implies that there is a $\mu \in \mathbb{R}_{\geq 0}$ such that

$$y_i = \mu \qquad \text{for all } i \in \mathcal{I}.$$

Since $\mathbf{A}$ is irreducibly diagonally dominant, there is an index $k \in \mathcal{I}$ such that

$$\sum_{\substack{j \in \mathcal{I} \\ j \neq k}} |a_{kj}| < |a_{kk}|, \qquad \beta := \frac{1}{|a_{kk}|} \sum_{\substack{j \in \mathcal{I} \\ j \neq k}} |a_{kj}| < 1,$$

and we find

$$\mu = y_k \leq (\widehat{\mathbf{M}}\mathbf{y})_k = \frac{1}{|a_{kk}|} \sum_{j \in \mathcal{I} j \neq k} |a_{kj}| y_j = \frac{1}{|a_{kk}|} \sum_{j \in \mathcal{I} j \neq k} |a_{kj}| \mu = \beta \mu.$$

Due to $\beta < 1$, we can conclude $\mu = 0$ and $\mathbf{y} = \mathbf{0}$.

Since we have $\mathbf{x} > \mathbf{0}$, this implies $\mathbf{x} \not\leq \widehat{\mathbf{M}}\mathbf{x} = \varrho(\widehat{\mathbf{M}})\mathbf{x}$, i.e., $\varrho(\widehat{\mathbf{M}}) < 1$.

Lemma A.14 yields $\varrho(\mathbf{M}) \leq \varrho(\widehat{\mathbf{M}}) < 1$. Due to Corollary 2.23, the Neumann series for $\mathbf{M}$ converges and $\mathbf{I} - \mathbf{M} = \mathbf{D}^{-1}\mathbf{A}$ is invertible, so $\mathbf{A}$ itself also has to be invertible. ■

Index

Bibliography

[1] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods.* Springer, 1994.

[2] J. M. Burgers. Mathematical examples illustrating relations occurring in the theory of turbulent fluid motion. *Verhandelingen der Koninklijke Nederlandse Akademie van Wetenschappen, Afdeeling Natuurkunde, Reihe 1*, 17(2):1–53, 1939.

[3] J. Céa. Approximation variationelle des problèmes aux limites. *Ann. Inst. Fourier*, 14(2):345–444, 1964.

[4] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100:32–74, 1928.

[5] J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Cambridge Philos. Soc.*, 43:50–67, 1947.

[6] G. Frobenius. Ueber Matrizen aus nicht negativen Elementen. *Sitzungsber. Königl. Preuss. Akad. Wiss.*, pages 456–477, 1921.

[7] W. Hackbusch. *Elliptic Differential Equations. Theory and Numerical Treatment.* Springer-Verlag Berlin, 1992.

[8] N. G. Meyers and J. Serrin. $H = W$. *Proc. Nat. Acad. Sci.*, 51(6):1055–1056, 1964.

[9] O. Perron. Zur Theorie der Matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[10] Y. Saad. *Iterative Methods for Sparse Linear Systems.* Society for Industrial Mathematics, 2nd edition, 2003.

[11] W. Walter. *Gewöhnliche Differentialgleichungen.* Springer, 1993.

[12] H. Wielandt. Unzerlegbare, nicht negative Matrizen. *Mathematische Zeitschrift*, 52:642–648, 1950.