Lecture Notes of the Course

Numerical Mathematics II for Engineers

Technische Universität Berlin
Winter Term 2020/21
lectured by Dr. Manuel Landstorfer

TEXed by Julia Ullrich,
revised by Dr. Raphael Kruse, Dr. Matthias Voigt, Dr. Dirk Peschka

Latest changes: February 13, 2021

Preface

These lecture notes are based on a course given by Dr. Raphael Kruse in the winter terms 2015/16 and 2016/17 at Technische Universität Berlin. It is mostly based on material from earlier courses (before 2013) "Numerical Mathematics II for Engineers" that have been taught by Prof. Dr. Jörg Liesen and from later courses by Dr. Martin Eigel (2014/15), Dr. Raphael Kruse (2015/16, 2016/17), Dr. Matthias Voigt (2018/19), Dr. Dirk Peschka (2013/14, 2019/20). This course is aimed at advanced students in engineering programs that already have some familiarity with standard concepts from calculus in several variables, linear algebra, and numerical mathematics. Additional knowledge of ordinary and partial differential equations is not necessary but may help to follow the course. The four main parts of this course cover

- terminology and some basic theory for partial differential equations,
- the finite difference method,
- the finite element method,
- iterative (incomplete) solvers for high-dimensional linear equations.

In parallel to the lecture, theoretical assignments as well as programming assignments are handed out and require knowledge of a programming language. In contrast to courses that are primarily aimed at mathematicians, the proofs are not always fully rigorous and theorems do not aim to cover the highest level of generality. Instead the focus lies on the explanation of the underlying ideas and how to implement those in practise. At the end of the course the students should have developed an understanding of different numerical methods for partial differential equations and how to properly apply those in basic settings. Further, we explain standard terminology for the finite difference and finite element methods, so that students can easily read and digest more specialized literature on this topic.

The first version of the lecture notes has been typed by Julia Ullrich, who attended this course in the winter term 2015/16. Dr. Raphael Kruse performed some corrections and reformulations during the following teaching term in 2016/17. He also likes to express his gratitude to Rouven Glauert, Phillip Kretschmer, and Amey Nandkumar Vasulkar and all other students in the term 2016/17, who helped to improve the presentation and to find errors and misprints in these lecture notes. Dr. Matthias Voigt performed some additional changes and corrections to this document during the winter term 2018/19. Dr. Dirk Peschka included changes and additional material during the winter term 2019/20. [1]

[1] **Important:** These notes may still contain typos and errors. Please send a mail to the lecturer in case you notice misprints or errors or if a formulation is unclear. Any assistance is highly appreciated.

Contents

I. Theory of Partial Differential Equations

I.1. Introduction

This first chapter seeks to familiarize the reader with different concepts and notation useful to study partial differential equations. Where appropriate we will abbreviate partial differential equation(s) with PDE(s). A partial differential equation consists of (a systems of) equations which relate an unknown function and its (partial) derivatives. As opposed to ordinary differential equations such as $y'(t) = F\big(y(t)\big)$ that mostly can be directly integrated, the nontrivial coupling of different partial derivatives makes the direct integration impossible most of the time. This is one of the properties that makes the analysis and numerical treatment of PDEs so interesting and challenging.

PDEs appear in many areas in the natural sciences and in engineering. Typical disciplines are solid & fluid mechanics, electrical engineering and electromagnetism, nonequilibrium thermodynamics, quantum field theories, and many more. Many laws in physics are expressed in terms of PDEs, for example: Maxwell's equation (in vacuum) for electromagnetic fields are given by

$$\begin{aligned} \nabla \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0}, \qquad & \nabla \times \mathbf{B} &= \mu_0 \mathbf{j} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}, \\ \nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} &= 0, \qquad & \nabla \cdot \mathbf{B} &= 0, \end{aligned} \tag{I.1a}$$

where one seeks the electrical and the magnetical (vector) field $\mathbf{E}(t, \mathbf{x})$ and $\mathbf{B}(t, \mathbf{x})$ for a given total charge density $\rho(t, \mathbf{x})$ and total current density $\mathbf{j}(t, \mathbf{x})$; the incompressible Navier-Stokes equations in fluid dynamics

$$\rho \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} \right) = -\nabla p + \mu \nabla^2 \mathbf{u} + \mathbf{f}, \qquad \nabla \cdot \mathbf{u} = 0, \tag{I.1b}$$

where ones seeks the velocity (vector) and pressure (scalar) fields $\mathbf{u}(t, \mathbf{x})$ and $p(t, \mathbf{x})$ for a given external (vector) force $\mathbf{f}(t, \mathbf{x})$; or the time-dependent Schrödinger equation from quantum mechanics

$$i\hbar \frac{\partial \Psi}{\partial t} = \left(\frac{-\hbar^2}{2m} \nabla^2 + V \right) \Psi \tag{I.1c}$$

where one seeks the complex-valued wave function $\Psi(t, \mathbf{x})$ for a given potential $V(t, \mathbf{x})$. A beautiful example of the pattern formation in the Navier-Stokes equation is shown in

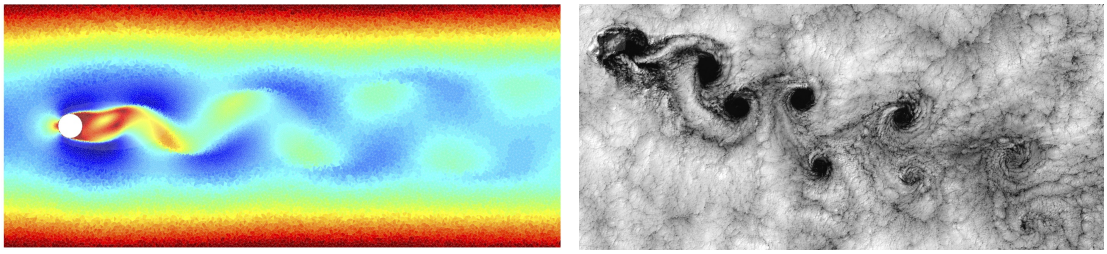


Figure I.1.: **(left)** Kármán vortex street as an instationary solution of the incompressible Navier-Stokes equation (I.1b) around an obstacle and **(right)** Landsat 7 image by NASA showing clouds near the Juan Fernandez Islands.

Figure I.1, where the longtime behavior at moderately large Reynolds numbers produces a repeating pattern of vortices detaching from an obstacle.

In the context of physics, PDEs are often derived from conservation laws and thereby express balance of mass, momentum or energy. In general, in a partial differential equation we seek functions depending on several variables by forming an equation from the function and its partial derivatives. In the examples above, the functions depend on both time $t \in \mathbb{R}$ and space $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$. The equations of (stationary) linear elasticity are

$$-\nabla \cdot (\mathbb{C} : \varepsilon) = \mathbf{f}, \qquad \text{where} \qquad \varepsilon = \tfrac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^\top), \tag{I.1d}$$

where we seek a stationary deformation field $\mathbf{u}(\mathbf{x})$, which only depends on space $\mathbf{x}$. Another example for a nonlinear PDE is the Cahn-Hilliard equation

$$\partial_t c = m\nabla^2(c^3 - c - \epsilon^2 \nabla^2 c), \tag{I.1e}$$

where we seek the phase-field $c(t, \mathbf{x})$. The Cahn-Hilliard equation is a well-studied example of a nonlinear PDE for diffusion and phase separation.

For each of the PDEs mentioned above lots of research has been devoted to the mathematical analysis, development of (sometimes highly specialized) numerical methods, model validation and so on, so that we can only expect to cover elementary but fundamental issues concerning PDE numerics.

The general framework for modeling with PDEs starting from real world problems and their numerical treatment is schematically shown in Figure I.2. For example, by using modeling principles from physics, i.e., conservation laws, variational principles and thermodynamic consistency, it is possible to model physical problems as a PDE. Using the theory of PDEs (which is not covered in too much detail here), one can state properties of the equations and their solutions. Most importantly, one can find conditions for *well-posedness* of the problem, meaning that solutions exist and are unique and that they depend continuously on the initial data. Since the PDE is still an infinite-dimensional problem, one performs a discretization of the problem to find an approximate solution to the problem in a finite-dimensional (but still high-dimensional) space. This leads to linear systems of equations with possibly millions of unknowns. On the other hand,

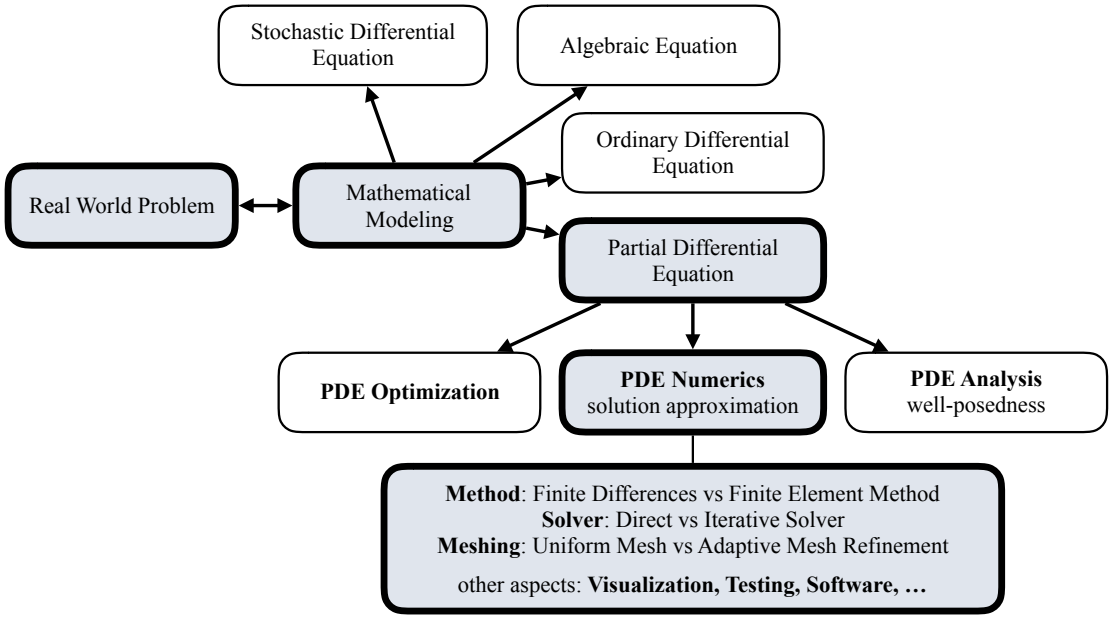


Figure I.2.: From Real World Problems to PDE Numerics

such linear systems typically have a lot of structure that can be exploited by numerical algorithms, e.g., sparsity or symmetry or linear systems.

It is important to know that in this general setting there are many sources for errors: The modeling step itself presents an unavoidable source of errors, since we always make certain idealizations about the behavior of a system. Model (order) reduction might be necessary to make a problem easier to solve in the context of a certain application but adds further modeling errors. Parameters and physical laws are determined by experiments, which have certain measurement errors and thereby lead to uncertainties in the solution process as well.

When solving the numerical problem, further errors are due to the discretization, algebraic errors as well as computational errors are present when solving a linear system on a computer due to the finite precision. Understanding the source and magnitude of such errors and developing error bounds and methods for their estimation are tough mathematical problems which are still a very active field of research.

In this course we will mainly discuss how to construct effective discretizations by employing the methods of finite differences and finite elements and we will also cover efficient solution methods for the resulting large systems of sparse linear equations. We will also discuss errors and discuss in what sense discrete solutions approximate the continuous problem.

I.2. Examples of PDEs

Above we have already seen a couple of important partial differential equations. Below we collect some further examples of PDEs, which will be also treated in this lecture. In the following $\Omega \subseteq \mathbb{R}^n$ denotes a spatial *domain*, i.e., an open, connected set. At least in a formal sense, a partial differential equation can be written as follows.

Definition I.1: Let $\Omega \subseteq \mathbb{R}^n$ be a *domain*. An expression of the form

$$F\left(D^k u(x), D^{k-1}u(x), \dots, Du(x), u(x), x\right) = 0 \quad \forall\, x \in \Omega,$$

where $D^j u$, $j = 1, \dots, k$ are the partial derivatives[1] of $u(x)$ of order j and where

$$F\colon \mathbb{R}^{(n^k)} \times \mathbb{R}^{(n^{k-1})} \times \dots \times \mathbb{R}^n \times \mathbb{R} \times \Omega \to \mathbb{R}$$

is given, is called a (scalar) *k-th order PDE* for the unknown $u\colon \Omega \to \mathbb{R}$. A function $u\colon \Omega \to \mathbb{R}$ that satisfies (I.1) is called a *solution of the PDE*. Later on we will discover that there are different concepts of solutions and that we need to narrow down the statement of the PDE to make it meaningful, i.e., to make the problem *well-posed*.

While this wonderful definition is found in every textbook on the subject, it is so general that it is rarely of any use beyond illustrating the general structure of a PDE. More concrete is the following definition of a linear PDE.

Definition I.2: A k-th order partial differential equation (I.1) of the form

$$\sum_{|\alpha| \le k} a_\alpha(x) D^\alpha u(x) = f(x),$$

with given coefficient functions $a_\alpha : \Omega \to \mathbb{R}$ and right-hand side $f : \Omega \to \mathbb{R}$ is called *linear*. For $f = 0$ the PDE is called *homogeneous*, otherwise *inhomogeneous*.

This definition is more useful since it considerably restricts the class of possible PDEs in (I.1). Furthermore it is clear that the linearity property allows to add an inhomogeneous and a homogeneous solution to obtain a new inhomogeneous solution of the PDE. Below we state some examples of linear PDEs.

a) Let $\Omega \subseteq \mathbb{R}^n$ and $f\colon \Omega \to \mathbb{R}$ be given domain and right-hand side. Then, the *Poisson equation* is given by the following 2-nd order PDE:

$$\begin{cases} \text{Find } u\colon \Omega \to \mathbb{R} \text{ such that} \\ -\Delta u(x) = f(x) \text{ for all } x \in \Omega. \end{cases} \tag{I.2}$$

If $f(x) = 0$ for all $x \in \Omega$ then the homogeneous problem reads

$$\begin{cases} \text{Find } u\colon \Omega \to \mathbb{R} \text{ such that} \\ -\Delta u(x) = 0 \text{ for all } x \in \Omega. \end{cases} \tag{I.3}$$

and is called the *Laplace equation* (or *homogeneous* Poisson equation). These equations have many applications. They appear, for example,

[1] D^j is understood in the sense of multiindices, e.g. cf. [Eva98]

- as a potential equation in fluid dynamics,
- in electrostatics,
- as a shallow slope approximation for minimal surfaces,
- in the membrane problem in mechanical engineering,
- as a stationary (time constant) solution for heat transport/diffusion.

For $\Omega \subset \mathbb{R}^2$ and $\Delta \equiv \nabla^2 = \operatorname{div}(\operatorname{grad}(\cdot)) = \partial_x^2 + \partial_y^2$, the Laplace equation also appears in complex analysis: Let $\Omega \subseteq \mathbb{C} \simeq \mathbb{R}^2$ be open and $f\colon \Omega \to \mathbb{C}$ be a complex-valued mapping. The mapping f is called (complex) differentiable in $z_0 \in \Omega \subseteq \mathbb{C}$ if the limit $\lim_{z\to z_0} \frac{f(z)-f(z_0)}{z-z_0}$ exists in $\mathbb{C}$. Then the limit is denoted, as usual, by $f'(z_0)$. In particular,

$$f'(z_0) = \lim_{h\to 0,\, h\in\mathbb{R}} \frac{f(z_0+h)-f(z_0)}{h} = \lim_{h\to 0,\, h\in\mathbb{R}} \frac{f(z_0+\mathrm{i}h)-f(z_0)}{\mathrm{i}h}. \tag{I.4}$$

We write $z = x + \mathrm{i}y$ for $x, y \in \mathbb{R}$ and $\operatorname{Re}(z) = x$, $\operatorname{Im}(z) = y$. Then we obtain $f(z) = f(x,y) = u(x,y) + \mathrm{i}v(x,y)$ where $u(x,y) = \operatorname{Re}(f(x,y))$, $v(x,y) = \operatorname{Im}(f(x,y))$. Inserting this into (I.4) yields

$$\begin{aligned}\lim_{h\to 0} \frac{1}{h} &\left(u(x_0+h,y_0) + \mathrm{i}v(x_0+h,y_0) - u(x_0,y_0) - \mathrm{i}v(x_0,y_0)\right) \\ &= \lim_{h\to 0} \frac{1}{\mathrm{i}h} \left(u(x_0,y_0+h) + \mathrm{i}v(x_0,y_0+h) - u(x_0,y_0) - \mathrm{i}v(x_0,y_0)\right).\end{aligned}$$

This leads to the equation

$$\frac{\partial u}{\partial x}(x_0,y_0) + \mathrm{i}\frac{\partial v}{\partial x}(x_0,y_0) = \frac{1}{\mathrm{i}}\frac{\partial u}{\partial y}(x_0,y_0) + \frac{\partial v}{\partial y}(x_0,y_0),$$

or, by comparing the real and imaginary parts separately,

$$\begin{aligned}\frac{\partial u}{\partial x}(x_0,y_0) - \frac{\partial v}{\partial y}(x_0,y_0) &= 0,\\ \frac{\partial v}{\partial x}(x_0,y_0) + \frac{\partial u}{\partial y}(x_0,y_0) &= 0.\end{aligned}$$

These two equations are called the *Cauchy-Riemann equations*.

To sum up, if $f\colon \Omega \to \mathbb{C}$, $f := u + \mathrm{i}v$ is complex differentiable, then it holds true that

$$u_x - v_y = 0 \text{ and } v_x + u_y = 0 \text{ in } \Omega.$$

Now, if we take the partial derivative with respect to x of the first equation and with respect to y of the second equation and sum both equations we obtain

$$u_{xx} - v_{yx} + u_{yy} + v_{xy} = 0.$$

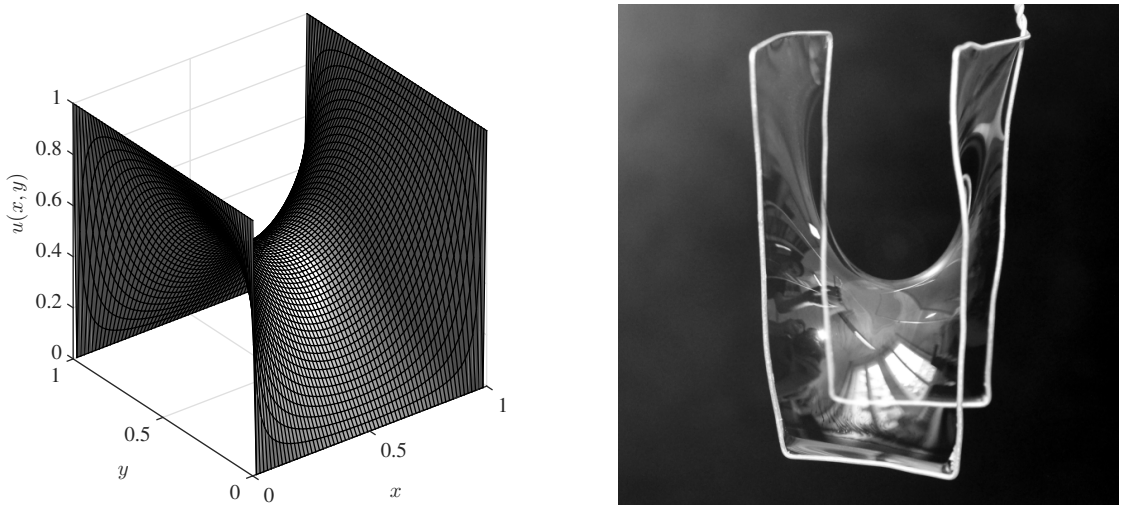


Figure I.3.: **(left)** Solution $u(x, y)$ of 2-dimensional homogeneous Poisson equation (I.5) on $\Omega = [0, 1]^2$ with boundary conditions $g(0, y) = g(1, y) = 1$ for $0 < y < 1$ and $g(x, 0) = g(x, 1) = 0$ for $0 < x < 1$ and **(right)** corresponding minimal (soap) surface on a wire. Note, the real minimal surface equation would be the nonlinear PDE $\nabla \cdot \left(\frac{\nabla u}{(1+|\nabla u|^2)^{1/2}}\right) = 0$.

Now, recall that the Theorem of Schwarz shows

$$v_{xy} = v_{yx},$$

since $v \in C^\infty(\mathbb{R}^2)$. From this it follows that

$$\Delta u = u_{xx} + u_{yy} = 0.$$

This gives the following result: The real part of every in Ω complex differentiable (holomorphic, analytic) function $f = u + \mathrm{i}v$ is a solution of the Laplace equation. Consequently, the solution to problem (I.3) is *not* unique.

The solutions $u\colon \Omega \to \mathbb{R}$ of the Laplace equation are often called (scalar) *potentials* or *harmonic functions*.

In order to obtain a *unique solution* we need to impose further conditions. In the case of the Poisson equation we usually impose *boundary conditions*. Thus, the problem (I.2) is extended as follows:

$$\begin{cases} \text{Find } u\colon \Omega \to \mathbb{R} & \text{such that} \\ -\Delta u(x) = f(x) & \text{for all } x \in \Omega, \\ u(x) = g(x) & \text{for all } x \in \partial\Omega, \end{cases} \tag{I.5}$$

where $g\colon \partial\Omega \to \mathbb{R}$ is defined on the boundary $\partial\Omega$ of Ω. To shorten the notation, we often suppress the explicit dependence of u and f on $x \in \Omega$. Hence, the following

problem is just a shorter version of (I.5):

$$\begin{cases} \text{Find } u\colon \Omega \to \mathbb{R} & \text{such that} \\ -\Delta u = f & \text{in } \Omega, \\ u = g & \text{on } \partial\Omega. \end{cases} \tag{I.6}$$

The boundary conditions in (I.5) and (I.6) are called *Dirichlet boundary conditions*.

b) Let $\Omega \subseteq \mathbb{R}^n$ be a domain (the physical space) and let $[0,T]$ be an interval (the time axis). The *heat equation* or *diffusion equation* is then given by the problem

$$\begin{cases} \text{Find } u\colon [0,T] \times \Omega \to \mathbb{R} & \text{such that} \\ \frac{\partial u}{\partial t} - \Delta u = f & \text{in } (0,T) \times \Omega. \end{cases} \tag{I.7}$$

Here, it is important to note that $\Delta u = \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2}$ does not contain derivatives with respect to time. In this setting, $u(t,x)$ can be interpreted as a temperature or particle density at position $x \in \Omega$ at time $t \in [0,T]$.

In order to ensure the uniqueness of the solution one usually imposes *initial conditions* and *boundary conditions*. For example, if $\Omega = (0,1)$, we might have the problem

$$\begin{cases} \text{Find } u\colon [0,T] \times \Omega \to \mathbb{R} & \text{such that} \\ \frac{\partial u}{\partial t} - \Delta u = f & \text{in } (0,T) \times (0,1), \\ u(t,0) = u_1(t) & \text{in } (0,T) \quad \text{(boundary conditions)}, \\ u(t,1) = u_2(t) & \text{in } (0,T) \quad \text{(boundary conditions)}, \\ u(0,x) = u_0(x) & \text{in } (0,1) \quad \text{(initial conditions)}. \end{cases} \tag{I.8}$$

Hereby, the mapping $u_0\colon \Omega \to \mathbb{R}$ denotes the initial condition. In heat conduction, one might interpret the Dirichlet boundary condition as an exterior heat source (or cooling device) that only affects the boundary of Ω. Of course, it might be possible that such a device is switched off or on, which is expressed in terms of the t-dependence of boundary conditions u_1 and u_2.

Alternatively, one might impose *Neumann boundary conditions* in order to model a perfectly isolated boundary (meaning: no heat conduction/no particle flux over the boundary). This leads us to the problem

$$\begin{cases} \text{Find } u\colon [0,T] \times \Omega \to \mathbb{R} & \text{such that} \\ \frac{\partial u}{\partial t} - \Delta u = f & \text{in } (0,T) \times (0,1), \\ \frac{\partial u}{\partial x}(t,0) = \frac{\partial u}{\partial x}(t,1) = 0 & \text{in } (0,T), \\ u(0,x) = u_0(x) & \text{in } (0,1). \end{cases} \tag{I.9}$$

c) Next, we introduce the *wave equation*. To this end, let $\Omega \subseteq \mathbb{R}^n$ be a domain. Then the linear wave equation is given by the PDE

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f \text{ in } (0,T) \times \Omega,$$

where $u\colon [0,T]\times\Omega\to\mathbb{R}$ is the unknown function. The 1-dimensional wave equation (the case $n=1$) simplifies to $u_{tt}-u_{xx}=f$.

The wave equation has been introduced by d'Alembert in 1746 to model a vibrating string. It has further applications, for instance in acoustics, electromagnetics, or solid mechanics.

As in the previous examples, the wave equation is not uniquely solvable without imposing further conditions. A typical setting, where one can expect a unique solution to the 1-dimensional wave equation with $\Omega=(0,1)$, is the following problem

$$\begin{cases} \text{Find } u\colon [0,T]\times\Omega\to\mathbb{R} & \text{such that} \\ u_{tt}-u_{xx}=f & \text{in } (0,T)\times(0,1), \\ u(t,0)=u_0,\ u(t,1)=u_1 & \text{in } (0,T) \text{ (boundary conditions)}, \\ u(0,x)=g_1(x) & \text{in } (0,1) \text{ (initial conditions)}, \\ u_t(0,x)=g_2(x) & \text{in } (0,1) \text{ (initial conditions)}. \end{cases} \tag{I.10}$$

d) Next we introduce the *linear transport equation*, which is probably the simplest among all presented linear PDEs. For this let $\Omega\subseteq\mathbb{R}$ be a domain and we seek $u\colon [0,T]\times\Omega\to\mathbb{R}$, which satisfies

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=0.$$

This model illustrates transport in the context of a partial differential equation.

e) As our final example we introduce the *(inviscid) Burgers equation* as a rather simple example for a *nonlinear partial differential equation*. For this let $\Omega\subseteq\mathbb{R}$ be a domain and we seek $u\colon [0,T]\times\Omega\to\mathbb{R}$, which satisfies

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=0.$$

This nonlinear PDE was first introduced by Bateman (1915) and studied later by Burgers (1948) in the context of turbulence. This inviscid variant is one example for a first order conservation law $\partial_t u+\partial_x(F(u))=0$, which even for smooth initial data can develop discontinuities after a finite time.

Now we show exemplarily how a PDE can be derived using variational arguments.

Example I.3 (Derivation of Minimal Surface PDE)**:** In this section we give a short derivation of a PDE by variational arguments. Quite some PDEs in physics can be derived from variational arguments (stationary action, minimal dissipation), often also conservation laws play an important role. We consider the minimal surface PDE, which is based on a surface $\Gamma\subset\mathbb{R}^{n+1}$ parametrized by a function $h:\Omega\subset\mathbb{R}^n\to\mathbb{R}$ so that

$$\Gamma=\{(x,z)\in\mathbb{R}^{n+1}: x\in\Omega, z=h(x)\in\mathbb{R}\}, \tag{I.11}$$

where we assume that the boundary values are fixed, i.e., $h(x) = g(x)$ for $x \in \partial\Omega$. The area A of the parametrized surface is given by

$$h \quad \mapsto \quad A[h] = \int_\Omega \sqrt{1 + |\nabla h|^2}\, \mathrm{d}x, \tag{I.12}$$

and is called a *functional*, i.e., for each function h we obtain the associated area as $A[h]$. In order to find the necessary condition for A to be minimal for a given surface h we consider perturbations

$$h_\delta(x) = h(x) + \delta u(x), \tag{I.13}$$

where $u = 0$ on $\partial\Omega$. Then the surface h is minimal, when the real-valued function

$$a(\delta) = A[h_\delta], \tag{I.14}$$

is minimal at $\delta = 0$ for all possible (and sufficiently smooth) perturbations u. We differentiate

$$\begin{aligned}
0 \overset{!}{=} a'(0) &= \lim_{\delta\to 0} \tfrac{1}{\delta} \left(A[h_\delta] - A[h]\right) \\
&= \int_\Omega \lim_{\delta\to 0} \frac{1}{\delta} \left(\sqrt{1 + |\nabla(h+\delta u)|^2} - \sqrt{1 + |\nabla h|^2}\right) \mathrm{d}x \\
&= \int_\Omega \frac{\nabla h \cdot \nabla u}{\sqrt{1 + |\nabla h|^2}}\, \mathrm{d}x \\
&= \int_\Omega -\nabla \cdot \left(\frac{\nabla h}{\sqrt{1 + |\nabla h|^2}}\right) u\, \mathrm{d}x
\end{aligned}$$

where we used a Taylor expansion around $\delta = 0$ and integration by parts and the boundary condition for u in the last step. Since this integral needs to vanish for all u, we obtain the minimal surface PDE

$$\nabla \cdot \left(\frac{\nabla h}{\sqrt{1 + |\nabla h|^2}}\right) = 0. \tag{I.15}$$

This list of partial differential equations is far from being extensive. For a somewhat longer list of PDEs including technical details we refer to the handbook by Zwillinger [Zwi98].

Note: The Definition I.1 of abstract PDEs is written for domains Ω. At first glance, this seems to exclude functions $u(t, x)$ that depend on both time $t \in [0, T]$ and space $x \in \bar{\Omega} \subset \mathbb{R}^n$. However, all the concepts should be understood by setting $\Omega = Q_T := [0, T] \times \bar{\Omega} \subset \mathbb{R}^{n+1}$ as the domain for the PDE. Then the notation Ω and $\bar{\Omega}$ is slightly ambiguous and should be deduced from the context. This also clarifies that the distinct properties of the variable t (time) should follow from the structure of the PDE rather from the naming convention in the expression F in Definition I.1.

I.3. Notation and Basic Terminology

Basic Notation:

a) **Integers**: By $\mathbb{N} = \{1, 2, \ldots\}$ we denote the set of all positive integers. If we also include the zero integer, we write $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

b) **Euclidean space**: By $(\mathbb{R}^n, \langle\cdot,\cdot\rangle, \|\cdot\|)$ with $n \in \mathbb{N}$, we denote the standard Euclidean vector space. More precisely, it is the n-dimensional vector space of real column vectors with the usual operations. For two vectors $x, y \in \mathbb{R}^n$ the standard inner product is defined as

$$\langle x, y\rangle = \sum_{j=1}^{n} x_j y_j = x \cdot y,$$

where $x = \big[x_1, \ldots, x_n\big]^{\mathsf{T}}$ and $y = \big[y_1, \ldots, y_n\big]^{\mathsf{T}}$. Hereby, $(\cdot)^{\mathsf{T}}$ denotes the transposed vector, that is

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \big[x_1, \ldots, x_n\big]^{\mathsf{T}}.$$

The standard Euclidean norm is given by

$$\|x\| = \sqrt{\langle x, x\rangle} = \left(\sum_{j=1}^{n} x_j^2\right)^{\frac{1}{2}}$$

for all $x \in \mathbb{R}^n$.

c) **Domain**: Considering a domain $\Omega \subset \mathbb{R}^n$, then $\partial\Omega$ denotes its *boundary*. Two points $x, y \in \Omega$ have the distance $\|x - y\| = \sqrt{\langle x - y, x - y\rangle}$. The domain is *open* (in the sense of metric spaces), if with $x \in \Omega$ there exists a real number $\varepsilon > 0$ such that all point $y \in \mathbb{R}^n$ with $\|x - y\| < \varepsilon$ are also in Ω. We denote $\overline{\Omega} = \Omega \cup \partial\Omega$ the *closure* of the domain/set Ω. The domain is a *Lipschitz domain*, if the boundary is (locally) the graph of a Lipschitz continuous function. Similarly we can define domains with other regularity properties. With $\nu : \partial\Omega \to \mathbb{R}^n$ we denote the *outer normal*. The domain $\Omega \subset \mathbb{R}^n$ is *bounded* if there exists an $R < \infty$ and $x \in \mathbb{R}^n$ such that $\|x - y\| \leq R$ for all $y \in \Omega$. Note that components of the coordinate vector $x \in \Omega$ will be denoted by x for $n = 1$, (x, y) for $n = 2$, and (x, y, z) for $n = 3$. In general we will use $x = (x_1, \ldots, x_n)^{\mathsf{T}}$ with lower indices.

d) **Function**: Let $u\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}^k$. For $k = 1$ we say *u is a scalar function*, for $k > 1$ a vector-valued function (in particular for $k = n$). We write $u \in C^k(\Omega)$ if u is k times differentiable and if all k-th partial derivatives are continuous functions.

e) **Partial derivatives**: Let $f\colon \Omega \to \mathbb{R}^k$ with a domain $\Omega \subseteq \mathbb{R}^n$ and $n, k \in \mathbb{N}$ be a mapping and let $x = \big[x_1, \dots, x_n\big]^\mathsf{T} \in \Omega$. If the limit $\lim_{h\to 0} \frac{1}{h}(f(x + he_m) - f(x))$ exists with $e_m = \big[0, \dots, 0, 1, 0, \dots, 0\big]^\mathsf{T} \in \mathbb{R}^n$ with 1 in the m-th component, $1 \leq m \leq n$, then f is called differentiable in x in direction e_m. The limit is called the *partial derivative* with respect to the variable x_m. It is usually denoted by

$$\frac{\partial f}{\partial x_m} \qquad \text{or} \qquad \partial_{x_m} f \qquad \text{or} \qquad f_{x_m}.$$

Similarly higher-order derivatives for $1 \leq i, j \leq n$ are written for example as

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \qquad \text{or} \qquad \partial_{x_i}\partial_{x_j} f = \partial_{x_i x_j} f \qquad \text{or} \qquad f_{x_i x_j}.$$

For $i = j$ we also write $\partial^2_{x_i} f$. Be careful not to confuse f_{x_i} with the i-th component of a vector f^i, which we denote by upper indices. Whenever this seems helpful, we will use boldface $\mathbf{f} : \Omega \to \mathbb{R}^n$ for vector or tensor fields. Note that while computing the partial derivative, all other arguments are considered fixed. This also implies that upon change of variables the chain rule needs to be employed in order to transform the derivative into the new variables.

- Example 1: Let $u\colon (0,T) \times \mathbb{R} \to \mathbb{R}$ be a mapping on $\Omega = (0,T) \times \mathbb{R}$. Then the two equations

$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial^2 u}{\partial x^2}(t,x) \quad \forall\, t \in (0,T),\, x \in \mathbb{R}$$

and

$$u_t(t,x) = u_{xx}(t,x) \quad \forall\, t \in (0,T),\, x \in \mathbb{R}$$

denote the same PDE.

- Example 2: Consider the change of variables/coordinates $F : \mathbb{R} \times (0, 2\pi) \to \mathbb{R}^2$ defined by $F(r,\phi) = r(\cos\phi, \sin\phi)^\mathsf{T}$ from polar to Cartesian coordinates and let $u : \mathbb{R}^2 \to \mathbb{R}$ a solution of the Poisson equation $-\nabla^2 u = f$ in Cartesian coordinates. Then $\bar{u} : \mathbb{R} \times (0, 2\pi) \to \mathbb{R}$ defined by $\bar{u}(r,\phi) = u(F(r,\phi))$ solves

$$-\Delta_{r,\phi}\bar{u} = -\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \bar{u}}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 \bar{u}}{\partial \phi^2}\right) = f(F(r,\phi)), \tag{I.16}$$

and $\Delta_{r,\phi}$ denotes the Laplacian (operator) in polar coordinates.

f) **Multiindex notation**: For a given *multiindex* $\alpha \in \mathbb{N}_0^n$ and $f\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}^k$ let

$$D^\alpha f(x) = \frac{\partial^{\alpha_1}}{\partial x_1}\frac{\partial^{\alpha_2}}{\partial x_2}\cdots\frac{\partial^{\alpha_n}}{\partial x_n} f(x) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n}} f(x) = \prod_{i=1}^{n} \frac{\partial^{\alpha_i}}{\partial x_i} f(x),$$

be a mixed derivative. The order of the multiindex (and the mixed partial derivative) is $|\alpha| = \alpha_1 + \ldots + \alpha_n$. Sometimes one also writes ∂^α instead of D^α. We also have $\alpha! = \alpha_1! \cdot \ldots \alpha_n!$ and $x^\alpha = x_1^{\alpha_1} \cdot \ldots \cdot x_n^{\alpha_n}$ for any $x \in \mathbb{R}^n$. For any $k \in \mathbb{N}$ we also denote the set of all partial derivatives of order k by $D^k f = \{D^\alpha f : |\alpha| = k\}$.

g) **Jacobian matrix**: Partial derivatives are understood component-wise, i.e., if we have a vector-valued function $\mathbf{f}(x) = \left[f^1(x), \ldots, f^k(x)\right]^\mathsf{T}$, then

$$\frac{\partial \mathbf{f}}{\partial x_m} = \begin{bmatrix} \frac{\partial f^1}{\partial x_m} \\ \vdots \\ \frac{\partial f^k}{\partial x_m} \end{bmatrix}.$$

The matrix containing all partial derivatives, that is

$$D^1\mathbf{f} \equiv D\mathbf{f} = \begin{bmatrix} \frac{\partial f^1}{\partial x_1} & \frac{\partial f^1}{\partial x_2} & \cdots & \frac{\partial f^1}{\partial x_n} \\ \frac{\partial f^2}{\partial x_1} & \frac{\partial f^2}{\partial x_2} & \cdots & \frac{\partial f^2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f^k}{\partial x_1} & \frac{\partial f^k}{\partial x_2} & \cdots & \frac{\partial f^k}{\partial x_n} \end{bmatrix},$$

is called *Jacobian matrix*. Note that $D\mathbf{f}\colon \Omega \to \mathbb{R}^{k\times n}$ is a matrix-valued mapping. Sometimes the symbols $J_\mathbf{f}$, $\vec{\nabla}\mathbf{f}$ or $\boldsymbol{\nabla}\mathbf{f}$ are also used to denote the *Jacobian* of $\mathbf{f}$.

h) **Differential operators**: If $k = 1$, that is $f\colon \Omega \to \mathbb{R}$, then the transposed Jacobian matrix in Euclidean coordinates is called the *gradient* of f, denoted by

$$\nabla f = \operatorname{grad}(f) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = (Df)^\mathsf{T}.$$

Formally, we define the *nabla operator* by

$$\nabla := \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{bmatrix}.$$

As indicated above, in different coordinates the gradient operator will transform accordingly. If $k = n$, then the Jacobian $D\mathbf{f}$ is a quadratic matrix and the trace of $D\mathbf{f}$ is called the *divergence*, given by

$$\operatorname{div}(\mathbf{f}) = \operatorname{tr}(D\mathbf{f}) = \frac{\partial f^1}{\partial x_1} + \ldots + \frac{\partial f^n}{\partial x_n} = \sum_{j=1}^{n} \frac{\partial f^j}{\partial x_j}.$$

For scalar functions, the divergence of the gradient gives the *Laplace operator*

$$\Delta f := \operatorname{div}(\operatorname{grad} f) = \sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2} f.$$

It is also quite common to write $\nabla^2 f$ instead of Δf. The Laplacian of a vector field $\mathbf{f} : \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ with components $\mathbf{f} = (f^x, f^y, f^z)$ is defined as

$$\Delta \mathbf{f} = \nabla(\nabla \cdot \mathbf{f}) - \nabla \times (\nabla \times \mathbf{f}),$$

and reduces to $\Delta \mathbf{f} = (\Delta f^x, \Delta f^y, \Delta f^z)^\mathsf{T}$ in Cartesian coordinates. Here, the differential operator $\nabla \times \mathbf{f} = \mathrm{curl}(\mathbf{f})$ denotes the curl of a vector field defined as

$$\nabla \times \mathbf{f} \equiv \mathrm{curl}(\mathbf{f}) = \begin{pmatrix} \partial_y f^z - \partial_z f^y \\ \partial_z f^x - \partial_x f^z \\ \partial_x f^y - \partial_y f^x \end{pmatrix}.$$

Another common differential operator is the directional derivative. Given a scalar function $f \colon \Omega \subset \mathbb{R}^n \to \mathbb{R}$ or a vector field $\mathbf{f} \colon \Omega \to \mathbb{R}^n$ and a vector field $\mathbf{u} : \Omega \to \mathbb{R}^n$, then the *directional derivative of* f *(or* $\mathbf{f}$*) in the direction of* $\mathbf{u}$ *is defined*

$$\mathbf{u}(x) \cdot \nabla f(x) = \sum_{i=1}^{n} \mathbf{u}^i(x) \frac{\partial f(x)}{\partial x_i}, \qquad \text{or} \qquad \mathbf{u}(x) \cdot \nabla \mathbf{f}(x) = \sum_{i=1}^{n} \mathbf{u}^i(x) \frac{\partial \mathbf{f}(x)}{\partial x_i}.$$

Alternatively, variants such as $\nabla_{\mathbf{u}} f$ or $Df(x)(\mathbf{u})$ can be found in literature.

I.4. Definitions and Classifications of PDEs

In this section we introduce some definitions and terminology as well as a classification of PDEs. In this lecture we are mostly concerned with PDEs of the following form.

Definition I.4: Let $\Omega \subseteq \mathbb{R}^n$ be a domain. A *linear, second-order PDE* for the unknown $u \colon \Omega \to \mathbb{R}$ is given by

$$-\sum_{i,k=1}^{n} a_{ik}(x) u_{x_i x_k}(x) + \sum_{i=1}^{n} b_i(x) u_{x_i}(x) + c(x)u(x) = f(x) \quad \forall\, x \in \Omega, \tag{I.17}$$

where a_{ik}, b_i, c, $f \colon \Omega \to \mathbb{R}$ are given functions for $i, k = 1, \ldots, n$. If a_{ik}, b_i, c are independent of $x \in \Omega$ we say that (I.17) has *constant coefficients*, otherwise it has *variable coefficients*. If $f(x) = 0$ for all $x \in \Omega$, then we say that (I.17) is *homogeneous*, else (there exists an $x \in \Omega$ with $f(x) \neq 0$) *inhomogeneous*. If one variable is interpreted as "time", we call the PDE *instationary*, otherwise it is called *stationary*.

Remark I.5: a) The PDE (I.17) is a linear combination of expressions of the unknown function u and its derivatives, which is why it is called linear.

b) Assuming that u is two times continuously differentiable, we have by the theorem of Schwarz that

$$u_{x_i x_k} = u_{x_k x_i}.$$

We can therefore always assume (without loss of generality) that in (I.17) it holds

$$a_{ik}(x) = a_{ki}(x) \quad \text{for all } x \in \Omega.$$

In this case, the matrix

$$A(x) = \begin{bmatrix} a_{11}(x) & \cdots & a_{1n}(x) \\ \vdots & & \vdots \\ a_{n1}(x) & \cdots & a_{nn}(x) \end{bmatrix}$$

is called the *diffusion matrix* of (I.17).

Since we have $a_{ki}(x) = a_{ik}(x)$ for all $x \in \Omega$ it follows that $A(x) = A(x)^\mathsf{T}$ is symmetric. Therefore, this matrix has n real eigenvalues for every $x \in \Omega$ (not necessarily distinct). This will be important in the following definition.

Definition I.6 (Classification of linear, second-order PDEs)**:** A linear, second-order PDE of the form (I.17) is called

a) *elliptic* in $x \in \Omega$ if $A(x)$ is definite, i. e., all eigenvalues of $A(x)$ are either strictly positive or strictly negative.

b) *hyperbolic* in $x \in \Omega$ if $A(x)$ has one strictly negative eigenvalue and $n-1$ strictly positive eigenvalues (or vice versa).

c) *parabolic* in $x \in \Omega$ if $A(x)$ has one eigenvalue equal to zero and $n-1$ strictly positive (or strictly negative) eigenvalues and $\operatorname{rank}\big(\begin{bmatrix} A(x) & b(x) \end{bmatrix}\big) = n$, where $b(x) = \begin{bmatrix} b_1(x), \ldots, b_n(x) \end{bmatrix}^\mathsf{T}$.

We say that the PDE is elliptic/hyperbolic/parabolic if it is elliptic/hyperbolic/parabolic in all $x \in \Omega$.

Remark I.7: a) The definition does not cover all possible situations of (I.17). If (I.17) is not elliptic/hyperbolic/parabolic we say that (I.17) is *unclassified*.

b) Note that the linear PDE

$$-(u_{x_1x_1} + u_{x_2x_1} + u_{x_2x_2}) = 0$$

does *not* give the diffusion matrix $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ but because of $u_{x_1x_2} = u_{x_2x_1}$ we obtain $A = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$.

Example I.8: a) The Poisson equation $-\Delta u = f$:

This is a linear second-order PDE, it is stationary and inhomogeneous (if $f \neq 0$). It has constant coefficients and the diffusion matrix reads

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The eigenvalues of this matrix are $\lambda_1 = \lambda_2 = 1$. Therefore, A is positive definite and, consequently, the Poisson equation is elliptic.

b) The heat equation $u_t - u_{xx} = f$:

This leads (with $u = u(t,x)$) to the coefficients $a_{11} = 0$, $a_{12} = 0$, $a_{21} = 0$, $a_{22} = 1$, $b_1 = 1$, $b_2 = 0$, $c = 0$. Consequently, the heat equation is a second-order linear PDE with constant coefficients. It is instationary and the diffusion matrix is given by

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

This matrix has the eigenvalues 0 and 1. Since

$$\operatorname{rank}\left(\begin{bmatrix} A & b \end{bmatrix}\right) = \operatorname{rank}\left(\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}\right) = 2,$$

we see that the heat equation is parabolic.

c) The wave equation $u_{tt} - u_{xx} = f \Leftrightarrow -(u_{tt} - u_{xx}) = -f$ for $u = u(t,x)$:

This is a linear 2nd order PDE, with constant coefficients. It is instationary, and inhomogeneous if $f \neq 0$. The coefficients of the diffusion matrix are $a_{11} = 1$, $a_{12} = a_{21} = 0$, $a_{22} = -1$. Thus

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Therefore, the wave equation is hyperbolic.

We often write (I.17) as $Lu = f$ where L is the linear differential operator

$$L = \underbrace{-\sum_{i,k=1}^{n} a_{ik}(\cdot)\frac{\partial^2}{\partial x_i \partial x_k}}_{\text{the main part of } L} + \sum_{i=1}^{n} b_i(\cdot)\frac{\partial}{\partial x_i} + c(\cdot)$$

Note: The operator L is linear, $L(\lambda u + v) = \lambda L(u) + L(v)$ for all $\lambda \in \mathbb{R}$, and all twice differentiable $u, v \colon \Omega \to \mathbb{R}$. In particular, if u, v are solutions of an elliptic problem $Lu = f_u$, $Lv = f_v$ with compatible boundary conditions, then $L(u+v) = f_u + f_v$ with appropriate boundary conditions. This concept is useful, because when considering discretized problems the operator L will turn into a finite-dimensional linear operator, i.e., a matrix.

Definition I.9 (Boundary conditions)**:** Let $\Omega \subseteq \mathbb{R}^n$ be a domain $\Gamma \subset \partial\Omega$ a part of the boundary.

i) We call the condition

$$u = g \qquad \text{on } \Gamma$$

Dirichlet boundary condition (for $g = 0$ *homogeneous*). If the first coordinate has the interpretation of time and $\Gamma = \{0\} \times \Omega$, then we call the condition an *initial condition* instead.

ii) We call the condition
$$\nu \cdot \nabla u = g \qquad \text{on } \Gamma$$
Neumann boundary condition (for $g = 0$ *homogeneous*).
Note: Here $\nu : \Gamma \to \mathbb{R}^n$ denotes the outer normal vector field.

iii) We call the condition
$$\alpha u + \beta \nu \cdot \nabla u = g \qquad \text{on } \Gamma$$
Robin/mixed boundary condition (for $g = 0$ *homogeneous*).

iv) Let $\Omega = [0, L_1] \times \ldots \times [0, L_n] \subset \mathbb{R}^n$. Then we call the conditions
$$u(L_1, x_2, \ldots, x_n) = u(0, x_2, \ldots, x_n), \quad u(x_1, L_2, x_3, \ldots, x_n) = u(x_1, 0, x_3, \ldots, x_n), \ldots$$
for u (and possibly its derivatives) *periodic boundary conditions*. This allows us to smoothly extend u to $\mathbb{R}^n$ via $u(x + \sum_{m=1}^n k_m L_m e_m) = u(x)$ for $x \in \Omega$, $k_m \in \mathbb{Z}$, and unit coordinate vector e_m.

v) Sometimes one is interested in solving a PDE on $\Omega = \mathbb{R}^n$. Then, it is possible to impose conditions of the form
$$\lim_{|x| \to \infty} u(x) \to u_0(x),$$
for some given $u_0(x)$, which is called a *far field condition*. Often one simply asks for solutions that *vanish at infinity*, i.e., $u_0 \equiv 0$. For example, the Stokes' paradox shows that it can depend on the dimension n whether such a condition leads to a well-posed problem.

Definition I.10 (Hyperbolic of first-order systems)**:** Let $\Omega \subset \mathbb{R}$ and $Q_T = (0, T) \times \Omega$. A system of (nonlinear) first-order PDEs
$$\partial_t \mathbf{u} + \mathbf{B}(\mathbf{u}) \partial_x \mathbf{u} = 0 \tag{I.18}$$
for $\mathbf{u} : Q_T \to \mathbb{R}^k$ is called *hyperbolic*, if $\mathbf{B}(\mathbf{u})$ has real eigenvalues. The transport equation $B(u) = a \in \mathbb{R}$ and the (inviscid) Burgers equation $B(u) = 2u$ are two important examples in the scalar case, i.e., for $k = 1$.

I.5. Well-Posedness and Classical Solution Concept

In this section we define what a well-posed problem is and we give some examples for illustration. First of all we need to agree on what we call *a solution*. Even though it would be nice if solutions would be smooth in the sense $u \in C^\infty(\Omega)$, but such a requirement is usually way too restrictive. With the mathematical formulation and tools we have defined already, it is reasonable to ask for k times continuously differentiable solutions $u \in C^k(\Omega)$ for a k-th order partial differential equation. For the second-order problems this leads to the following definition.

Definition I.11 (Classical solution)**:** Any twice continuously differentiable function $u \in C^2(\Omega)$ satisfying the second-order PDE defined in I.4 (possibly with additional initial/boundary conditions on $\partial\Omega$) is called a *classical solution*.

We will see that even this requirement is sometimes too strong, but we will work with classical solutions for the moment. Now we need to see what really constitutes a proper PDE formulation, which allows us to give meaning to the solution concept. Here we follow the concept of well-posedness put forward by Jacques Hadamard.

Definition I.12 (Hadamard well-posedness)**:** A PDE with initial/boundary conditions is called *well-posed* if the following conditions are satisfied:

a) *Existence of solutions:* There exists at least one solution.

b) *Uniqueness of solutions:* There is at most one solution.

c) *Stability:* The solution behavior changes continuously with the *data*.

A PDE which is not well-posed is called *ill-posed*.

Let explain and discuss this concept: The term data here consitutes initial/boundary conditions and other input parameters, e.g., the right-hand-side $f(x)$ in I.4. Continuous dependence on the data means that small changes in the data (in an appropriate norm) produce small changes in the solution (in an appropriate norm). If no solutions exist, then we might have set too many conditions or have a too restrictive solution concept. If infinitely many solutions exist, then we might have set too few conditions or have a too relaxed solution concept. If the solution behavior does not depend continously on the data, then small approximation errors (in particular in the numerical approximation) potentially lead to large errors in our prediction. In the following we give examples for ill-posed PDEs.

Example I.13 (Wrong sign)**:** Consider the PDE $\partial_x^2 u + u = 0$ and let $u(0) = 0$.

i) Let $\Omega = (0, \pi/2)$. With $u(\pi/2) = 1$ the unique solution is $u(x) = \sin(x)$.

ii) Let $\Omega = (0, \pi)$. With $u(\pi) = 1$ there are no solutions. On the other hand, if we require $u(\pi) = 0$ then infinitely many solutions $u(x) = a\sin(x)$ with $a \in \mathbb{R}$ exist.

iii) If on the other hand we consider $-\partial_x^2 u + u = 0$ then we get $u(x) = c_1 \sinh(x)$ where c_1 is determined by the second boundary condition (well-posed).

Example I.14 (Forward/backward heat equation)**:** Consider the PDE $\partial_t u + k\Delta u = 0$ on $Q_T = (0, T) \times \Omega$ with initial conditions at $t = 0$ and boundary conditions on $(0, T) \times \partial\Omega$.

i) For $k < 0$ the equation is well-posed (heat equation).

ii) For $k > 0$ the equation is ill-posed.
For $\Omega = (0, 1)$ and $u(t, 0) = u(t, 1) = 0$ consider initial data $u(0, x) = n^{-1}\sin(n\pi x)$, which is an eigenfunction of the operator $-k\partial_x^2$ with eigenvalue $\lambda_n = k\pi^2 n^2 > 0$.

The solution of the problem is given by $u(t,x) = \exp(\lambda_n t)u(0,x)$. While for $n \to \infty$ the initial data is arbitrarily close to zero, the solution becomes arbitrarily large at any finite time t.

Example I.15 (Nonsmooth solution)**:** Consider the Burgers equation $\partial_t u + \partial_x\left(u^2\right) = 0$ on $Q_T = (0,T) \times (0,2\pi)$ with periodic boundary conditions $u(t,0) = u(t,2\pi)$ and $T = 1$. The initial data are $u(0,x) = 1 + \cos(x)$. The solution shown in Figure I.4 is smooth and follows the characteristic $u(t,x(t)) = u_0(x_0)$ where $x(t) = x_0 + 2u_0(x_0)t$ until it becomes multivalued and the numerical solution develops a jump discontinuity which is determined by the Rankine–Hugoniot condition. The numerical solutions is computed via the MATLAB code in Listing I.1.

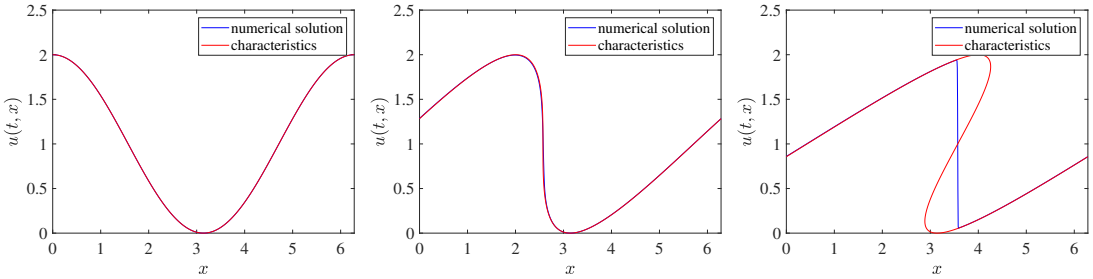


Figure I.4.: Solution of the Burgers equation at **(left)** $t = 0$ **(middle)** $t = 0.5$ and **(right)** $t = 1$.

Listing I.1: MATLAB code for Burgers equation

```
% solves du/dt + d(f(u))/dx = 0 for Burgers equation
% where f(u)=u^2 via up-wind discretization and with
% periodic boundary conditions.

L  = 2*pi;      % spatial domain (0,L)
T  = 1.0;       % time domain (0,T)
Nx = 1024;      % spatial resolution
Nt = 2000;      % temporal resolution

x = linspace(0,2*pi,Nx); % grid
u = 1+cos(x) ;           % initial data

% time-step size
dt = T/Nt;
dx = L/(Nx-1);

% solve inviscid Burgers equation via up-wind
for i=1:2001
    % plot
    if (mod(i,100)==1)
        plot(x,u,'b-');
        hold on
    end
    % compute flux f=u^2 and up-wind derivative dfdx
    f    = u.^2;
    dfdx = (f(1:end)-f([end 1:end-1]))/dx;
    u    = u - dt*dfdx;
```

```
end
hold off
```

Listing I.2: Python code for Burgers equation

```
# run: "python burgers.py"
# solves du/dt + d(f(u))/dx = 0 for Burgers equation
# where f(u)=u^2 via up-wind discretization and with
# periodic boundary conditions.
import numpy as np
import matplotlib.pyplot as plt

L = 2*(np.pi)  # spatial domain (0,L)
T = 1.0        # time domain (0,T)
Nx = 1024      # spatial resolution
Nt = 2000      # temporal resolution

x = np.linspace(0, L, Nx+1)  # grid
u = 1+np.cos(x)              # initial data

dt = T/Nt  # time step
dx = L/Nx  # grid size

# solve inviscid Burgers equation via up-wind
for i in range(Nt):
    # compute flux f=u^2 and up-wind derivative dfdx
    f = u*u
    dfdx = (f-f[np.r_[Nx, 0:Nx]])/dx
    u = u - dt*dfdx

    # create plot every 100 steps
    if (i % 100 == 1):
        plt.plot(x, u)

# add labels to plot
plt.xlabel('x')
plt.ylabel('solution u')
plt.show()
```

Example I.16 (Missing regularity)**:** Consider the problem $-\Delta u = 0$ in $\Omega = (0,1)^2$ with $u = g = x^2$ on $\partial\Omega$. Note that $\partial_x^2 g = 2$ and $\partial_y^2 g = 0$ and $u = g$ on $\partial\Omega$. Hence, $-\Delta u = 2 \neq 0$ on $\partial\Omega$ but $-\Delta u = 0$ in Ω. This shows that $u \notin C^2(\bar{\Omega})$.

Example I.17 (Wrong boundary conditions)**:** Consider the Laplace equation

$$-(u_{xx} + u_{yy}) = 0 \quad \text{in } \Omega = \mathbb{R} \times (0,T)$$

as instationary initial value problem with the "time" y. For some $n > 0$ let the two initial conditions

$$\begin{aligned} u(x,0) &= \frac{\sin(nx)}{n}, \\ u_y(x,0) &= 0 \end{aligned}$$

be given. Then

$$u(x,y) = \frac{\cosh(ny)\sin(nx)}{n} = \frac{1}{n}\frac{\mathrm{e}^{ny} + \mathrm{e}^{-ny}}{2}\sin(nx)$$

is a solution. We see this since

$$u_{xx}(x,y) = (\cosh(ny)\cos(nx))_x = -n\cosh(ny)\sin(nx),$$
$$u_{yy}(x,y) = \frac{\sin(nx)}{2}\left(\mathrm{e}^{ny} - \mathrm{e}^{-ny}\right)_y = n\cosh(ny)\sin(nx),$$

so it holds $u_{xx} + u_{yy} = 0$. The initial conditions are also satisfied.

The solution grows as e^{ny} for growing n. Therefore, there are arbitrarily small initial values $u(x,0) = \frac{\sin(nx)}{n}$ such that there are arbitrarily large solutions for some fixed $y \in (0,T)$. On the other hand, the PDEs

$$-(v_{xx} + v_{yy}) = 0 \quad \text{in } \Omega = \mathbb{R} \times (0,T)$$

with the initial conditions

$$v(x,0) = 0,$$
$$v_y(x,0) = 0$$

has a solution $v(x,y) = 0$. Therefore, for arbitrarily small changes in the initial condition, the differences of the solutions u and v can be arbitrarily large. Thus the solution is not stable against perturbations of the initial values and the problem is *ill-posed*.

We observed that unclassified problems can be well-posed (or not), but showing this might depend on the specific problem at hand. Furthermore, also parabolic, elliptic,hyperbolic problems can be ill-posed, if the boundary/initial conditions are imposed incorrectly. This shows

- Boundary/initial conditions have a significant impact on the solution and whether or not a PDE is well-posed.
- The classification helps setting up well-posed PDE problems.
- Different types (classes) of PDEs require different numerical solution methods.

That is why the proper characterization of second-order linear PDEs is so important. We have the following rule of thumb for well-posed second-order PDEs:

- elliptic: PDE + boundary conditions,
- parabolic: PDE + boundary conditions + one initial condition,
- hyperbolic: PDE + boundary conditions + two initial conditions.

The boundary conditions can be, for example, of Dirichlet, of Neumann, or of Robin type. We finally comment on the difference between parabolic and hyperbolic PDEs.

Remark I.18: Consider the wave equation

$$u_{tt} - u_{xx} = 0 \quad \text{in } \Omega = (0,1) \times (0,T),$$

which is hyperbolic according to Example I.8 c). It can be shown that all its solutions attain the form

$$u(x,t) = \varphi(x+t) + \psi(x-t), \tag{I.19}$$

where φ and ψ are two arbitrary twice continuously differentiable functions.

Proper initial conditions are of the form $u(x,0) = g_1(x)$, $u_t(x,0) = g_2(x)$ for all $x \in (0,1)$. With (I.19) it follows that

$$\begin{aligned} g_1(x) &= u(x,0) = \varphi(x) + \psi(x), \\ g_2(x) &= u_t(x,0) = \varphi'(x+t) + \psi'(x-t)\big|_{t=0} = \varphi'(x) - \psi'(x). \end{aligned} \tag{I.20}$$

The first equation gives $g_1'(x) = \varphi'(x) + \psi'(x)$. After some rearrangements in the second equation we further get

$$\begin{aligned} \varphi'(x) &= g_2(x) + \psi'(x) = g_2(x) + g_1'(x) - \varphi'(x), \\ \psi'(x) &= \varphi'(x) - g_2(x) = g_1'(x) - \psi'(x) - g_2(x). \end{aligned}$$

Therefore, we obtain

$$\varphi'(x) = \frac{1}{2}\left(g_2(x) + g_1'(x)\right), \quad \psi'(x) = \frac{1}{2}\left(g_1'(x) - g_2(x)\right).$$

Therefore, φ and ψ are completely specified in the interval $(0,1)$ up to the two integration constants. In fact, by (I.20) the integration constants have to cancel each other. To obtain the whole solution u in Ω, the functions φ and ψ also have to be specified in the intervals $[1, 1+T)$ and $(-T, 0]$, respectively. This can be achieved by imposing additional boundary conditions such as $u(0,t) = u_0(t)$ and $u(1,t) = u_1(t)$ for all $t \in (0,T)$ and some functions u_0, $u_1 : (0,T) \to \mathbb{R}$.

Now consider the heat equation

$$u_t - u_{xx} = 0 \quad \text{in } \Omega = (0,1) \times (0,T),$$

which is parabolic according to Example I.8 b). Here only one initial condition

$$u(x,0) = g_1(x) \quad \text{for } x \in (0,1)$$

may be prescribed, since we already have

$$u_t(x,0) = u_{xx}(x,0) = g_1''(x),$$

where the last equality follows from the initial condition and the assumption that g_1 is twice continuously differentiable. This means that $u_t(x,0)$ is already prescribed by the choice of g_1.

I.6. Nondimensionalization of PDEs

In general, both the solution $u(x)$ and the variables x in a PDE have certain physical dimensions, which are certain powers of the base units for length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity as shown in Table I.6, i.e., the units of each physical quantity X can be written in the form $[X] = L^{n_1} M^{n_2} T^{n_3} I^{n_4} \Theta^{n_5} N^{n_6} J^{n_7}$ for a unique choice of numbers $(n_1, ..., n_7)$. Additionally, the statement of the partial differential equation comes with certain parameters, which also have certain physical dimensions. A quantity Y without units for which $[Y] = 1$ is called *nondimensional*.

name	unit	unit name	symbol
length	m	meter	L
mass	kg	kilogram	M
time	s	second	T
electric current	A	ampere	I
thermodynamic temperature	K	kelvin	Θ
amount of substance	mol	mole	N
luminous intensity	cd	candela	J

Table I.1.: Base units in the SI system.

As an example, consider the following convection-diffusion equation

$$\partial_t u + \nabla \cdot (u\mathbf{v}) = \nabla \cdot \big(D \nabla u\big), \qquad \text{in } Q_T = (0, T_\infty) \times \Omega, \tag{I.21a}$$

$$u = g, \qquad \text{in } (0, T_\infty) \times \partial\Omega, \tag{I.21b}$$

$$u(t = 0, \cdot) = u_0, \qquad \text{in } \Omega, \tag{I.21c}$$

with boundary conditions $g : \partial\Omega \to \mathbb{R}$ and inital conditions $u_0 : \Omega \to \mathbb{R}$. In this equation $u : Q_T \to \mathbb{R}$ represents the density of particles at $(t, x) \in Q_T$. The physical dimension of the density is $[u] = N \cdot L^{-n}$, i.e., amount of substance per volume. The time has units $[t] = T$ and space has units $[x] = L$. The physical dimensions of the diffusion constant D and the convection velocity $\mathbf{v}$ are $[D] = L^2 \cdot T^{-1}$ and $[\mathbf{v}] = L \cdot T^{-1}$. While the classification of PDEs for $D > 0$ determines this to be a parabolic equation, often also the relative magnitude of terms is important for the qualitative behavior of solutions. Often this knowledge is essential for the choice of the numerical discretization.

Nondimensionalization Method 1: A straightforward method for the nondimensionalization of a PDE such as (I.21) is to express all dimensional quantities (solution, variables, parameters) in multiples of their physical dimension expressed in the SI units shown in Table I.6. For example, we define

$$u(t, x) = N \cdot L^{-n} \bar{u}(\bar{t}, \bar{x}), \qquad \mathbf{v}(t, x) = L \cdot T^{-1} \bar{\mathbf{v}}(\bar{t}, \bar{x}), \qquad t = T\bar{t}, \qquad x = L\bar{x}, \tag{I.22}$$

where all $\bar{u}, \bar{\mathbf{v}}, \bar{x}, \bar{t}$ carry no physical dimension anymore and $T = 1\text{s}$, $N = 1\text{mol}$, $L = 1\text{m}$. Inserting these definitions into (I.21a) we obtain

$$\partial_{\bar{t}}\bar{u} + \bar{\nabla} \cdot (\bar{u}\bar{\mathbf{v}}) = DTL^{-2}\bar{\nabla}^2\bar{u}, \tag{I.23}$$

where the differential operators $\partial_{\bar{t}}, \bar{\nabla}$ act on the dependence of $\bar{u}, \mathbf{v}$ on $\bar{t}, \bar{x}$. Equivalently we reformulate the boundary and initial conditions. The nondimensional quantity $\bar{D} = DTL^{-2}$ expresses the diffusion constant, with the choice of T, N, L above expressed in SI units. However, in no way are the parameters adjusted to the problem, i.e., L does not relate to the size of the domain or the size of typical features and L/T does not relate to the magnitude velocity $\mathbf{v}$. Hence, also the magnitude of $\bar{D}$ does not carry any viable information about the importance of diffusion in comparison with convection.

Nondimensionalization Method 2: Now we propose a problem-adjusted nondimensionalization of (I.21). Therefore, we choose again

$$u(t,x) = N \cdot L^{-n}\bar{u}(\bar{t},\bar{x}), \qquad \mathbf{v}(t,x) = L \cdot T^{-1}\bar{\mathbf{v}}(\bar{t},\bar{x}), \qquad t = T\bar{t}, \qquad x = L\bar{x}, \tag{I.24}$$

but now set L as a typical size in the problem, e.g. the domain size $L = \max_{x,y\in\Omega} \|x-y\|$. Furthermore we assume that the velocity $\mathbf{v}$ has typical values V, so that we can define a characteristic time scale as $T = L/V$. Then we obtain

$$\partial_{\bar{t}}\bar{u} + \bar{\nabla} \cdot (\bar{u}\bar{\mathbf{v}}) = \bar{D}\bar{\nabla}^2\bar{u}, \tag{I.25}$$

where

$$\bar{D} = \frac{D}{LV} \equiv \text{Pe}^{-1}, \tag{I.26}$$

where Pe is the so-called Péclet number, which characterizes the ratio of advective transport and diffusive transport. In many engineering applications the Péclet number can be quite large, which potentially leads to so-called singularly perturbed problems which require careful numerical treatment. The concept of corresponding boundary layers was first introduced by Ludwig Prandtl in the context of fluid flows.

I.7. Solution Strategies, Exact Solutions, Solution Operators

Even though in the lecture we will see many explicit expressions or representations of solutions to a PDE, in real-life situations this is hopeless or very unlikely. However, in the following we mention some general strategies to solve or simplify PDEs, some explicit exact solutions or solution operators.

I.7.1. Solution strategies

Numerical methods

Numerical methods, as treated in this lecture, often rely on transforming nonlinear problem into a sequence of linear problems (Newton or fixed-point iterations) and on transforming infinite-dimensional problems into finite-dimensional problems.

Besides finite difference and finite element methods, there are many more discretization methods for partial differential equations. Most noteworthy perhaps are:

- finite volume methods: express conservation laws on surfaces (**robust**),
- spectral methods: higher-order methods based on Fourier transform (**precise**),
- boundary element method: efficient solution of simple elliptic problems (**fast**),
- method of lines/symplectic integrators/gradient flows: time integration,
- variational inequalities: nonsmooth problems,
- particle-based methods/discrete element methods: fancy.

Finite differences and finite elements should be the first methods to learn because they are the most versatile and applicable in almost every situation. While finite volume methods are particularly strong for certain problems with large Péclet numbers (convection dominated), the resulting problems can also be addressed by these two methods.

Special solutions

While special solutions do not help to solve a general PDE problem, they can convey some information about the general behavior of solutions. Two types of special solutions that deserve a special mentioning are *self-similar solutions* and *traveling-wave solutions*.

Definition I.19 (Self-similar solution)**:** A solution $u : (0,\infty) \times \mathbb{R}^n \to \mathbb{R}$ to an instationary PDE problem of the form

$$u(t,x) = t^\alpha U(\eta), \qquad \eta = xt^\beta, \tag{I.27}$$

is called a *self-similar* solution. As has been pointed out by Zel'dovich and studied by Barenblatt, depending on how α, β are determined, self-similar solutions can be of *first* or *second kind*.

Definition I.20 (Traveling-wave solution)**:** A solution $u : (0,\infty) \times \mathbb{R} \to \mathbb{R}$ to an instationary PDE problem of the form

$$u(t,x) = U(\eta), \qquad \eta = x - ct, \tag{I.28}$$

is called a *traveling-wave solution* moving with speed c.

Such a behavior is interesting since often one might also be able to show convergence to a self-similar/traveling-wave solutions for general initial data, which then determines the long-time behavior of solutions. In particular, such solutions are also found for nonlinear equations, which otherwise might not admit other simple (closed form) solutions.

Integral transformations

Integral transformation such as Fourier or Laplace transform can significantly simplify a PDE. For example: Let $u : (0, 2\pi) \to \mathbb{R}$ and consider its Fourier series

$$u(x) = \sum_k a_k \exp(ikx). \tag{I.29}$$

Now compute the second derivative with respect to x

$$\partial_x u(x) = \sum a_k (ik) \exp(ikx),$$
$$\partial_x^2 u(x) = \sum a_k (-k)^2 \exp(ikx),$$

which shows that differentiation becomes a multiplication with (ik) in Fourier space, which can be readily inverted. However, this is restricted to simple boundary conditions and appropriate domains Ω. However, so-called spectral methods use this property in conjunction with fast Fourier transform to compute fast and accurate solutions to certain classes of PDEs.

Separation of variables

When seeking a function of several variables, e.g., $u(t, x)$, then the ansatz $u(t, x) = f(t)g(x)$ is called separation of variables. This can sometimes significantly reduce the effort of solving the corresponding PDE. For example: Consider the heat equation

$$\partial_t u - k u'' = 0, \qquad \text{in } Q_T = (0, T) \times (0, L),$$

with initial data $u(0, x) = u_0(x)$ at $t = 0$ and boundary conditions $u(t, 0) = u(t, L) = 0$. Inserting the separation of variables ansatz into the equation gives

$$\frac{f'(t)}{k f(t)} = \frac{g''(x)}{g(x)}.$$

Since the left side only depends on t and the right on x, both must be equal and equal to, say, $-\lambda$. Therefore

$$f'(t) = -\lambda k f(t), \qquad g''(x) = -\lambda g(x).$$

Using the boundary conditions one can show $\lambda > 0$ and hence $f(t) = f_0 \exp(-\lambda k t)$ and $g(x) = g_1 \sin(\sqrt{\lambda} x)$. Boundary conditions require $\sqrt{\lambda} = n\pi/L$ for any integer $n \in \mathbb{N}$. A general solution is then of the form

$$u(t, x) = \sum_{n=1}^{\infty} a_n \sin\big(\tfrac{\pi n}{L} x\big) \exp\big(-\frac{n^2 \pi^2 k t}{L^2}\big),$$

where the $a_n \in \mathbb{R}$ are given by the Fourier expansion (sine expansion) of the initial data.

Change of variables

When a given PDE problem has a certain symmetry, it can be useful to use this symmetry and transform the variables to respect this symmetry. For example: Consider the domain $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ and Helmholtz's equation

$$-\Delta u = \lambda u, \qquad \text{in } \Omega,$$

with $u = 0$ on $\partial\Omega$. Using the Laplace operator in polar coordinates from (I.16) and a separation ansatz $u(r, \phi) = \sum_{n\geq 0} a(r\sqrt{\lambda})\cos(n\phi) + b_n(r\sqrt{\lambda})\sin(n\phi)$ produces Bessel's differential equation for $a(x), b(x)$, i.e.,

$$x^2 a'' + xa' + (x^2 - n^2)a = 0,$$

and solutions are given by Bessel functions of the first kind $J_n(x)$.

I.7.2. Method of characteristics

We consider the (nonlinear) first-order PDE, in particular the strictly hyperbolic equation in conservation form, where we seek $\mathbf{u} : (0, t) \times \mathbb{R} \to \mathbb{R}^k$ such that

$$\partial_t \mathbf{u} = -\partial_x \mathbf{f}(\mathbf{u}) = \mathbf{B}(\mathbf{u})\partial_x \mathbf{u}, \tag{I.30}$$

for given $\mathbf{f} : \mathbb{R}^k \to \mathbb{R}^k$ and initial data $\mathbf{u}(t = 0, \cdot) = \mathbf{u}_0$. We aim at solving (I.30) by converting it into an ODE. Therefore, consider a curve $\xi : (0, T) \to \mathbb{R}^k$ and consider $z(t) = \mathbf{u}\big(t, \xi(t)\big)$ and $p = \partial_x \mathbf{u}(t, \xi(t))$. Then we have

$$z'(t) = \partial_t \mathbf{u}(t, x) + \xi'(t) \cdot \nabla \mathbf{u}.$$

For given initial data $\xi(0) = x_0$ and $z(0) = \mathbf{u}_0(x_0)$ the solution of the ODEs

$$\xi'(t) = \mathbf{B}(z(t)), \qquad z'(t) = 0,$$

gives $z(t) = \mathbf{u}_0(x_0)$. When $\xi(t)(x_0)$ is invertible, a solution of the PDE is given by

$$\mathbf{u}(t, \xi(t)) = \mathbf{u}_0(x_0), \qquad \xi(t) = \mathbf{B}(\mathbf{u}_0(x_0))t + x_0. \tag{I.31}$$

Example I.21 (Burgers equation)**:** Consider the inviscid Burgers equation, where $k = 1$ and $f(u) = u^2$ and $B(u) = 2u$. Hence we get the solution

$$u(t, \xi(t)) = u_0(x_0), \qquad \xi(t) = 2u_0(x_0)t + x_0, \tag{I.32}$$

as shown in Figure I.4 for $u_0(x_0) = 1 + \cos(x_0)$.

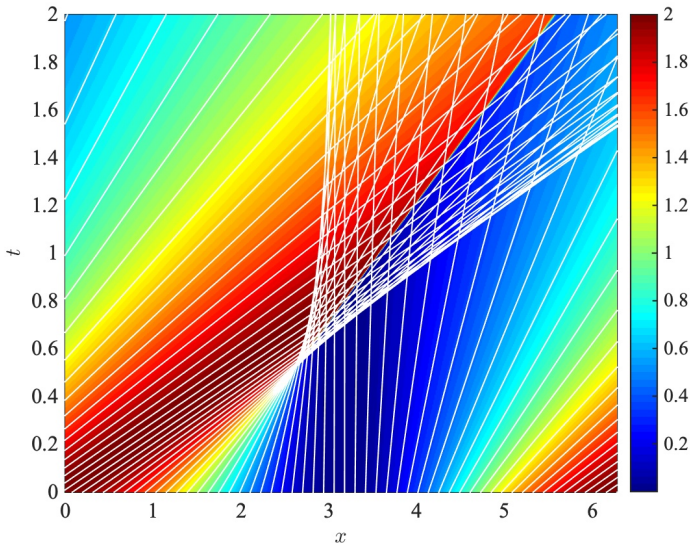


Figure I.5.: Numerical solution of Burgers equation with shading showing $u(t,x)$ at (t,x) and white lines are the characteristic curves $(t,\xi(t))$ in a periodic domain.

I.7.3. Exact solutions

In the following we present some further expressions for exact solutions of the PDEs introduced before.

Example I.22 (Homogeneous transport equation)**:** The method of characteristics applied to the transport equation $\partial_t u + a\partial_x u = 0$ with initial data $u(t=0,\cdot) = u_0$ gives the general solution

$$u(t,x) = u_0(x - at), \tag{I.33}$$

for $(t,x) \in \mathbb{R} \times \mathbb{R}$.

Example I.23 (Inhomogeneous transport equation)**:** The method of characteristics applied to the transport equation $\partial_t u + a\partial_x u = f$ with initial data $u(t=0,\cdot) = u_0$ gives the general solution

$$u(t,x) = u_0(x - at) + \int_0^t f(\tau, x + a(\tau - t))\,\mathrm{d}\tau \tag{I.34}$$

for $f : Q_T \to \mathbb{R}$ and $(t,x) \in Q_T = \mathbb{R} \times \mathbb{R}$.

Definition I.24 (Fundamental solution of Laplace's equation)**:** Let $\Omega = \mathbb{R}^n$. The singular function $G : \Omega \to \mathbb{R}$ of the form

$$G(x) = \begin{cases} \frac{1}{2}|x| & n = 1 \\ -\frac{1}{2\pi}\log|x| & n = 2 \\ \frac{1}{n(n-2)\alpha(n)}\frac{1}{|x|^{n-2}} & n \geq 3 \end{cases} \tag{I.35}$$

is the *fundamental solution of Laplace's equation*. One can verify that $-\nabla^2 G = \delta_0$, where δ_0 denotes the Dirac-δ distribution on Ω.

Example I.25 (Poisson equation)**:** When we define $u : \mathbb{R}^n \to \mathbb{R}$ by the convolution

$$u(x) = \int_{\mathbb{R}^n} G(x-y) f(y) \mathrm{d}y, \tag{I.36}$$

then u solves the Poisson equation $-\nabla^2 u = f$ for sufficiently smooth (and fast decaying) $f : \mathbb{R}^n \to \mathbb{R}$. For $n \geq 3$, any bounded solution is (up to an additive constant) of this form. This constant can be fixed by a far field condition $u(x) \to 0$ as $|x| \to \infty$.

Example I.26 (Homogeneous heat equation)**:** With $u(t,x) = t^{-\alpha} w(t^{-\beta}|x|)$ for $\alpha = n/2$ and $\beta = 1/2$ one can verify that $w(s) = \exp\left(-\frac{1}{4}s^2\right)$ is a self-similar solution of the heat equation. This motivates to define

$$G(t,x) = \begin{cases} \frac{1}{(4\pi t)^{n/2}} \exp(-\frac{|x|^2}{4t}) & t > 0 \\ 0 & t < 0 \end{cases} \tag{I.37}$$

as the fundamental solution of the heat equation. One can verify that $(\partial_t - \Delta) G = \delta_0$, where δ_0 denotes the Dirac-δ distribution on $Q_T = (0,T) \times \Omega$. Therefore

$$u(t,x) = \int_\Omega G(t, x-y) u_0(y) \mathrm{d}y, \tag{I.38}$$

is a solution of the homogeneous heat equation with inital data u_0.

Definition I.27 (Duhamel's principle)**:** Consider an instationary, inhomogeneous, linear PDE of the form

$$\partial_t u - Lu = f, \qquad \text{in } Q_T = (0,T) \times \Omega,$$

with homogeneous Dirichlet boundary conditions $u = 0$ on $\Gamma = \partial\Omega$ and homogeneous initial data $u(0,x) = 0$. Then Duhamel's principles formally gives the solution

$$u(t,x) = \int_0^t (P^s f)(t,x) \, \mathrm{d}s, \tag{I.39}$$

where $u_s = P^s f$ is the solution operator on $Q_{s,T} = (s,T) \times \Omega$ for the problem

$$\partial_t u_s - L u_s = 0, \qquad \text{in } Q_{s,T},$$

with homogeneous Dirichlet boundary conditions $u_s = 0$ on $\Gamma = \partial\Omega$ and inhomogeneous initial data $u_s(t = s, x) = f(s,x)$.

Example I.28 (Inhomogeneous heat equation)**:** Using the solution of the homogeneous heat equation and Duhamel's principle, the inhomogeneous heat equation is solved by

$$u(t,x) = \int_0^t \int_{\mathbb{R}^n} G(t-\tau, x-y) f(\tau, y) \mathrm{d}y \, \mathrm{d}\tau.$$

Example I.29 (Homogeneous wave equation)**:** For given $Q_T = (0,T) \times \mathbb{R}$ consider the homogeneous wave equation

$$u_{tt} = u_{xx} \qquad \text{on } (0,T) \times \mathbb{R},$$

with initial conditions $u(0,x) = f(x)$ and $u_t(0,x) = g(x)$ for $x \in \mathbb{R}$. Using change of variables $\xi = x + t$ and $\eta = x - t$ and $u(t,x) = \bar{u}(\xi,\eta)$ we obtain $u_{tt} - u_{xx} = \bar{u}_{\xi\eta} = 0$, which we can solve in general using $\bar{u}(\xi,\eta) = \phi(\xi) + \psi(\eta)$. Using the initial conditions we get $f = \phi + \psi$ and $g = \phi' - \psi'$. Thus,

$$\phi(\xi) = \tfrac{1}{2} f(\xi) + \tfrac{1}{2} \int_{x_0}^{\xi} g(r)\mathrm{d}r, \quad \psi(\eta) = \tfrac{1}{2} f(\eta) - \tfrac{1}{2} \int_{x_0}^{\eta} g(r)\mathrm{d}r,$$

for arbitrary x_0, which results in d'Alembert's formula

$$u(t,x) = \tfrac{1}{2}(f(x+t) + f(x-t)) + \tfrac{1}{2} \int_{x-t}^{x+t} g(r)\mathrm{d}r. \tag{I.40}$$

In principle this discussion could be extended towards aspects of:

- **uniqueness**: maximum principles & variational techniques are used,
- **other boundary conditions**: give slightly or considerably different expressions,
- **more complex domains**: possible for boundary element methods.

For a somewhat longer discussion we refer to the textbook by Evans [Eva98].

I.8. Summary and Concluding Remarks

What we learned in the chapter is that in practice, some real world/life problems can be formulated mathematically in terms of PDEs. The process of transforming the problem into a PDE is sometimes refered to as *PDE modeling*. Often, the modeling employs certain physical assumptions and conservation laws. However, as not every PDE statement makes sense, also the mathematical problem statement must be well-posed in order to be ready for numerical discretization.

Example I.30 (Exemplary real world problem)**:** Consider the problem: How much energy is lost, if you leave a window in your house open? A sketch showing the geometry of the house is given in Figure I.6

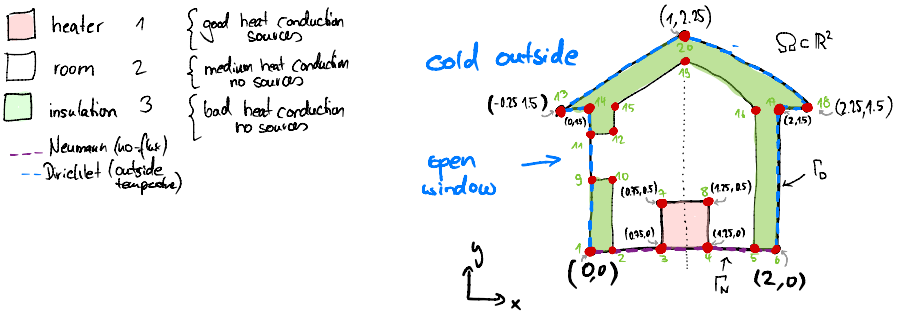


Figure I.6.: Geometric model of (diffusive) heat transport in house.

For the sake of simplicity we make the following assumptions:

- The house is two-dimensional and the problem stationary, i.e., $\Omega \subset \mathbb{R}^2$.
- Heat transport is diffusive (more likely conductive due to large Péclet number).
- The walls are either cold (Dirichlet boundary conditions) or insulating (homogeneous Neumann boundary conditions).
- The heater provides a certain constant output power density, so that we reach a comfortable temperature of $T_{\text{inside}} = 19°C$ inside, while having $T_{\text{outside}} = 0°C$.

Thereby, we expect to solve the following PDE problem

$$\begin{aligned} \nabla \cdot \mathbf{q}(x) &= f(x), && \text{balance of heat flux } \mathbf{q} \text{ and production } f, \\ \mathbf{q}(x) &= -k(x)\nabla T(x), && \text{Fouriers law with heat conductivity } k, \\ T(x) &= T_{\text{outside}}, && \text{on outside walls } \Gamma_{\text{D}}, \\ n \cdot \nabla T(x) &= 0, && \text{on insulating walls } \Gamma_{\text{N}}, \end{aligned}$$

where $f(x) = f_i$ and $k(x) = k_i$ for $x \in \Omega_i$ and $i \in \{\text{room}, \text{heater}, \text{insulator}\}$. In this example we assume all quantitites are nondimensional and

$$k_{\text{insulator}} = 0.1, \qquad k_{\text{room}} = 1, \qquad k_{\text{heater}} = 2, \qquad f_{\text{heater}} = 180,$$

and $f_{\text{room}} = f_{\text{insulator}} = 0$. In Figure I.7 the construction of the geometry Ω, the computational mesh, and the disjoint subdomains $\Omega = \cup_i \Omega_i$ is shown.

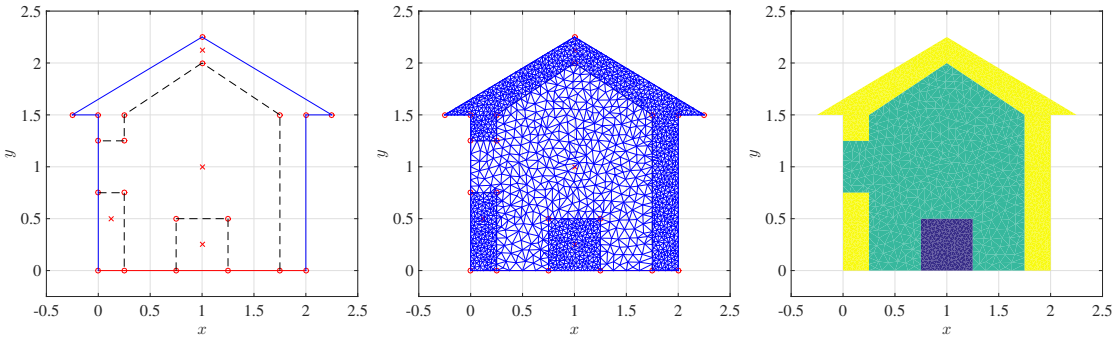


Figure I.7.: **(left)** CAD geometry description from 20 vertices (red dots) with line segments for interior interfaces (black dashed), Dirichlet boundary (blue full) and insulating boundary (red full) **(middle)** computational mesh (triangulation **(right)** distinct subdomains Ω_i encoded with different colors.

Below, in Figure I.8 we show the numerical solution of the heat flow problem (computed using P_2 FEM in Matlab with 6,356 unknowns).

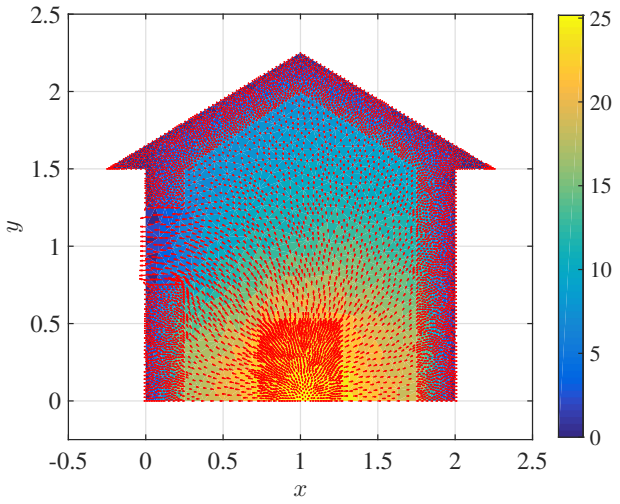


Figure I.8.: Numerical solution of the stationary heat flow problem (linear, elliptic PDE) showing temperature $T : \Omega \to \mathbb{R}$ using shading/color and the heat flux $\mathbf{q} : \Omega \to \mathbb{R}^2$ using red arrows.

From the heat flow we can compute the energy loss per time, i.e., power P, as

$$P = -\int_{\partial\Omega} \mathbf{q} \cdot n \, \mathrm{d}A = -\int_{\Omega} \nabla \cdot \mathbf{q} \, \mathrm{d}x = |\Omega_{\text{heater}}| f_{\text{heater}},$$

which one would have to redimensionalize again. Of course, without a closed window the energy loss is smaller, since one can maintain a comfortable average room temperature of $|\Omega_{\text{room}}|^{-1} \int_{\Omega_{\text{room}}} T(x) \, \mathrm{d}x = T_{\text{inside}} = 19°C$ at a lower heating power f_{heater}.

Listing I.3: MATLAB: vertices (x, y) and connectivity `xy_poly` of house geometry. The three columns in `xy_poly` indicate the starting and ending vertex of a connection and its id $= \{-1, 0, 1\}$ for edges of type $\{\text{Internal}, \text{Neumann}, \text{Dirichlet}\}$.

```
x=[0.00  0.25  0.75  1.25  1.75  2.00  ...
   0.75  1.25  0.00  0.25  0.00  0.25  ...
  -0.25  0.00  0.25  1.75  2.00  2.25  ...
   1.00  1.00];

y=[0.00  0.00  0.00  0.00  0.00  0.00  ...
   0.50  0.50  0.75  0.75  1.25  1.25  ...
   1.50  1.50  1.50  1.50  1.50  1.50  ...
   2.00  2.25];

xy_poly=[ 1   2   0; 2   3   0; 3   4   0; 4   5   0; 5   6   0; 3   7  -1; 7   8  -1;...
          8   4  -1; 5  16  -1;16  19  -1;19  15  -1;15  12  -1;12  11  -1; 9  10  -1;...
         10   2  -1; 6  17   1;17  18   1;18  20   1;20  13   1;13  14   1;14  11   1;...
         11   9   1; 9   1   1];
```

II. Finite Difference Methods

II.1. Introduction

The finite difference method (FDM) is a widely used approach to solve ODEs and PDEs on a computer. In this chapter we will first develop the underlying methodology for elliptic two-point boundary value problems. Then we generalize to problems on higher dimensional domains and discuss the treatment of more general boundary conditions. Finally, we also discuss an application of the finite difference method to the numerical approximation of hyperbolic problems.

In order to explain the basic concept of the finite difference method we first consider a simple domain $\overline{\Omega} = [0, L_1] \times \ldots \times [0, L_n] \subset \mathbb{R}^n$ which depending on n would be called interval $n = 1$, square/rectangle $n = 2$, cube/cuboid $n = 3$, tesseract $n = 4$, or for general n a hypercube/hyperrectangle. For simplicity we also assume that all length are the same and equal to $L_i = 1$, so that scalar functions are defined $u : [0, 1]^n \to \mathbb{R}$.

For example let $\overline{\Omega} = [0, 1]^2$. Instead of evaluting u at every point $x \in [0, 1]^2$ we seek approximations $u_{i,j}$ of the function $u(x_{i,j})$ evaluated at finitely many points $x_{i,j} = (ih, jh) \in \mathbb{R}^2$ for $i, j = 0, \ldots, N + 1$ and $h = 1/(N + 1)$ denotes the mesh size of the grid for given $N \in \mathbb{N}$. The corresponding 36 grid points for $N = 4$ are show in Fig. II.1

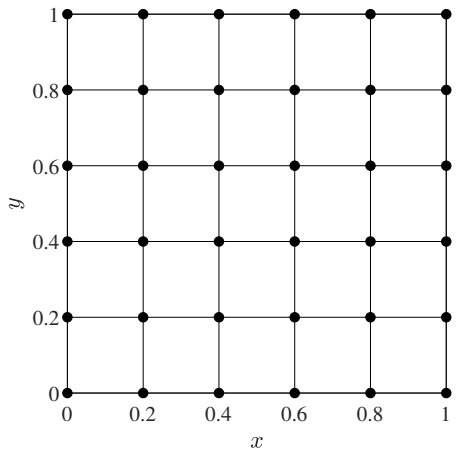


Figure II.1.: Example discretization mesh

The basic idea of the method will be to replace differential operators by a finite difference approximation, which will lead to a large sparse system of linear equations for the vector $u_{i,j}$. In this section we will also introduce the basic framework to show, under

rather strong assumptions, the convergence of solutions to the exact solution of the PDE. We start this endeavor by considering a one-dimensional elliptic boundary value problem (BVP).

II.2. One-Dimensional Elliptic BVP

In the previous section we introduced the linear, second-order PDE problem

$$Lu = f, \qquad \text{in } \Omega \subset \mathbb{R}^n, \tag{II.1}$$
$$u = g, \qquad \text{on } \partial\Omega, \tag{II.2}$$

with Dirichlet condition specified on the boundary $\partial\Omega$. For simplicitly we will focus on problems with constant coefficients and on simple domains, i.e., hyperrectangles of the form $\overline{\Omega} = [0, L_1] \times [0, L_n] \subset \mathbb{R}^n$ or even with $L_i = 1$ for $i = 1 \dots n$. To introduce the main concepts we start in one spatial dimension $n = 1$.

Consider the following two-point boundary value problem (BVP): Set $\Omega = (0, 1)$ and $\overline{\Omega} = \Omega \cup \partial\Omega = [0, 1]$. Find a function $u \colon \overline{\Omega} \to \mathbb{R}$ with

$$\begin{cases} -a(x)u''(x) + b(x)u'(x) + c(x)u(x) = f(x) & \text{in } \Omega = (0,1), \\ u(0) = \alpha, \\ u(1) = \beta, \end{cases} \tag{II.3}$$

where $\alpha, \beta \in \mathbb{R}$, $a, b, c, f : \Omega \to \mathbb{R}$ with $a > 0$ and $c \geq 0$ are given.

Remark II.1: Strictly speaking, this problem depends on only one variable and therefore should be considered an ODE. However, since we impose boundary conditions at $x = 0$ and $x = 1$ this problem can not be simply integrated. Still, this type of problem is often solved by ODE-type methods using so-called *shooting methods*. Nevertheless, this problem is very well-suited to introduce the main ideas of finite differences.

The general *idea* of the finite difference method is to replace all derivatives in the BVP by suitable approximation using difference quotients, for instance

$$u'(x) \approx \frac{u(x+h) - u(x)}{h} =: D^+u(x)$$

for $x \in \overline{\Omega}$ and a sufficiently small $h > 0$ such that $x + h \in \overline{\Omega}$ as well. Examples of difference quotients are

$$D^+u(x) := \frac{u(x+h) - u(x)}{h} \qquad \text{(“forward difference quotient”)},$$
$$D^-u(x) := \frac{u(x) - u(x-h)}{h} \qquad \text{(“backward difference quotient”)},$$
$$D^0u(x) := \frac{u(x+h) - u(x-h)}{2h} \qquad \text{(“central difference quotient”)}.$$

The small positive parameter $h > 0$ is called the *step size*, *grid or mesh size* of the difference quotient. Again let us stress that the difference quotients are only well-defined if $x \pm h \in \overline{\Omega}$. Note that the difference operators are linear in the sense that we have, for example,

$$D^+[\lambda u + v](x) = \lambda D^+ u(x) + D^+ v(x) \quad \text{(similarly for } D^-, D^0)$$

for all continuous functions u, $v \colon \overline{\Omega} \to \mathbb{R}$. For the approximation of the second derivative we apply two of the operators. The most classical one is

$$\begin{aligned} D^+D^-u(x) &= D^+\left(\frac{u(x) - u(x-h)}{h}\right) \\ &= \frac{u(x+h) - u(x)}{h^2} - \frac{u(x) - u(x-h)}{h^2} \\ &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}. \end{aligned}$$

This converges with order 2 to the second derivative of u provided that $u \in C^4(\overline{\Omega})$ as we will see further below. Moreover, it holds true that

a) $D^0 u(x) = \frac{1}{2}(D^+ + D^-)u(x)$,

b) $D^+D^-u(x) = D^-D^+u(x)$,

For the error analysis we recall the following result.

Theorem II.2 (Taylor's formula)**:** Let $I \subset \mathbb{R}$ be an open interval and $u \in C^{r+1}(\overline{I})$, that is u is $(r+1)$-times continuously differentiable on I and continuous on the closed interval $\overline{I}$. Then, for every x, $y \in I$ it holds

$$u(y) = \sum_{k=0}^{r} \frac{u^{(k)}(x)}{k!}(y-x)^k + R$$

with the Lagrange form of the remainder $R = \frac{u^{(r+1)}(\xi)}{(r+1)!}(y-x)^{r+1}$ and $\xi \in [\min(x,y), \max(x,y)]$. In general if $u : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ we can also use the previous multiindex notation to show

$$u(x+h) = \sum_{\alpha, |\alpha| \le r} \frac{D^\alpha u(x)}{\alpha!} h^\alpha + R$$

with a similar remainder term R.

Now we can use difference formulas to replace the operators in (II.3). But which finite difference should we use? For example, if $y = x + h \in I$, then we get

$$u(x+h) = u(x) + u'(x)h + \frac{1}{2}u''(x)h^2 + \cdots + \frac{u^{(k)}(x)}{k!}h^k + R$$

with $R = \frac{u^{(k+1)}(\xi)}{(k+1)!}h^{k+1}$ for some $\xi \in [x, x+h]$. Hence, from this we obtain with $k = 1$

$$\underbrace{D^+u(x) = \frac{u(x+h) - u(x)}{h}}_{\text{difference operator}} \overset{\text{Taylor } (n=1)}{=} u'(x) + \underbrace{\frac{1}{2}u''(\xi)h}_{\text{"error"}}.$$

Theorem II.3: Let $I \subseteq \mathbb{R}$ be an open interval and let $[x-h, x+h] \subseteq \overline{I}$. Then it holds:

a) If $u \in C^2(\overline{I})$, then we have

$$D^+u(x) = u'(x) + hR_1$$

with $|R_1| \leq \frac{1}{2}\max_{\xi\in[x,x+h]}|u''(\xi)|$ and

$$D^-u(x) = u'(x) + hR_2$$

with $|R_2| \leq \frac{1}{2}\max_{\xi\in[x-h,x]}|u''(\xi)|$.

b) If $u \in C^3(\overline{I})$, then we have

$$D^0u(x) = u'(x) + h^2R_3$$

with $|R_3| \leq \frac{1}{6}\max_{\xi\in[x-h,x+h]}|u'''(\xi)|$.

c) If $u \in C^4(\overline{I})$, then we have

$$D^-D^+u(x) = u''(x) + h^2R_4$$

with $|R_4| \leq \frac{1}{12}\max_{\xi\in[x-h,x+h]}|u^{(4)}(\xi)|$.

Remark II.4: Sometimes it makes sense to iterate discrete difference operators to obtain higher order derivatives, e.g., D^+D^- or $(D^+D^-)^2$, but sometimes the results might not be as useful, e.g., $(D^0)^2$. The safest method is usually to approximate a derivative using Taylor expansions using some construction principle (order of approximation, compactness of stencil), as we will see later. For example consider $(D^0)^2$:

$$D^0u(x) = \frac{u(x+h) - u(x-h)}{2h}$$
$$\big(D^0(D^0u)\big)(x) = \frac{u(x+2h) - 2u(x) + u(x-2h)}{4h^2}$$

which is still a second-order approximation of $u''(x)$, but is now defined on a smaller domain and will cause problems with implementation near the boundary.

Discretization strategy for the BVP

Now we discuss the discretization strategy for the BVP (II.3): For simplicity, we assume constant coefficients and as before $\Omega = (0,1) \in \mathbb{R}$.

Step 1: We replace $\Omega = (0,1)$ and $\overline{\Omega} = [0,1]$ by uniform meshes/grids, i.e., finite sets of points in $\overline{\Omega}$ with distance/step size/grid size $h = \frac{1}{N+1}, N \in \mathbb{N}$, covering the domain and obtain

$$\Omega_h = \{h, 2h, \cdots, Nh\}, \quad \overline{\Omega}_h = \{0, h, 2h, \cdots, Nh, \underbrace{(N+1)h}_{=1}\}, \quad \Gamma_h = \{0,1\},$$

as shown examplarily in Fig. II.2 for $N = 7$. The subscript h highlights the discrete approximation with grid size h. E.g. in MATLAB this is realized via the command `L=1;xh = linspace(0,L,N+2);h=L/(N+1);` for a domain $\Omega = (0,L)$ with $L = 1$. We enumerate points in $\overline{\Omega}_h$ using $x_n = nh$ for $0 \leq N \in \mathbb{N}_0 \leq (N+1)$. Note that in MATLAB enumeration of vector indices starts at $n = 1$, i.e., $x_0 =$`xh(1)`$, \ldots, x_{N+1} =$`xh(N+2)`.

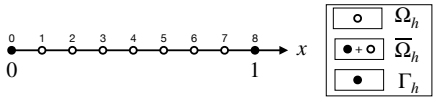


Figure II.2.: Example discretization 1D mesh

Step 2: We replace $u : \overline{\Omega} \to \mathbb{R}$ by a grid function $u_h : \overline{\Omega}_h \to \mathbb{R}$. Such a function can be interpreted/represented as a vector $u_h = (u_h(x_0), \ldots, u_h(x_{N+1}))^\top \in \mathbb{R}^{N+2}$. Now we approximate the derivatives of u by difference quotients in terms of u_h: For x_n this gives

$$\begin{aligned} u'(x_n) &\approx \frac{u_h(x_n + h) - u_h(x_n - h)}{2h} = \frac{u_h(nh + h) - u_h(nh - h)}{2h} \\ &= \frac{u_h(x_{n+1}) - u_h(x_{n-1})}{2h} \\ &= D^0 u_h(x_n). \end{aligned}$$

Then, instead of the continuous problem of finding a function $u\colon \overline{\Omega} \to \mathbb{R}$ with

$$\begin{cases} -au''(x) + bu'(x) + cu(x) = f(x) \quad \text{in } \Omega, \\ u(0) = \alpha, \\ u(1) = \beta, \end{cases} \tag{II.P}$$

we solve the following discrete problem (let $h = \frac{1}{N+1}, N \in \mathbb{N}$) of finding a grid function $u_h : \overline{\Omega}_h \to \mathbb{R}$ such that

$$\begin{cases} -aD^+D^- u_h(x) + bD^0 u_h(x) + cu_h(x) = f(x) \quad \text{for } x \in \Omega_h, \\ u_h(0) = \alpha, \\ u_h(1) = \beta. \end{cases} \tag{II.DP}$$

In fact, (II.DP) is a linear system of equations. Then $x_0 = 0$ and $x_{N+1} = 1$. For the unknowns $u_h(x_0), \dots, u_h(x_{N+1})$ we have to solve

$$-a\frac{u_h(x_{n+1}) - 2u_h(x_n) + u_h(x_{n-1})}{h^2} + b\frac{u_h(x_{n+1}) - u_h(x_{n-1})}{2h} + cu_h(x_n) = f(x_n),$$

for each $n \in \{1, \dots, N\}$. Further, the boundary conditions yield the conditions $u_h(x_0) = \alpha$ and $u_h(x_{N+1}) = \beta$. Since the discretization only uses relations between $x_h = nh$ and its neighbors at $(n \pm 1)h$, this results in a tridiagonal, sparse system of linear equations.

Step 3: Next, we write this system of linear equations in terms of a matrix-vector product as (for simplicity let $b = c = 0$)

$$\begin{aligned} &-\frac{a}{h^2}\begin{bmatrix} -\frac{h^2}{a}u_h(x_0) \\ u_h(x_0) - 2u_h(x_1) + u_h(x_2) \\ \vdots \\ \vdots \\ u_h(x_{N-1}) - 2u_h(x_N) + u_h(x_{N+1}) \\ -\frac{h^2}{a}u_h(x_{N+1}) \end{bmatrix} \\ &= -\frac{a}{h^2}\begin{bmatrix} -\frac{h^2}{a} & 0 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & \ddots & \ddots & & \\ & & \ddots & \ddots & -1 & \\ & & & 1 & -2 & 1 \\ & & & & 0 & -\frac{h^2}{a} \end{bmatrix}\begin{bmatrix} u_h(x_0) \\ u_h(x_1) \\ \vdots \\ \vdots \\ u_h(x_N) \\ u_h(x_{N+1}) \end{bmatrix} \overset{!}{=} \begin{bmatrix} \alpha \\ f(x_1) \\ \vdots \\ \vdots \\ f(x_N) \\ \beta \end{bmatrix}. \end{aligned}$$

Note that this system consists of $N + 2$ equations with $N + 2$ unknowns. However, the equations for the boundary conditions can be easily eliminated, so that we have a system of N equations with N unknowns $u_h(x_1), \dots, u_h(x_N)$. For this we insert $u_h(x_0) = \alpha$ in the second equation and $u_h(x_{N+1}) = \beta$ in the second last equation and move them to the right hand side.

Then the *reduced linear system* (with eliminated boundary conditions) reads

$$-\frac{a}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}\begin{bmatrix} u_h(x_1) \\ u_h(x_2) \\ \vdots \\ \vdots \\ u_h(x_N) \end{bmatrix} \overset{!}{=} \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ \vdots \\ f(x_N) \end{bmatrix} + \begin{bmatrix} \alpha\frac{a}{h^2} \\ 0 \\ \vdots \\ 0 \\ \beta\frac{a}{h^2} \end{bmatrix}.$$

In the case that $b \neq 0$ or $c \neq 0$ we obtain the corresponding reduced matrix-vector system

in the same way as

$$\left(-\frac{a}{h^2} \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} + \frac{b}{2h} \begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 0 \end{bmatrix} + c\mathbb{I}_N \right) \begin{bmatrix} u_h(x_1) \\ u_h(x_2) \\ \vdots \\ \vdots \\ u_h(x_N) \end{bmatrix}$$

$$= \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ \vdots \\ f(x_N) \end{bmatrix} + \begin{bmatrix} \alpha\left(\frac{a}{h^2} + \frac{b}{2h}\right) \\ 0 \\ \vdots \\ 0 \\ \beta\left(\frac{a}{h^2} - \frac{b}{2h}\right) \end{bmatrix}. \quad \text{(II.4)}$$

A simple numerical solution of the Poisson problem with homogeneous Dirichlet boundary conditions solved with this finite difference method is shown in Fig. II.3.

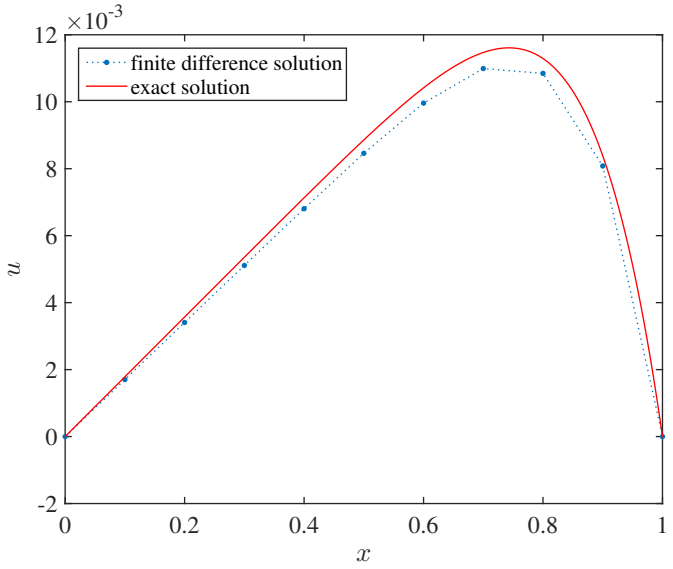


Figure II.3.: Comparison of the numerical solution of $L_h u_h = f$ on $\overline{\Omega}_h$ generated from $\overline{\Omega} = [0, 1]$ on a coarse grid with $h = 1/(N + 1) = 1/10$ and parameters $a = 1, \alpha = \beta = b = c = 0$ with $f(x) = x^6$ with exact solution $u(x) = 1/56(x - x^8)$. Dots indicate the position of the grid point.

Definition II.5 (Compact notation for difference stencils)**:** For short, we often write $Lu = f$ with $L = -a\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b\frac{\mathrm{d}}{\mathrm{d}x} + c$ for the BVP. In the same way we write $L_h u_h = f_h$ where $u_h = \big[u_h(x_1), \ldots, u_h(x_N)\big]^\mathsf{T} \in \mathbb{R}^N$ and $f_h \in \mathbb{R}^N$ denotes the vector on the right hand side of (II.4) that includes the inhomogeneity f and the boundary values α, $\beta \in \mathbb{R}$. The matrix $L_h \in \mathbb{R}^{N\times N}$ is given by

$$L_h = -\frac{a}{h^2} \underbrace{(1,\,-2,\,1)}_{\text{difference stencil}} + \frac{b}{2h}(-1,\,0,\,1) + c(0,\,1,\,0) \in \mathbb{R}^{N\times N},$$

where we write $(d_1,\, d_2,\, d_3)$ for the tridiagonal matrix

$$\begin{bmatrix} d_2 & d_3 & & & \\ d_1 & d_2 & d_3 & & \\ & \ddots & \ddots & \ddots & \\ & & d_1 & d_2 & d_3 \\ & & & d_1 & d_2 \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

which is a Toeplitz matrix for constant grid size.

To sum up, the exact solution $u\colon \overline{\Omega} \to \mathbb{R}$ solves $Lu = f$ with

$$L = -a\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b\frac{\mathrm{d}}{\mathrm{d}x} + c$$

with additional boundary conditions on $\partial\Omega$.

For the FDM we first choose a step size $h = \frac{1}{N+1}$ with arbitrary $N \in \mathbb{N}$ and obtain an equidistant grid $\Omega_h = \{h,\, 2h,\, \ldots,\, Nh\} = \{x_1,\, \ldots,\, x_N\}$. The extended grid $\overline{\Omega}_h$ also includes the boundary of Ω, that is $\overline{\Omega}_h = \{0,\, h,\, \ldots,\, Nh,\, (N+1)h\} = \{0,\, 1\} \cup \Omega_h$. Then we determine a grid function $u_h\colon \overline{\Omega}_h \to \mathbb{R}$ that solves $L_h u_h = f_h$, where we usually interpret $u_h = \big[u_h(x_1),\, \ldots,\, u_h(x_N)\big]^\mathsf{T} \in \mathbb{R}^N$ as a vector. The matrix $L_h \in \mathbb{R}^{N\times N}$ is given by

$$L_h = -\frac{a}{h^2}(1,\,-2,\,1) + \frac{b}{2h}(-1,\,0,\,1) + c(0,\,1,\,0) \in \mathbb{R}^{N\times N}$$

and the inhomogeneity

$$f_h = \big[f(h),\, \ldots,\, f(Nh)\big]^\mathsf{T} + \big[\alpha\left(\tfrac{a}{h^2} + \tfrac{b}{2h}\right),\, 0,\, \ldots,\, 0,\, \beta\left(\tfrac{a}{h^2} - \tfrac{b}{2h}\right)\big]^\mathsf{T} \in \mathbb{R}^N.$$

Note that at this point we did not impose any conditions that ensure the (unique) solvability of the discrete problem $L_h u_h = f_h$. We will address this question later.

II.3. Difference Stencils

In general we will not derive operators for derivatives by iterating D^+, D^-, D^0 but develop finite difference formulas based on the question at hand. The general tool to derive these approximations will always be Taylor's theorem. A formalization of finite difference formulas can be performed introducing difference stencils.

II.3.1. General Difference Stencils on Uniform Meshes

In the general n-dimensional case with uniform mesh we can write stencils for a r-th order differential operator $\Delta_{h,r}$ and in particular for the Laplacian $\Delta_h := \Delta_{h,2}$ as

$$\Delta_{h,r} u(x_h) = \frac{1}{h^r} \sum_{\alpha} s_\alpha u_h(x + \alpha h) \tag{II.5}$$

using integer stencil indices $\alpha \in \mathbb{Z}^n$ (similar to multiindices), where we define the shift $x+\alpha h = (x_1+\alpha_1 h, \ldots, x_n+\alpha_n h)^\top \in \overline{\Omega}_h$. With $k(s_\alpha) = \sum_{\alpha, s_\alpha \neq 0} 1$ we denote the number of points contributing to the stencil and say $\Delta_{h,r}$ is a k-point stencil. A difference stencil is called *compact*, if $s_\alpha = 0$ for $\max_i |\alpha_i| > 1$. In particular in one and two spatial dimensions compact difference stencils are represented using

$$\Delta_{h,r} = \frac{1}{h^r} \begin{pmatrix} s_{-1} & s_0 & s_{+1} \end{pmatrix}, \qquad \Delta_{h,r} = \frac{1}{h^r} \begin{pmatrix} s_{-1,-1} & s_{0,-1} & s_{+1,-1} \\ s_{-1,0} & s_{0,0} & s_{+1,0} \\ s_{-1,+1} & s_{0,+1} & s_{+1,+1} \end{pmatrix}, \tag{II.6}$$

respectively. We draw the compact 9-point stencil in 2D as shown in Fig. II.4.

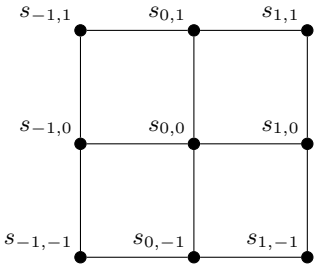


Figure II.4.: General compact 9-point stencil in 2D

Two simple examples are the compact 3-point stencil for the Laplacian or the central difference for the first derivative in 1D

$$\Delta_h = \frac{1}{h^2} \begin{pmatrix} 1 & -2 & 1 \end{pmatrix}, \qquad D^0 = \Delta_{h,1} = \frac{1}{2h} \begin{pmatrix} -1 & 0 & 1 \end{pmatrix}. \tag{II.7}$$

II.3.2. Advanced Stencils in One Dimension

Now we consider the construction principle behind advanced stencils in one spatial dimension. In particular we are interested in the approximation of the following expressions

$$u^{(r)}(x), \qquad (a(x)u'(x))' \quad \text{and} \quad a(x)u''(x),$$

and the extension to non-uniform spatial meshes. We assume that $a(x)$ is a given smooth function $a : \Omega \to \mathbb{R}$ that we can evaluate at arbitrary points $x \in \Omega$. We will be focussed on symmetric stencils.

Approximation of $u^{(r)}$

Assume that $u_h : \Omega_h \to \mathbb{R}$, a point $x_n \in \overline{\Omega}_h$, and $m \in \mathbb{N}_0$ are given so the m neighbors of x_n to either side $x_{n-m}, x_{n-m+1}, \ldots, x_{n+m-1}, x_{n+m}$ are in $\overline{\Omega}_h$ as well. For each x_{n+j} with $j \in \mathbb{Z}$ and $-m \leq j \leq m$ we can make a Taylor expansion around x_n of the form

$$u(x_{n+j}) = u(x_n) + u'(x_n)jh + u''(x_n)\frac{(jh)^2}{2!} + \ldots + u^{(k-1)}(x_n)\frac{(jh)^{k-1}}{(k-1)!} + R_{j,k} \tag{II.8}$$

so that with $k = 2m+1$ we have $2m+1$ equations for the $2m+1$ unknowns $h^r u^{(r)}(x_n)$ for $0 \leq r \leq 2m$ built from the coefficients in (II.8). Constructing the corresponding matrix $S_k \in \mathbb{R}^{k\times k}$ results in the linear system of equations

$$\begin{pmatrix} u_h(x_{n-m}) \\ u_h(x_{n-m+1}) \\ \vdots \\ u_h(x_{n+m-1}) \\ u_h(x_{n+m}) \end{pmatrix} = S_k \begin{pmatrix} u(x_n) \\ hu'(x_n) \\ h^2u''(x_n) \\ \vdots \\ h^{k-1}u^{(k-1)}(x_n) \end{pmatrix}, \tag{II.9}$$

where S_k^{-1} provides the desired symmetric stencils s_α for derivatives $\Delta_{h,r}$ in one spatial dimension. Examples for k-point stencils approximating different derivatives r using different number of neighbors m are shown in Tab. II.1. The remainders from Taylor's theorem are $R_{j,k} = \frac{(jh)^k}{(k)!}u^{(k)}(\xi)$ for some $\xi \in [x_n - mh, x_n + mh]$.

r	m	s_{-3}	s_{-2}	s_{-1}	s_0	s_{+1}	s_{+2}	s_{+3}
1	1			$-1/2$	0	$1/2$		
1	2		$1/12$	$-2/3$	0	$2/3$	$-1/12$	
1	3	$-1/60$	$3/20$	$-3/4$	0	$3/4$	$-3/20$	$1/60$
2	1			1	-2	1		
2	2		$-1/12$	$4/3$	$-5/2$	$4/3$	$-1/12$	
2	3	$1/90$	$-3/20$	$3/2$	$-49/18$	$3/2$	$-3/20$	$1/90$
3	2		$-1/2$	1	0	-1	$1/2$	
3	3	$1/8$	-1	$13/8$	0	$-13/8$	1	$-1/8$
4	2		1	-4	6	-4	1	
4	3	$-1/6$	2	$-13/2$	$28/3$	$-13/2$	2	$-1/6$

Table II.1.: Difference stencils for derivatives of different order

Example II.6 (Stencils with $m = 0$ and $m = 1$)**:**

$$S_0 = (1), \quad S_3 = \begin{pmatrix} 1 & -1 & 1/2 \\ 1 & 0 & 0 \\ 1 & 1 & 1/2 \end{pmatrix} \Rightarrow S_3^{-1} = \begin{pmatrix} 0 & 1 & 0 \\ -1/2 & 0 & 1/2 \\ 1 & -2 & 1 \end{pmatrix} = \begin{pmatrix} (0,1,0) \\ hD^0 \\ h^2D^+D^- \end{pmatrix}$$

II.3.3. Approximation of $(\widehat{a}(x)u'(x))'$ and $a(x)u''(x)$

First of all, clearly for elliptic problems $Lu = f$ we can, in principle, always represent an operator $(\widehat{a}(x)u'(x))'$ by using the product rule and defining $a(x) = \widehat{a}(x)$ and $b(x) = \widehat{a}'(x)$ in the standard form. The reason for directly discretizing the operator $(a(x)u'(x))'$ is that it often appears in the 1D equivalent of time-dependent problems of the form

$$\partial_t u(t,x) - \nabla \cdot (a(x)\nabla u(t,x)) = 0, \tag{II.10}$$

which with homogeneous Neumann boundary condition and using Gauss's theorem implies a conservation of $\int_\Omega u(t,x)\,\mathrm{d}x$ over time. Having such a property on the discrete level is often tied to the spatial discretization of the operator $(a(x)u'(x))'$. Therefore, first of all we can define the flux

$$\mathbf{q}_{n+1/2} = a(x_{n+1/2})\frac{u_h(x_{n+1}) - u_h(x_n)}{x_{n+1} - x_n}$$

where $x_{n+1/2} = \frac{1}{2}(x_n + x_{n+1})$ is the point, where we evaluate $a(x)$ at. With Taylor expansion we can check that $\mathbf{q}_{n+1/2} = a(x)u'(x) + \mathcal{O}(h^2)$ for $x = x_{n+1/2}$. Similarly we can check that

$$\begin{aligned}(a(x_n)u'(x_n))' &\approx \frac{\mathbf{q}_{n+1/2} - \mathbf{q}_{n-1/2}}{x_{n+1/2} - x_{n-1/2}} \\ &= \tfrac{2}{x_{n+1}-x_{n-1}}\left[a(x_{n+1/2})\tfrac{u_h(x_{n+1})-u_h(x_n)}{x_{n+1}-x_n} - a(x_{n-1/2})\tfrac{u_h(x_n)-u_h(x_{n-1})}{x_n-x_{n-1}}\right].\end{aligned} \tag{II.11}$$

is a second-order approximation for sufficiently smooth a, u and for uniform meshes. Please observe that using $\delta x_n = (x_{n+1/2} - x_{n-1/2})$ the numerical integration gives

$$\sum_{n=1}^{N}(a'(x_n)u'(x_n))'(\delta x_n) = \mathbf{q}_{N+1/2} - \mathbf{q}_{1/2} = 0, \tag{II.12}$$

because $\mathbf{q}_{N+1/2} = \mathbf{q}_{1/2} = 0$ for homogeneous Neumann boundary conditions. This property would imply the discrete conservation law for the integral

$$\sum_{n=1}^{N} u_h(t,x_n)(\delta x_n), \tag{II.13}$$

even though we did not yet consider time-discretizations of parabolic PDEs.

Finally, the difference quotient for $a(x)u''(x)$ is, for uniform meshes, simply given by

$$a(x_n)u''(x_n) = a(x_n)\frac{u_h(x_{n+1}) - 2u_h(x_n) + u_h(x_{n+1})}{h^2}, \tag{II.14}$$

where higher-order difference stencils even for non-uniform meshes can be derived as sketched before using Taylor's theorem.

II.3.4. Alternative Difference Stencils for Laplace Operator

The standard 7-point stencil in 3D is defined by $s_{0,0,0} = -6$ and $s_{\pm 1,0,0} = s_{0,\pm 1,0} = s_{0,0,\pm 1} = 1$ as it is shown in panel Fig. II.5(**a**). Another compact stencil, i.e., a stencil with $s_\alpha = 0$ for $|\alpha_i| > 1$, in 2D is shown in Fig. II.5(**b**) where $s_{0,0,0} = -8/3$ and $s_{i,j,k} = 1/3$ otherwise for $-1 \le i, j, k \le 1$. And in Fig. II.5(**c**) we show another example of a non-compact 9-point stencil in 2D. The standard stencil for the Laplacian in arbitrary dimension n is given by $s_\alpha = -2n$ for $\alpha = 0$ and $s_\alpha = 1$ for any α with $\sum_i |\alpha_i| = 1$.

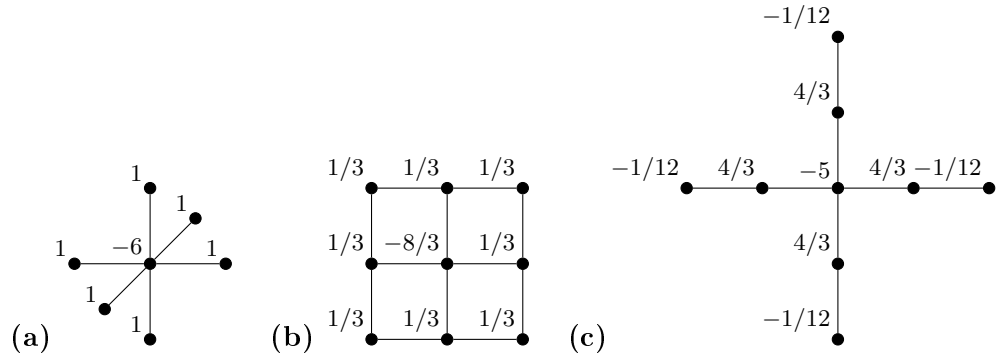


Figure II.5.: (**a**) standard 7-point difference stencil in 3D (**b**) alternative compact 9-point stencil in 2D (**c**) alternative noncompact 9-point stencil in 2D with consistency order 4, assuming $u \in C^6(\overline{\Omega})$.

One can show that in two dimensions there is no compact 9-point stencil of consistency order 3. However, under special assumptions one can modify the stencil and right-side to obtain a higher convergence order, e.g.,

$$\frac{-1}{6h^2}\begin{pmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{pmatrix} u_h = \frac{1}{6}\begin{pmatrix} & 1/2 & \\ 1/2 & 4 & 1/2 \\ & 1/2 & \end{pmatrix} f, \tag{II.15}$$

is of consistency order 4, if we assume $u \in C^6(\overline{\Omega})$. However, except for domains with periodic boundary conditions, such operators are rarely used since they can not be used near the boundary.

II.4. Convergence of the Elliptic BVP

We first investigate the question how close the approximation u_h is to the exact solution u? Of course, it does not make sense to compare a finite dimensional vector u_h with a function u, since both are quite different mathematical objects.

Therefore, when applying the finite difference method we always introduce an error, which is called the *discretization error*. In order to be able to compare u and u_h and estimate the discretization we restrict the exact solution to the grid Ω_h and then compare

the restriction with u_h in a suitable norm. For this, let $R_h\colon C(\overline{\Omega}) \to \mathbb{R}^N, h = \frac{1}{N+1}$, be the *restriction operator* on the grid Ω_h defined by

$$R_h w := \big[w(h), \ldots, w(Nh)\big]^{\mathsf{T}} \in \mathbb{R}^N.$$

Note that R_h takes a function as an argument and returns a vector containing the function values at the grid points.

In order to evalute the discretization error, let us briefly recall what a norm is.

Definition II.7 (norm on a vector space)**:** A *norm* on an $\mathbb{R}$-($\mathbb{C}$-) vector space V is a mapping $\|\cdot\|\colon V \to \mathbb{R}$ such that

a) $\|v\| \geq 0$ for all $v \in V$ and $\|v\| = 0 \Leftrightarrow v = 0$;

b) $\|\alpha v\| = |\alpha|\|v\|$ for all $\alpha \in \mathbb{R}$ (or $\mathbb{C}$) and $v \in V$;

c) $\|v + w\| \leq \|v\| + \|w\|$ for all v, $w \in V$ (triangle inequality).

Every norm on a vector space induces a norm on the corresponding space of all linear mappings/operators mapping into this vector space. The following theorem explains this for square matrices.

Theorem II.8: Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Then the mapping $|\!|\!|\cdot|\!|\!| : \mathbb{R}^{d\times d} \to \mathbb{R}$ given by

$$|\!|\!|A|\!|\!| = \sup_{v\in\mathbb{R}^d,\, \|v\|=1} \|Av\| = \sup_{v\in\mathbb{R}^d,\, v\neq 0} \frac{\|Av\|}{\|v\|}$$

for $A \in \mathbb{R}^{d\times d}$ is a norm on $\mathbb{R}^{d\times d}$. This norm is called the *matrix norm induced by* $\|\cdot\|$. For all A, $B \in \mathbb{R}^{d\times d}$ and $v \in \mathbb{R}^d$ it holds

a) $|\!|\!|A \cdot B|\!|\!| \leq |\!|\!|A|\!|\!| \cdot |\!|\!|B|\!|\!|$,

b) $\|Av\| \leq |\!|\!|A|\!|\!| \cdot \|v\|$.

Now, for every $h = \frac{1}{N+1}$, $N \in \mathbb{N}$, let $\|\cdot\|_h$ be a norm on $\mathbb{R}^N$. The family $(\|\cdot\|_h)_{h>0}$ is called a *system of norms*. Examples are the *maximum norm on* $\mathbb{R}^N$ given by

$$\|v\|_{\infty,h} := \max_{1\leq i\leq N} |v_i|, \quad \text{where} \quad v = \big[v_1, \ldots, v_N\big]^{\mathsf{T}} \in \mathbb{R}^N;$$

or the *discrete L_2-norm* on $\mathbb{R}^N$

$$\|v\|_{2,h} = \left(h \sum_{i=1}^{N} |v_i|^2\right)^{\frac{1}{2}}.$$

Definition II.9: Let $(\|\cdot\|_h)_{h>0}$ be a system of norms. A finite difference method $L_h u_h = f_h$ approximating $Lu = f$ with exact solution u is called

a) *consistent* w. r. t. $(\|\cdot\|_h)_{h>0}$, if $\|f_h - L_h R_h u\|_h \to 0$ as $h \to 0$,

b) *consistent of order* $p > 0$ w. r. t. $(\|\cdot\|_h)_{h>0}$, if $\|f_h - L_h R_h u\|_h \in \mathcal{O}(h^p)$ as $h \to 0$,

c) *convergent* w. r. t. $(\|\cdot\|_h)_{h>0}$, if $\|u_h - R_h u\|_h \to 0$ as $h \to 0$,

d) *convergent of order* $p > 0$ w. r. t. $(\|\cdot\|_h)_{h>0}$, if $\|u_h - R_h u\|_h \in \mathcal{O}(h^p)$ as $h \to 0$.

In the above definition we made use of the Landau $\mathcal{O}$-notation. Let us recall this. Assume that $f, g : \mathbb{R} \to \mathbb{R}$ are two functions. In general, we write

$$f(x) \in \mathcal{O}(g(x)) \text{ for } x \to a,$$

if and only if there exist $\delta > 0$ and $M > 0$ such that

$$|f(x)| \leq M \cdot |g(x)| \quad \forall\, x \in [a - \delta, a + \delta].$$

More specifically, in the situation of Definition II.9 d) this means that there exist $h_0 > 0$ and $M > 0$ such that

$$\|u_h - R_h u\|_h \leq M h^p \quad \text{for all } 0 \leq h \leq h_0.$$

Remark II.10: The notions of consistency and convergence are closely related (see Theorem II.12) but take slightly different view points:

a) Consistency: How well does the exact solution u solve the discrete problem $L_h u_h = f_h$ as $h \to 0$?

b) Convergence: How close are the discrete solution u_h and the exact solution u at the grid points as $h \to 0$?

Note that for consistency it is not necessary to know whether the discrete problem $L_h u_h = f_h$ actually has a (unique) solution.

Definition II.11: A finite difference method $L_h u_h = f_h$ is called *stable* with respect to a given system of norms $(\|\cdot\|_h)_{h>0}$, if $L_h \in \mathbb{R}^{N \times N}$ is invertible and if there exists $C > 0$ with

$$\left\| L_h^{-1} \right\|_h \leq C < \infty$$

for all $h = \frac{1}{N+1}$ with sufficiently large $N \in \mathbb{N}$. In particular, the constant C is independent of the step size h.

Theorem II.12 ("Consistency + Stability = Convergence")**:** Let the discretization of the elliptic BVP $L_h u_h = f_h$ be consistent and stable with respect to a given system of norms $(\|\cdot\|_h)_{h>0}$. Then the corresponding FDM is convergent with respect to $(\|\cdot\|_h)_{h>0}$. Furthermore, if the FDM is consistent of order $p > 0$ with respect to $(\|\cdot\|_h)_{h>0}$, then it is convergent of order $p > 0$ with respect to the same system of norms.

Proof. Since L_h is invertible, we have

$$u_h = L_h^{-1} f_h.$$

Inserting this into the discretization error yields

$$\begin{aligned} \|u_h - R_h u\|_h &= \|L_h^{-1} f_h - R_h u\|_h \\ &= \|L_h^{-1} f_h - L_h^{-1} L_h R_h u\|_h \\ &\leq \underbrace{|||L_h^{-1}|||_h}_{\leq C} \|f_h - L_h R_h u\|_h. \end{aligned}$$

Now, if the FDM is consistent then

$$\|f_h - L_h R_h u\|_h \to 0 \quad \text{for } h \to 0.$$

Thus, the FDM is also convergent. Moreover, if the FDM is consistent of order p, then the same computation shows

$$\|u_h - R_h u\|_h \leq C \|f_h - L_h R_h u\|_h \in \mathcal{O}(h^p) \quad \text{for } h \to 0.$$

Hence, the FDM is also convergent of order p. □

Theorem II.13: Assume that the exact solution to (II.P) is four times continuously differentiable ($u \in C^4(\overline{\Omega})$), then the FDM (II.DP) is consistent of order 2 w. r. t. the system of maximum norms ($\|v\|_h = \max_{1 \leq i \leq N} |v_i|$).

Proof. This follows directly from Theorem II.3 since

$$\begin{aligned} D^0 u(x) &= u'(x) + \mathcal{O}(h^2), \\ D^- D^+ u(x) &= u''(x) + \mathcal{O}(h^2) \end{aligned}$$

for all $x \in \Omega_h$. □

Example II.14: Consider

$$\begin{cases} -\frac{1}{2} u''(x) = x(1-x) & \text{on } \Omega = (0,1), \\ u(0) = u(1) = 0. \end{cases}$$

By integrating the ODE twice, the exact solution is found to be

$$u(x) = \frac{1}{6} x^4 - \frac{1}{3} x^3 + \frac{1}{6} x, \quad x \in \Omega.$$

The FDM approximation is given by

$$L_h u_h = -\frac{1}{2h^2}(1, -2, 1) u_h = f_h$$

with $f_h = \big[x_1(1-x_1), \ldots, x_N(1-x_N)\big]^\mathsf{T}$, where $x_j = jh$ and $h = \frac{1}{N+1}$.

Observe that the exact solution is a polynomial which for this particular example is four times continuously differentiable. Hence, Theorem II.13 ensures the consistency of order 2. Further, numerical experiments indicate

$$\big\|\!\big|L_h^{-1}\big|\!\big\|_{\infty,h} \leq \frac{1}{4}.$$

The method is stable (at least in the experiments) and, hence, convergent of order 2. Further below we will give a theoretical proof that the method is indeed stable.

Remark II.15: Why is the order of convergence actually important? Let $\varepsilon > 0$ be a given error tolerance. How should we choose $N \in \mathbb{N}$ (or h) such that $\|u_h - R_h u\|_h \in \mathcal{O}(\varepsilon)$?

If $\|u_h - R_h u\|_h \in \mathcal{O}(h^p)$, we need to choose $h \in \mathcal{O}\Big(\varepsilon^{\frac{1}{p}}\Big)$ or $N \in \mathcal{O}\Big(\varepsilon^{-\frac{1}{p}}\Big)$. Therefore, if p is larger, N can be chosen smaller which is less expensive.

Thus, if two stable numerical methods with the same computational cost are given, we should prefer the method with the higher order of convergence since it may provide the same accuracy with a larger (= less expensive) choice of the step size.

Example II.16: For given $a \in (0,\infty)$ consider the BVP

$$\begin{cases} -au''(x) - u'(x) = 0 \quad \text{in } \Omega = (0,1), \\ u(0) = 0, \\ u(1) = 1. \end{cases} \tag{II.16}$$

The unique exact solution to (II.16) is given by

$$u(x,a) = \frac{1 - \mathrm{e}^{-\frac{x}{a}}}{1 - \mathrm{e}^{-\frac{1}{a}}}$$

for all $x \in (0,1)$ and $a \in (0,\infty)$. In this example we study the effect of the parameter a on the exact and the discrete solution. For instance, Figure II.6 shows the exact solution for two different values of a.

In this situation, the standard FDM is given by (with $h = \frac{1}{N+1}$)

$$\begin{cases} -aD^-D^+u_h(x,a) - D^0u_h(x,a) = 0 \quad \text{for } x \in \Omega_h, \\ u_h(0,a) = 0, \\ u_h(1,a) = 1. \end{cases}$$

It can be shown that

$$u_h(jh,a) = \frac{1 - r^j}{1 - r^{N+1}} \quad \text{with } r = \frac{2a-h}{2a+h}$$

is the discrete solution. Numerical experiments indicate that the FDM is stable. Thus, we can safely expect that the FDM is convergent of order 2.

Now, we have the following two observations:

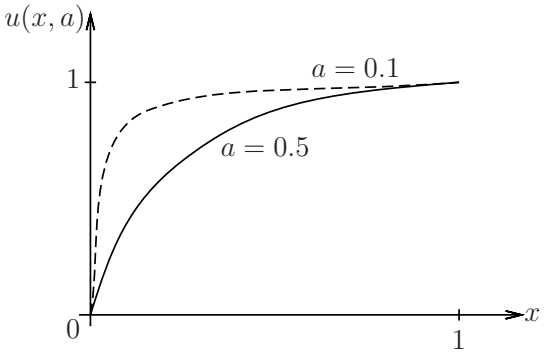
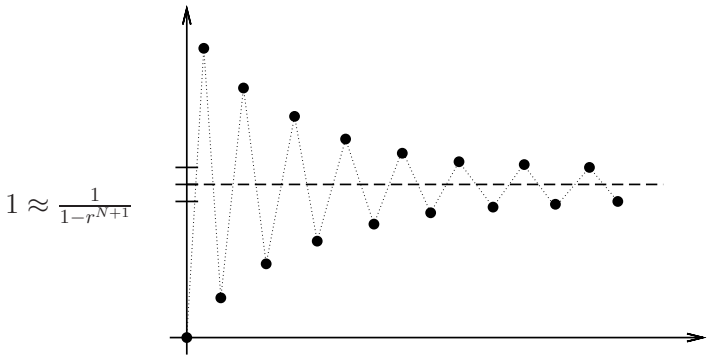


Figure II.6.: Exact solution to (II.16) for two different values of the parameter a.

a) If $h > 2a$, then it holds that $r < 0$. Moreover, $u_h(jh, a) = \frac{1}{1-r^{N+1}} - r^j \frac{1}{1-r^{N+1}}$ and thus the discrete solution has oscillations since r^j has alternating sign. As a consequence, the discrete solution does not look like the exact solution at all. To avoid these oscillations, we need to choose N sufficiently large, but then solving the linear system becomes more expensive!



b) If $a = \frac{1}{N+1} = h$ for some large $N \in \mathbb{N}$ then the FDM solution at $x_1 = h$ (the first grid point) becomes

$$u_h(x_1, a) = \frac{1 - \frac{1}{3}}{1 - (\frac{1}{3})^{N+1}} \approx 1 - \frac{1}{3} \quad \text{for large } N.$$

On the other hand, the exact solution is

$$u(h, a) = \frac{1 - \mathrm{e}^{-1}}{1 - \mathrm{e}^{-(N+1)}} \approx 1 - \mathrm{e}^{-1}.$$

Then the relative error at $x_1 = h$ is

$$\frac{|u(h,a) - u_h(h,a)|}{|u(h,a)|} \approx 0.05465 \approx 5\ \% \text{ off the exact solution at the first grid point.}$$

This holds true for *any* choice of the step size h!

Note that this does not contradict the convergence results from this section, since a varies with h. However, the convergence theorem is only valid for a fixed parameter value a which is independent of h and for $h \to 0$.

The above two observations lead to the following conclusions:

- Never couple free problem parameters with your numerical step size!
- In practice, the BVP/PDE can require very large values for $N \in \mathbb{N}$ in order to obtain a "good" approximation.
- There is not one $N \in \mathbb{N}$ that gives a good approximation for every BVP/PDE. Even if everything is correctly implemented and all assumptions of the convergence theorem are satisfied the numerical solution can still be far-off the exact solution if the step size h is not small enough. Do not blindly trust a numerical solution (in particular in critical applications).
- Compare two numerical solutions with different step sizes, say h and $\frac{h}{2}$. If they look completely different, one should not trust them, but take smaller values for h. In particular checking if the error decreases as predicted by theory can be a good indicator.
- Oscillations in elliptic/parabolic problems usually indicate that N is not large enough!

Let us conclude this subsection with a final remark on the involved norms.

Remark II.17: Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two norms on $\mathbb{R}^N$. Then there exists $m > 0$, $M > 0$ such that

$$m\|w\|_a \leq \|w\|_b \leq M\|w\|_a.$$

In this sense, all norms on $\mathbb{R}^N$ are *equivalent*. For example, a constant M_{ab} with $\|w\|_a \leq M_{ab}\|w\|_b$ for all $w \in \mathbb{R}^N$ is given in the following table:

a \ b	1	2	∞
1	1	$\sqrt{N}$	N
2	1	1	$\sqrt{N}$
∞	1	1	1

$$\|w\|_1 = \sum_{i=1}^{N} |w_i|,$$

$$\|w\|_2 = \left(\sum_{i=1}^{N} |w_i|^2 \right)^{\frac{1}{2}},$$

$$\|w\|_\infty = \max_{1 \leq i \leq N} |w_i|.$$

The same holds true for matrix norms, the respective constants are listed in the following table:

a \ b	1	2	∞	F
1	1	$\sqrt{N}$	N	$\sqrt{N}$
2	$\sqrt{N}$	1	$\sqrt{N}$	$\sqrt{N}$
∞	N	$\sqrt{N}$	1	$\sqrt{N}$
F	$\sqrt{N}$	$\sqrt{N}$	$\sqrt{N}$	1

Why is this important to us? Consider a system of norms $(\|\cdot\|_{1,h})_{h>0}$ and define $\|w\|_{2,h} := \frac{1}{h^3}\|w\|_{1,h}$. Thus, the discretization error may vanish with order 2 w. r. t. $(\|\cdot\|_{1,h})_{h>0}$ but it may be divergent w. r. t. $(\|\cdot\|_{2,h})_{h>0}$ since

$$\|u_h - R_h u\|_{2,h} = \frac{1}{h^3}\|u_h - R_h u\|_{1,h} \leq C\frac{1}{h},$$

which blows up for $h \to 0$. So when writing about convergence, it is important to agree on the family of norms!

II.5. Higher-Dimensional Elliptic BVP

We consider the higher-dimensional Poisson equation on hypercubes $\Omega = (0,1)^n$

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = g, & \text{on } \partial\Omega, \end{cases} \tag{II.17}$$

again with Dirichlet boundary conditions on $\partial\Omega$. We start by considering the two-dimensional extension of the previous considerations.

II.5.1. Two dimensions

We follow the same strategy as in 1D:
Step 1: Discretize Ω by a mesh Ω_h:

$$\begin{aligned} \Omega_h &= \{(kh, lh) \mid k, l = 1, \ldots, N\} \quad \text{where } h = \frac{1}{N+1},\ N \in \mathbb{N}, \\ \overline{\Omega}_h &= \{(kh, lh) \mid k, l = 0, \ldots, N+1\}, \\ \Gamma_h &= \overline{\Omega}_h \backslash \Omega_h \quad \Leftrightarrow \quad \overline{\Omega}_h = \Omega_h \cup \Gamma_h. \end{aligned}$$

Step 2: Approximation of partial derivatives by difference operators: Fix $x = \begin{bmatrix} x_1, & x_2 \end{bmatrix}^\mathsf{T} \in \Omega_h$. We want to approximate Δu by D^+, D^- etc. Define $w_1(x_1) := u(x_1, x_2)$ (for fixed x_2). Then we have

$$\begin{aligned} D^- D^+ w_1(x_1) &= \frac{w_1(x_1+h) - 2w_1(x_1) + w_1(x_1-h)}{h^2} \\ &= w_1''(x_1) + \mathcal{O}(h^2), \end{aligned}$$

if $u \in C^4(\overline{\Omega})$ which implies $w_1 \in C^4([0,1])$. In the same way, for $w_2(x_2) := u(x_1, x_2)$:

$$\begin{aligned} D^- D^+ w_2(x_2) &= \frac{w_2(x_2+h) - 2w_2(x_2) + w_2(x_2-h)}{h^2} \\ &= w_2''(x_2) + \mathcal{O}(h^2), \end{aligned}$$

if $u \in C^4(\overline{\Omega})$ which implies $w_2 \in C^4([0,1])$.

With this we get

$$\begin{aligned} \Delta u = u_{x_1x_1} + u_{x_2x_2} &= w_1''(x_1) + w_2''(x_2) \\ &\approx D^- D^+ w_1(x_1) + D^- D^+ w_2(x_2) \\ &= \frac{w_1(x_1+h) - 2w_1(x_1) + w_1(x_1-h)}{h^2} + \frac{w_2(x_2+h) - 2w_2(x_2) + w_2(x_2-h)}{h^2} \end{aligned}$$

Replace w_1, w_2 by u:

$$\Delta u \approx \frac{u(x_1+h, x_2) + u(x_1, x_2+h) - 4u(x_1, x_2) + u(x_1-h, x_2) + u(x_1, x_2-h)}{h^2}$$

This is called *5-point difference stencil* for the Laplace operator $-\Delta u$ in 2D. In order to symbolize the 5-point stencil we draw it as shown in Fig. II.7, where the weights $1/h^2$ are usually omitted, since a uniform mesh width is assumed.

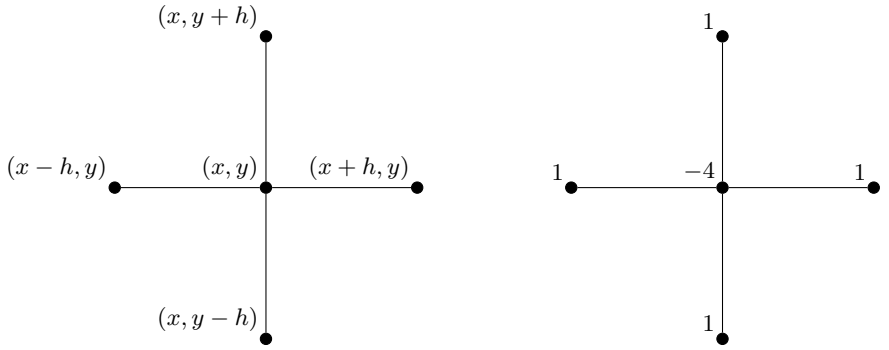


Figure II.7.: 5-point difference stencil for $-\Delta u$ in two spatial dimensions showing (left) the vertices/grid points and (right) the corresponding weights.

Step 3: Derivation of the system of linear equations: Let $u_h : \overline{\Omega}_h \to \mathbb{R}$ be a grid function which approximates u on Ω_h, i. e., $u_h(x) \approx u(x)$ for all $x = \begin{bmatrix} x_1, & x_2 \end{bmatrix}^\mathsf{T} \in \Omega_h$. Consider

$$\begin{cases} -\Delta u = f & \text{in } \Omega = (0,1) \times (0,1), \\ u = g, & \text{on } \partial\Omega = \Gamma. \end{cases} \tag{II.PDir}$$

For short, we write:

$$\begin{aligned} u_{k,l} &:= u_h(kh, lh), \quad k, l = 0, \ldots, N+1, \\ f_{k,l} &:= f(kh, lh), \quad k, l = 1, \ldots, N, \\ g_{k,l} &:= g(kh, lh), \quad (kh, lh) \in \Gamma_h. \end{aligned}$$

By replacing $-\Delta u$ by the five point difference stencil we obtain

$$\begin{cases} -\frac{1}{h^2}(u_{k+1,l} + u_{k,l+1} - 4u_{k,l} + u_{k-1,l} + u_{k,l-1}) = f_{k,l}, & k, l = 1, \ldots, N, \\ u_{k,l} = g_{k,l} \quad \text{for } (kh, lh) \in \Gamma_h. \end{cases}$$

This leads to a system of linear equations on the $(N+2)^2$ unknowns $u_{k,l}$, $k, l = 0, \ldots, (N+1)$. As in 1D, we can eliminate the Dirichlet boundary conditions and write this as $L_h u_h = f_h$. Then, $L_h \in \mathbb{R}^{N^2 \times N^2}$, $u_h \in \mathbb{R}^{N^2}$, $f_h \in \mathbb{R}^{N^2}$. Alternatively we can explicitly include the boundary conditions, which leads to $L_h \in \mathbb{R}^{(N+2)^2 \times (N+2)^2}$, $u_h \in \mathbb{R}^{(N+2)^2}$, $f_h \in \mathbb{R}^{(N+2)^2}$.

Book-keeping of degrees of freedoms is important: Keep track of which entry in $u_h \simeq u_{k,l}$ corresponds to a component of the solution vector u_h. For this we need a one-to-one mapping $(k, l) \to i_{k,l}$ such that $(u_h)_{i_{k,l}} = u_{k,l}$. The standard way is to use the lexicographical ordering, shown below in Fig. II.8 for $N = 4$

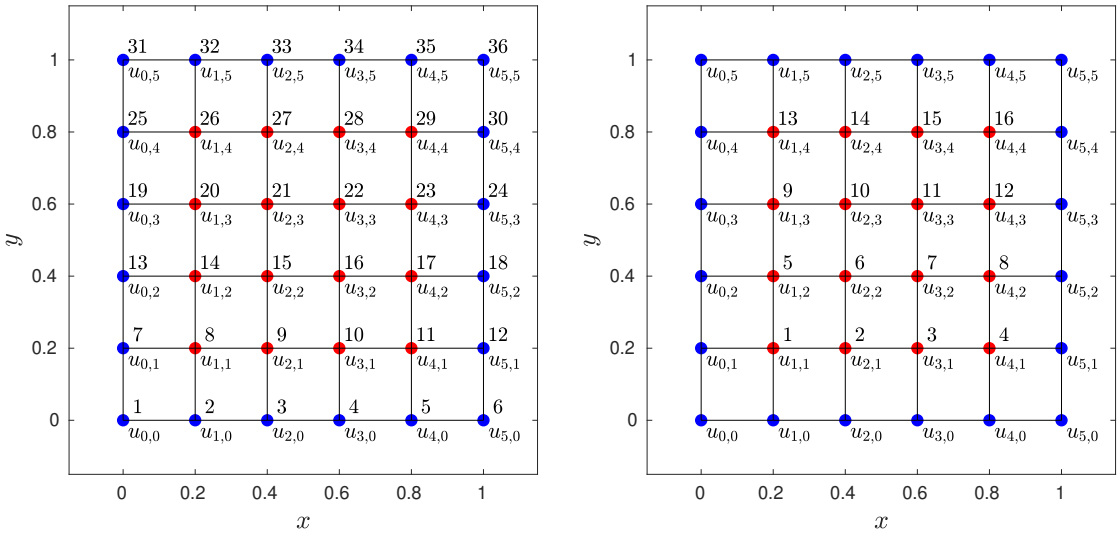


Figure II.8.: Lexicographical order of $u_{k,l} = u_h(kh, lh)$ for degrees of freedom in Ω_h (•) and on Γ_h (•) shown (**left**) for u_h defined on $\overline{\Omega}_h$ and (**right**) for reduced u_h defined on Ω_h.

In two dimensions this corresponds to the index mapping $i_{k,l} = 1 + k + l(N+2)$ for $u_h : \overline{\Omega}_h \to \mathbb{R}$ and $i_{k,l} = k + (l-1)N$ for $u_h : \Omega_h \to \mathbb{R}$. The main point here is that this easily allows us to compute the indices of neighbors by considering $i_{k\pm1,l}$ and $i_{k,l\pm1}$ needed for the implementation of the difference stencil. Then we get for $u_h : \overline{\Omega}_h \to \mathbb{R}$

$$u_h = \left[u_{0,0}, u_{1,0}, \ldots, u_{N+1,0}, u_{0,1}, u_{1,1}, \ldots, u_{N+1,1}, u_{0,2}, \ldots, u_{N+1,N+1}\right]^{\mathsf{T}} \in \mathbb{R}^{(N+2)^2}.$$

or for the reduced grid function $u_h : \Omega_h \to \mathbb{R}$

$$u_h = \left[u_{1,1}, u_{2,1}, \ldots, u_{N,1}, u_{1,2}, u_{2,2}, \ldots, u_{N,2}, u_{1,3}, \ldots, u_{N,N}\right]^{\mathsf{T}} \in \mathbb{R}^{N^2}.$$

The five point stencil then results into the matrix

$$L_h = -\frac{1}{h^2}\begin{bmatrix} T & I_N & & & \\ I_N & T & I_N & & \\ & \ddots & \ddots & \ddots & \\ & & I_N & T & I_N \\ & & & I_N & T \end{bmatrix} \in \mathbb{R}^{N^2 \times N^2},$$

where $h = \frac{1}{N+1}$ and $I_N \in \mathbb{R}^{N \times N}$ denotes the identity matrix and the matrix T is given by

$$T = \begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{bmatrix} \in \mathbb{R}^{N \times N}.$$

The vector f_h on the right hand side is given by

$$f_h = \left[f_{1,1}, \ldots, f_{N,1}, f_{1,2}, \ldots, f_{N,2}, f_{1,3}, \ldots, f_{N,N}\right]^\mathsf{T} + \frac{1}{h^2}\left[g_{0,1} + g_{1,0}, g_{2,0}, \ldots, g_{N,0} + g_{N+1,1}, \ldots\right]^\mathsf{T} \in \mathbb{R}^{N^2}.$$

Note that f_h has to use the same lexicographical ordering as u_h. The second part consists of the eliminated boundary values of u. It is (in general) nonzero for any inner grid point that is close to the boundary (meaning that at least one neighbouring grid point lies on the boundary Γ_h). An examplary solution of the Poisson problem is shown in Fig. II.9.

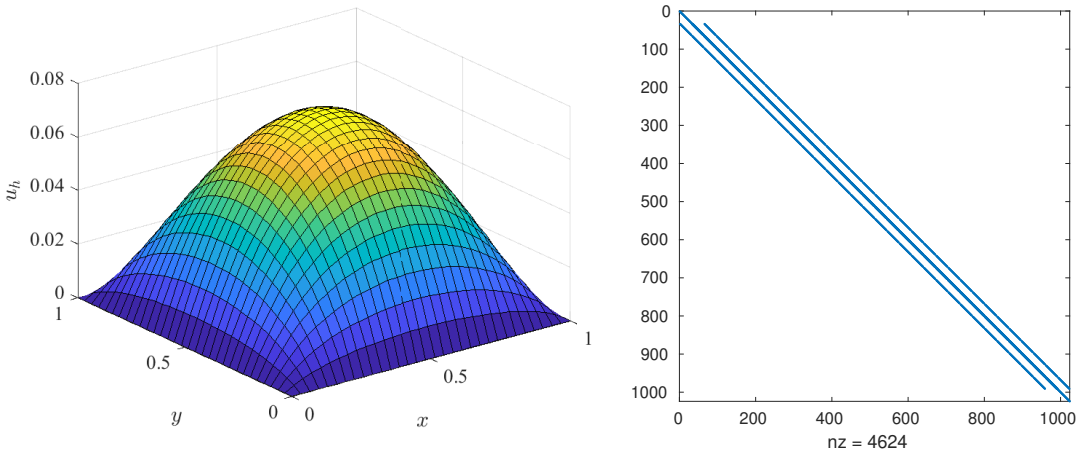


Figure II.9.: **(left)** Numerical solution u_h of the 2D Poisson problem $L_h u_h = f$ with homogeneous Dirichlet boundary conditions on $\Omega = (0,1)^2$ with $f(x) = 1$ using the standard 5-point stencil for $Lu = -\Delta u$. **(right)** Sparsity pattern of the matrix $L_h \in \mathbb{R}^{(N+2)^2 \times (N+2)^2}$ (before reduction) with $N = 30$.

Next we will perform an error analysis. The definitions of consistency, stability, and convergence carry over to the 2D case. From the derivation of the 2D-FDM we directly get the following result.

Lemma II.18: Assume that $u \in C^4(\overline{\Omega})$ is the exact solution to (II.PDir). Let $L_h u_h = f_h$ be the linear system for the 5-point-stencil FDM. Then it holds:

a) The matrix $L_h \in \mathbb{R}^{N^2 \times N^2}$ is symmetric.

b) The FDM is consistent of order 2 with respect to the maximum norms on $\mathbb{R}^N$.

It remains to prove stability, for which we can proceed using two similar line of arguments using M-matrices or the discrete maximum principle.

Proof via M-matrices

For this, we need the following definition:

Definition II.19: Let $A = [a_{ij}]_{i,j=1}^N$ be a matrix. If

a) $a_{ii} > 0 \; \forall \, i = 1, \ldots, N$ and $a_{ij} \leq 0 \; \forall \, i \neq j$,

b) $\det(A) \neq 0$,

c) $A^{-1} \geq 0$ (that is all entries in A^{-1} are non-negative),

then A is called an *M-matrix*.

Let $\mathbb{1} := \big[1, \ldots, 1\big]^\mathsf{T} \in \mathbb{R}^N$. Again, similarly as in the above definition, we understand "$\leq$" entrywise in the following.

Theorem II.20: Let $A \in \mathbb{R}^{N \times N}$ be an M-matrix. If there exists a vector $w \in \mathbb{R}^N$ with $Aw \geq \mathbb{1}$, then it holds

$$\big\| A^{-1} \big\|_\infty \leq \|w\|_\infty .$$

Proof. Let $y = \big[y_1, \ldots, y_N\big]^\mathsf{T} \in \mathbb{R}^N$ be arbitrary. We write $|y|$ for the vector with entries $|y| = \big[|y_1|, \ldots, |y_N|\big]^\mathsf{T} \in \mathbb{R}^N$. Then it holds true that

$$|y| \leq \|y\|_\infty \cdot \mathbb{1} \leq \|y\|_\infty \cdot Aw \tag{II.18}$$

by our assumption on $w \in \mathbb{R}^N$. Since A is an M-matrix, we have $A^{-1} \geq 0$. From this and the triangle inequality it follows

$$\big|A^{-1} y\big|_i = \left| \sum_{j=1}^N \big[A^{-1}\big]_{ij} y_j \right| \leq \sum_{j=1}^N \big[A^{-1}\big]_{ij} |y_j| = \big[A^{-1} |y|\big]_i .$$

for each component $i = 1, \ldots, N$. In vector notation with $\leq$ understood entrywise, this reads

$$\left|A^{-1}y\right| \leq A^{-1}|y| \leq A^{-1}\|y\|_\infty Aw = \|y\|_\infty w,$$

where we inserted (II.18). After taking the maximum norm we therefore obtain

$$\left\|A^{-1}y\right\|_\infty = \max_{1\leq i\leq N} \left|A^{-1}y\right|_i \leq \|w\|_\infty \|y\|_\infty,$$

or, equivalently,

$$\frac{\left\|A^{-1}y\right\|_\infty}{\|y\|_\infty} \leq \|w\|_\infty.$$

Now observe that the right hand side is independent of y. Since y was arbitrary we can therefore take the supremum over all y. This yields

$$\left|\left|\left|A^{-1}\right|\right|\right|_\infty = \sup_{y\in\mathbb{R}^N, \|y\|_\infty>0} \frac{\left\|A^{-1}y\right\|_\infty}{\|y\|_\infty} \leq \|w\|_\infty$$

as claimed. □

Theorem II.21: Let

$$L_h = -\frac{1}{h^2}\begin{bmatrix} T & I_N & & & \\ I_N & T & I_N & & \\ & \ddots & \ddots & \ddots & \\ & & I_N & T & I_N \\ & & & I_N & T \end{bmatrix} \in \mathbb{R}^{N^2\times N^2},$$

$$T = \begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{bmatrix} \in \mathbb{R}^{N\times N}$$

be the matrix of the 5-point difference stencil. Then

a) L_h is an M-matrix.

b) $\left|\left|\left|L_h^{-1}\right|\right|\right|_\infty \leq \frac{1}{8}$ for all $h > 0$ (thus, L_h is stable).

Proof. a) The first condition for an M-matrix is satisfied by L_h. (We will not show the other two properties.)

b) Let $w(x_1,x_2)=\frac{x_1-x_1^2}{2}=\frac{1}{2}x_1(1-x_1)$. Let $w_h=R_hw$ be the restriction of w to Ω_h. The entries of $w_h\in\mathbb{R}^{N^2\times N^2}$ are ordered lexicographically and consist of $w_{k,l}=w(kh,lh)$. In the interior of the grid, a typical entry in the vector L_hw_h is

$$\begin{aligned}
&-\frac{1}{h^2}(w_{k+1,l}+w_{k,l+1}-4w_{k,l}+w_{k-1,l}+w_{k,l-1})\\
&=\frac{1}{h^2}(2w_{k,l}-w_{k-1,l}-w_{k+1,l}) \quad (w \text{ independent of } x_2\colon\ w_{k,l-1}=w_{k,l}=w_{k,l+1})\\
&=\frac{1}{h^2}\left(2\frac{kh-k^2h^2}{2}-\frac{(k-1)h-(k-1)^2h^2}{2}-\frac{(k+1)h-(k+1)^2h^2}{2}\right)\\
&=\frac{1}{2h^2}\big(-2k^2h^2+(k-1)^2h^2+(k+1)^2h^2\big)\\
&=-\frac{1}{2}\big(2k^2-(k-1)^2-(k+1)^2\big)\\
&=-\frac{1}{2}\left(2k^2-(k^2-2k+1)-(k^2+2k+1)\right)=-\frac{1}{2}(2k-2k-1-1)\\
&=1.
\end{aligned}$$

In the same way we verify that $-\frac{1}{h^2}(\ldots)>1$ when we are close to the boundary. Hence we get $L_hw_h\geq\mathbb{1}$. Theorem II.20 then shows

$$\left\|\left\|L_h^{-1}\right\|\right\|_\infty\leq\|w_h\|_\infty=\max_{k,l=1,\ldots,n}|w(kh,lh)|\leq\max_{x\in[0,1]}\frac{x-x^2}{2}=\frac{1}{8}. \qquad \square$$

Proof via maximum principle

For the following arguments we will make some use of the particular structure of the standard difference stencil of the discrete Laplace operator in nD, where $s_\alpha=-2n$ for $\|\alpha\|=0$ and $s_\alpha=1$ for $\|\alpha\|=1$. Assuming additionally that the domain $\overline{\Omega}_h$ is discretely connected, i.e., all points can be reached by moving the difference stencil, allows us to make the following two statements.

Theorem II.22 (Discrete maximum principle)**:** Let $u_h:\overline{\Omega}_h\subset\mathbb{R}^n\to\mathbb{R}$ a grid function with $\Delta_hu_h\geq 0$ on Ω_h using the standard stencil. Then $\max_{\Omega_h}u_h\leq\max_{\Gamma_h}u_h$ and equality holds if and only if u_h is constant.

Proof. If the maximum is attained on Γ_h the proof is finished. Now suppose the contrary $\max_{\Omega_h}u_h>\max_{\Gamma_h}u_h$ and let $x_0\in\Omega_h$ be a point where the maximum is attained and x_i for $i=1\ldots 2n$ are its nearest neighbors. Using the standard stencil we have

$$2nu_h(x_0)=\sum_{i=1}^{2n}u_h(x_i)-h^2\underbrace{\Delta_hu_h(x_0)}_{\geq 0}\leq\sum_{i=1}^{2n}u_h(x_i)\leq 2nu_h(x_0),$$

where in the last step we used $u_h(x_i)\leq u_h(x_0)$. This implies $u_h(x_i)=u_h(x_0)$ for all neighbors. Successively (proof by induction) repeat the argument with neighbors over

the whole domain (since domain is discretely connected) to find u_h is constant on $\overline{\Omega}_h$, contradicting the assertion and finishing the first part of the proof. The second part of the proof works analogously. □

Theorem II.23 (Existence and uniqueness of discrete Poisson problem)**:** There is a unique solution $u_h : \overline{\Omega}_h \to \mathbb{R}$ to the discrete problem

$$\begin{aligned} -\Delta_h u_h &= f, \qquad \text{in } \Omega_h, \\ u_h &= g, \qquad \text{on } \Gamma_h. \end{aligned}$$

Proof. We only need to show $-\Delta_h$ is nonsingular, i.e., $-\Delta_h u_h = 0$ in Ω_h, $u_h = 0$ on Γ_h if and only if $u_h = 0$. Apply maximum principle to u_h and $-u_h$ and get $0 \leq u_h \leq 0$ and hence $u_h = 0$. □

Theorem II.24 (Discrete continuous dependence on data)**:** Let u_h solve

$$\begin{aligned} -\Delta_h u_h &= f, \qquad \text{in } \Omega_h \subset \mathbb{R}^n, \\ u_h &= g, \qquad \text{on } \Gamma_h, \end{aligned}$$

then

$$\max_{\overline{\Omega}_h} |u_h| \leq C \max_{\Omega_h} |f| + \max_{\Gamma_h} g,$$

for a constant $C \in \mathbb{R}$ independent from f, g (and of course h).

Proof. Consider a domain $\overline{\Omega}_h$ contained in ball of radius R around x_0. Defining an auxilliary function $\phi(x) = R^2 - \|x - x_0\|^2$, then we have $0 \leq \phi(x) \leq R^2$ for all points $x \in \overline{\Omega}_h$. Furthermore, by the properties of the discrete Laplacian (exact for polynomials) we have $-\Delta_h \phi = 2n$. Now define the function

$$v(x) = \max_{\Gamma_h} |g| + \tfrac{1}{2n} \phi(x) \max_{\Omega_h} |f|.$$

By construction we have $-\Delta_h v(x) = \max_{\Omega_h} |f| \geq f(x) = -\Delta_h u_h(x)$. This implies

$$\Delta_h (u_h - v) = (\max_{\Omega_h} |f|) - f \geq 0,$$

for all points in Ω_h and $u_h - v \leq 0$ for all points on Γ_h by construction. Applying the maximum principle to $w = (u_h - v)$ we find $\max_{\Omega_h} w \leq \max_{\Gamma_h} w \leq 0$, such that $\max_{\overline{\Omega}_h} u_h \leq \max_{\overline{\Omega}_h} v$. Using a similar argument with $-v$ we obtain

$$\max_{\overline{\Omega}_h} |u_h| \leq \max_{\overline{\Omega}_h} v = \max_{\Gamma_h} |g| + \max_{x \in \overline{\Omega}_h} \left[\frac{\phi(x)}{2n} \max_{\Omega_h} |f| \right] \leq C \max_{\Omega_h} |f| + \max_{\Gamma_h} g$$

with $C = R^2/(2n)$ due to $0 \leq \phi \leq R^2$. □

Corollary II.25: The constant C in Thm. II.24 also provides the constant for the stability estimate $\left\|\left\|L_h^{-1}\right\|\right\|_h \leq C$. On $\Omega = (0,1)^2$ we have $C = R^2/(2n) = 1/8$.

In the following, we will analyze the order of convergence by numerical experiments. The order of convergence of a numerical method can be visualized or determined in numerical experiments. For this, assume that the discretization error satisfies

$$\operatorname{err}(h) = \|u_h - R_h u\| = Ch^p$$

We want to determine the order p. Therefore, we compute the errors for two different (sufficiently small) step sizes h_1, h_2. This gives

$$\log(\operatorname{err}(h_1)) = \log(C) + \log(h_1^p) = \log(C) + p\log(h_1),$$

and therefore, we get

$$\frac{\log(\operatorname{err}(h_1)) - \log(\operatorname{err}(h_2))}{\log(h_1) - \log(h_2)} = p =: \operatorname{eoc}(h_1, h_2).$$

If the exact solution is unknown, then one often replaces $R_h u$ by $R_h u_{\widetilde{h}}$, where $u_{\widetilde{h}}$ is a discrete solution with a much smaller step size $\widetilde{h} \ll h$.

II.5.2. Index Order and Matrices on Structured and on General Domains

Structured Domains

The reduced form of the matrix L_h with lexicographical ordering in this section has an even more special structure, namely, it can be written in terms of Kronecker or tensor products which we will briefly discuss now.

Let $S = -\frac{1}{h^2}(1, -2, 1) \in \mathbb{R}^{N\times N}$ be the tridiagonal matrix from Section II.1. We can write $L_h = (I_N \otimes S) + (S \otimes I_N)$. Here, $\otimes$ is the *tensor product* or *Kronecker product* of two matrices. For $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{p\times q}$ it is defined by

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & & \vdots \\ a_{n1}B & \cdots & a_{nm}B \end{bmatrix} \in \mathbb{R}^{(np)\times(mq)}.$$

This leads to

$$I_N \otimes S = \begin{bmatrix} S & & 0 \\ & \ddots & \\ 0 & & S \end{bmatrix} \in \mathbb{R}^{N^2\times N^2},$$

$$S \otimes I_N = -\frac{1}{h^2}\begin{bmatrix} -2I_N & 1I_N & & & \\ 1I_N & -2I_N & 1I_N & & \\ & \ddots & \ddots & \ddots & \\ & & 1I_N & -2I_N & 1I_N \\ & & & 1I_N & -2I_N \end{bmatrix} \in \mathbb{R}^{N^2\times N^2}.$$

Here, the matrix $I_N \otimes S$ can be interpreted as a discretization of the differential operator $-\frac{\partial^2}{\partial x_1^2}$ on the grid Ω_h, whereas $S \otimes I_N$ is interpreted as a discretization of the operator $-\frac{\partial^2}{\partial x_2^2}$ on the grid Ω_h. Therefore, and since $T = S - 2I_N$, $L_h = (I_N \otimes S) + (S \otimes I_N)$ is indeed a discretization of the negative Laplace operator $-\Delta = -\left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}\right)$.

This is a very helpful observation for assembling the matrix L_h (see the built-in functions "`kron`" in `MATLAB` and "`scipy.sparse.kron`" in `python`).

The Kronecker product has the following useful properties:

a) $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ for matrices A, B, C, D of conforming dimensions,

b) $(A \otimes B)^\mathsf{T} = A^\mathsf{T} \otimes B^\mathsf{T}$,

c) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, if both A and B are invertible.

This has a lot of advantages in scientific computing:

- If $L_h = A \otimes B$ has a tensor structure, we only need to save $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{p\times q}$. This only occupies
$$\text{const.} \cdot (nm + pq) \text{ bytes},$$
while storing the full matrix $M = \mathbb{R}^{np\times mq}$ needs const. $\cdot\, (nmpq)$ bytes.
- Clever algorithms like the *fast Fourier transformation (FFT)* [HB09, Sec. 94] or the *alternating directions implicit (ADI) iteration* [Wac13] often decrease the number of arithmetical operations significantly if a tensor structure is found.

General Domains

The matrix structure above is valid for lexicographical ordering and simple domains, i.e., boxes $\Omega = (0, L)^n$ discretized with uniform meshes. Now we consider a method that allows us to solve the Poisson equation with Dirichlet boundary conditions on arbitrary domains. Assuming the continuous dependence on the data also holds for perturbations of the domain shape, we consider $\overline{\Omega}_h$ to be an approximation of the actual shape $\overline{\Omega}$ in the sense that points in Γ_h are not contained in Γ_h. We consider the problem

$$-\Delta u = f \qquad \text{in } \Omega \subset \mathbb{R}^n, u = g \qquad\qquad \text{on } \Gamma = \partial\Omega,$$

where $f, g : \Omega \to \mathbb{R}$ are sufficiently smooth. Furthermore, assume $\overline{\Omega} \subset [0, 1]^n$, which we can always obtain for bounded domains by a shift and rescaling of the original domain. Now define $\overline{\Omega}_h = \Omega \cap h\mathbb{Z}^n$, i.e.,

$$\overline{\Omega}_h = \{x = (i_1 h, \ldots, i_n h) \in \mathbb{R}^n : x \in \Omega, i_k \in \mathbb{Z}\} \tag{II.19}$$

for $h = 1/(N + 1)$ as usual. Correspondingly, for the standard $(2n + 1)$ stencil a point x is in Ω_h, if all direct neighbors $x \pm e_k h$ for $k = 1, ..., n$ are contained in $\overline{\Omega}_h$, i.e.,

$$\Omega_h = \{x \in \overline{\Omega}_h : \text{if all neighbors } x \pm he_k \in \overline{\Omega}_h\}, \qquad \Gamma_h = \overline{\Omega}_h \setminus \Omega_h. \tag{II.20}$$

such that $\overline{\Omega}_h = \Omega_h \cup \Gamma_h$.

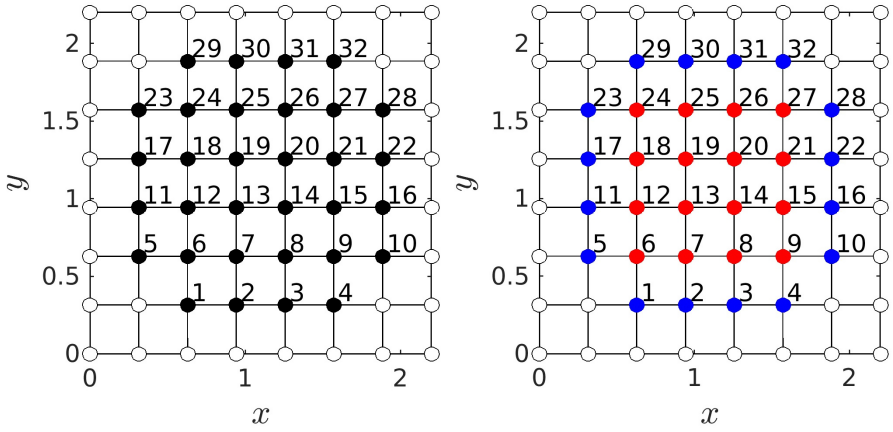


Figure II.10.: Lexicographical ordering on a non-box 2D domain $\overline{\Omega}_h$ **(left)** showing points in $\overline{\Omega}_h$ and **(right)** showing points in Ω_h (red) and in Γ_h (blue).

The definition of Ω_h, Γ_h in (II.21) changes for a general compact stencil to

$$\Omega_h = \{x \in \overline{\Omega}_h : x \pm \alpha h \in \overline{\Omega}_h, -1 \leq \alpha_i \leq 1\}, \tag{II.21a}$$

$$\Gamma_h = \overline{\Omega}_h \setminus \Omega_h. \tag{II.21b}$$

This allows us to discretize the Poisson problem with Dirichlet boundary conditions as

$$-\Delta_h u_h = f \qquad \text{in } \Omega_h, \tag{II.22}$$

$$u_h = g \qquad \text{on } \Gamma_h, \tag{II.23}$$

where we need to evaluate g on Γ_h, which does not necessarily overlap with Γ. This is possible since here $g : \overline{\Omega} \to \mathbb{R}$ instead of the previous $g : \Gamma \to \mathbb{R}$ and since $\overline{\Omega}_h \subset \overline{\Omega}$. Using the discrete continuous dependence on data with $w_h = u_h - R_h u$ we get

$$\max_{\overline{\Omega}_h} |w_h| \leq C \max_{\Omega_h} |f - L_h R_h u| + \max_{\Gamma_h} |w_h|. \tag{II.24}$$

The first term goes to zero because of the consistency of the difference stencil. The second term goes to zero for Lipschitz continuous solution and boundary data since we have $|w_h| \leq Lh$ on Γ_h. For problems with general domains with Dirichlet boundary conditions we need to require $\|g - R_h u_h\|_{\Gamma,h} \to 0$ as part of the consistency requirement as well.

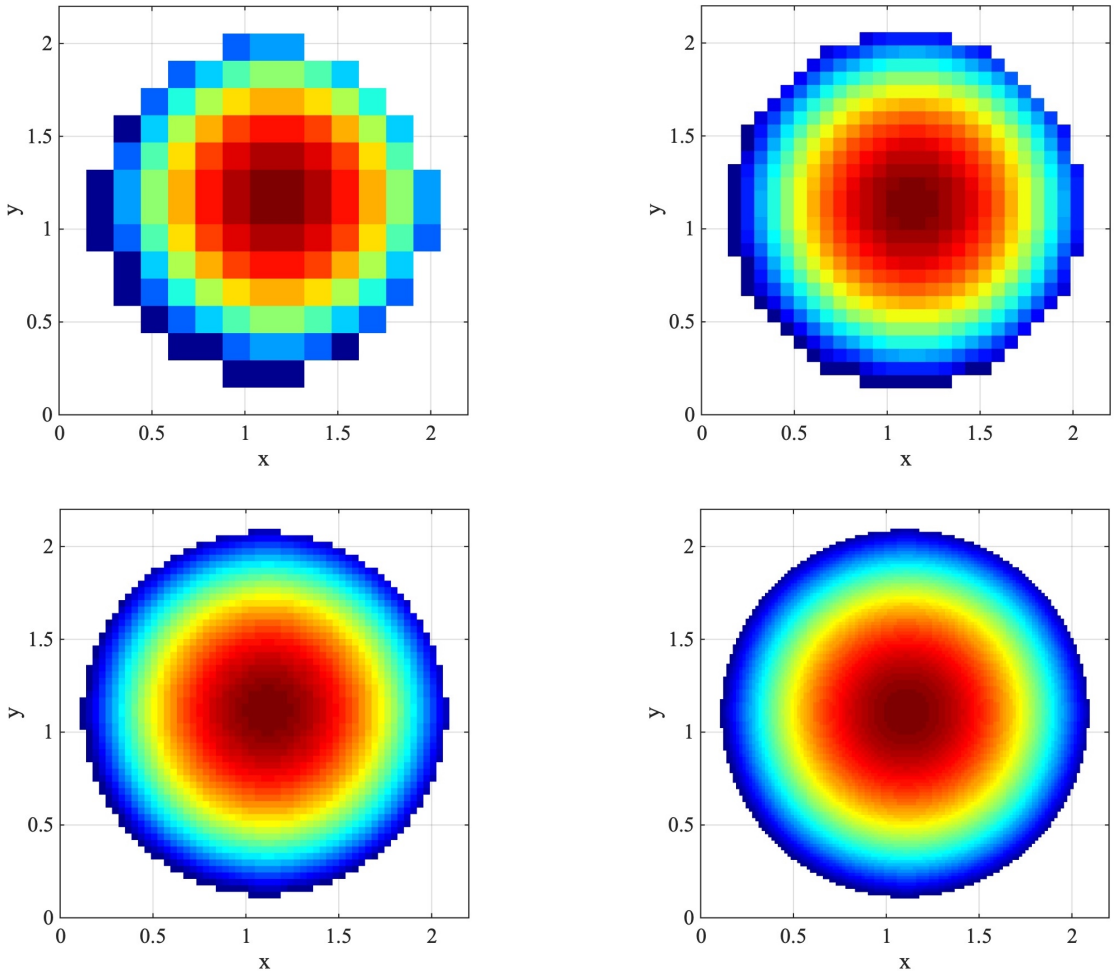


Figure II.11.: Series of numerical solutions with $f = 1$, $g = 0$ on $\Omega = \{x \in \mathbb{R}^2 : \|x - \frac{1}{2}(L, L)^\top\| < 1\}$ with $L = 2.2$ for $N + 2 = \{16, 32, 64, 128\}$ with exact solution $u(x) = (1 - \|x - \frac{1}{2}(L, L)^\top\|^2)/4$ with the error $\epsilon(h) = \|u_h - R_h u\|_{h,\infty} = 10^{-2}(5.4, 2.8, 1.6, 0.8)$ showing the linear order of convergence.

Similar to the maximum principle and M-matrices, by solving $Lw = 1$ with homogeneous Dirichlet boundary conditions we can determine $\left|\left|\left|L_h^{-1}\right|\right|\right|_{h,\infty} = \max_{\Omega_h} w$, which for the approximation of the disc in Fig. II.11 is $\{0.213, 0.232, 0.241, 0.246\}$ and obviously converges to the theoretical bound $\frac{R^2}{2n} = \frac{1}{4}$ from below. Clearly, for the elliptic problem on a disc the stability constant is optimal.

II.6. Boundary Conditions for Elliptic BVPs

II.6.1. Neumann Boundary Conditions

In this section, we want to find an approximation of the solution to the Poisson equation with pure Neumann boundary conditions, that is

$$\begin{cases} -\Delta u = f & \text{in } \Omega \subset \mathbb{R}^n, \\ \frac{\partial u}{\partial \nu} = g & \text{on } \Gamma = \partial\Omega. \end{cases} \tag{NP}$$

First of all we make the simple observation that, should a solution u exist, then also $\bar{u} = u + c$ for $c \in \mathbb{R}$ will be solution. In order to check if a solution exists, integrate (NP) over the domain, i.e.,

$$\int_\Omega f \,\mathrm{d}x = \int_\Omega (-\Delta u)\,\mathrm{d}x \overset{\text{Gauss}}{=} \int_\Gamma -\frac{\partial u}{\partial \nu}\,\mathrm{d}A = -\int_\Gamma g\,\mathrm{d}A, \tag{II.25}$$

which turns out to be a nontrivial condition on the data. This leads us to conclude that solutions of (NP) can only exist (but are nonunique), if the data f, g satisfy the *solvability condition* (II.25). However, even with such a solvability condition, this solution is not unique. Hence the problem is ill-posed. If we want to ensure well-posedness of the problem with Neumann conditions, we might consider the modified problem, where we additionally ask the solution to satisfy

$$C[u] = \int_\Omega u\,\mathrm{d}x = \sigma, \tag{II.26}$$

which removes the ambiguity of $\bar{u} = u + c$ being also a solution (often we set $\sigma = 0$). However, in order to find the corresponding PDE we observe that (NP) can be found by minimizing the functional

$$A[u] = \int_\Omega \left(\tfrac{1}{2}|\nabla u(x)|^2 - f(x)u(x)\right)\,\mathrm{d}x - \int_\Gamma g(x)u(x)\,\mathrm{d}A, \tag{II.27}$$

over all suitable functions u, without imposing conditions on Γ. The computation is similar to the derivation of the minimal surface PDE

$$\begin{aligned} 0 \overset{!}{=} \frac{\mathrm{d}}{\mathrm{d}s} A[u + sv] &= \int_\Omega \nabla u \cdot \nabla v - fv\,\mathrm{d}x - \int_\Gamma gv\,\mathrm{d}A \\ &\overset{\text{Green}}{=} \int_\Omega \big[-\Delta u - f\big]v\,\mathrm{d}x + \int_\Gamma \big[\frac{\partial u}{\partial \nu} - g\big]v\,\mathrm{d}A = 0 \end{aligned}$$

for all functions v. In the second line we used Green's first identity (integration by parts). In order to minimize A subject to the constraint $C[u] = 0$ from (II.26) we introduce $J[u, \lambda] = A[u] + \lambda C[u]$ using the scalar Lagrange multiplier $\lambda \in \mathbb{R}$. The necessary condition to be satisfied is obtained by differentiating with respect to u and λ

and we obtain

$$0 \overset{!}{=} \frac{\mathrm{d}}{\mathrm{d}s} J[u+sv,\lambda] = \int_\Omega \nabla u \cdot \nabla v - fv + \lambda v \,\mathrm{d}x - \int_\Gamma gv \,\mathrm{d}A$$
$$0 \overset{!}{=} \frac{\mathrm{d}}{\mathrm{d}s} J[u,\lambda+s] = \int_\Omega u \,\mathrm{d}x.$$

Again, integrating by parts we find this is equivalent to the modified form of the Poisson problem with Neumann boundary conditions

$$-\Delta u + \lambda = f \quad \text{in } \Omega \subset \mathbb{R}^n, \qquad \frac{\partial u}{\partial \nu} = g \quad \text{on } \Gamma = \partial\Omega, \quad \int_\Omega u \,\mathrm{d}x = \sigma. \tag{II.28}$$

Then there are two alternatives:

1. The data (f, g) satisfies the solvability condition: Then $\lambda = 0$ and u, the *unique* solution of the modified problem, also solves the original problem.

2. The data (f, g) does not satisfy the solvability condition: Then $\lambda \neq 0$ such that u is the *unique* solution of the *modified* problem with data $(\widetilde{f}, g)$, where $\bar{f} = f - \lambda$ satisfies the solvability condition.

Now lets consider the numerical discretization of this problem. As usual, we replace derivatives by difference quotients. This strategy also applies to the boundary conditions. To develop this idea we first consider the 1D problem

$$\begin{cases} -u''(x) = f(x) & \text{in } \Omega = (0,1), \\ u'(0) = g_0,\ u'(1) = g_1 & \text{for } g_0,\, g_1 \in \mathbb{R}. \end{cases} \tag{P1D}$$

Let $h = \frac{1}{N+1}$, $N \in \mathbb{N}$, and $\Omega_h = \{jh \mid 1 \leq j \leq N\}$. The boundary is then $\Gamma_h = \{0, 1\}$. The second order derivative is replaced by the usual 3-point difference stencil

$$-\frac{1}{h^2}\begin{pmatrix} 1 & -2 & 1 \end{pmatrix} u_h = f \qquad \text{in } \Omega_h, \tag{II.29}$$

For the boundary conditions, we can use either central or non-central differences. First we consider the non-central case

$$u'(0) \approx D^+ u_h(0) = \frac{u_h(h) - u_h(0)}{h} = g_0,$$
$$u'(1) \approx D^- u_h(1) = \frac{u_h(1) - u_h(1-h)}{h} = g_1,$$

or equivalently,

$$\frac{1}{h} g_0 = \frac{1}{h^2}(u_h(h) - u_h(0)), \qquad \frac{1}{h} g_1 = \frac{1}{h^2}(u_h(1) - u_h(1-h)). \tag{II.30}$$

Writing all these equations in matrix-vector form yields the extended system

$$-\frac{1}{h^2}\underbrace{\begin{bmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix}}_{\in\mathbb{R}^{(N+2)\times(N+2)}} \underbrace{\begin{bmatrix} u_h(0) \\ u_h(h) \\ \vdots \\ u_h(1-h) \\ u_h(1) \end{bmatrix}}_{\in\mathbb{R}^{N+2}} = \underbrace{\begin{bmatrix} -\frac{1}{h}g_0 \\ f(h) \\ \vdots \\ f(Nh) \\ \frac{1}{h}g_1 \end{bmatrix}}_{\in\mathbb{R}^{N+2}}$$

In order to get the reduced system, we eliminate $u_h(0)$, $u_h(1)$. With (II.30) we obtain

$$u_h(0) = u_h(h) - hg_0, \quad u_h(1) = u_h(1-h) + hg_1.$$

Now we insert the first expression into the 3-point stencil and get

$$\begin{aligned} f(h) &= -\frac{1}{h^2}(u_h(0) - 2u_h(h) + u_h(2h)) \\ &= -\frac{1}{h^2}(u_h(h) - hg_0 - 2u_h(h) + u_h(2h)). \end{aligned}$$

This gives

$$-\frac{1}{h^2}(-u_h(h) + u_h(2h)) = f(h) - \frac{1}{h}g_0.$$

Analogously, we obtain

$$-\frac{1}{h^2}(u_h((N-1)h) - u_h(Nh)) = f(Nh) + \frac{1}{h}g_1.$$

In matrix-vector formulation this results in the linear system of equations

$$\underbrace{-\frac{1}{h^2}\begin{bmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix}}_{=:L_h\in\mathbb{R}^{N\times N}} \underbrace{\begin{bmatrix} u_h(h) \\ \vdots \\ \\ \vdots \\ u_h(Nh) \end{bmatrix}}_{=:u_h\in\mathbb{R}^N} = \underbrace{\begin{bmatrix} f(h) \\ \vdots \\ \\ \vdots \\ f(Nh) \end{bmatrix} + \begin{bmatrix} -\frac{1}{h}g_0 \\ 0 \\ \vdots \\ 0 \\ \frac{1}{h}g_1 \end{bmatrix}}_{=:f_h\in\mathbb{R}^N}$$

We make the following observations:

a) The matrix L_h is symmetric.

b) The matrix L_h is not invertible, since

$$L_h \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

This corresponds to the fact that the pure Neumann boundary conditions problem is ill-posed. In the FDM, the constant functions represented by $\alpha\mathbb{1}$, are eigenfunctions of L_h with eigenvalue zero. Therefore, we get $\{\alpha\mathbb{1} \mid \alpha \in \mathbb{R}\} \subseteq \ker L_h$ and therefore, $\operatorname{rank} L_h \leq N-1$. Moreover, not every vector f_h even admits a solution of $L_h u_h = f_h$, namely if $f_h \notin \operatorname{im} L_h$. By adding all entries in $L_h u_h$, we obtain a necessary solvability condition as

$$h \sum_{n=1}^{N} f(x_n) = -(g_1 - g_0), \tag{II.31}$$

corresponding to (II.25).

However, the discrete solvability condition (II.31) may not be satisfied after discretizing the continuous solvability condition.

Question: How do we attempt solve the linear system $L_h u_h = f_h$, if L_h is not invertible? This problem can be bypassed by extending the matrix L_h and the vectors u_h and f_h in the following way. For $\lambda, \sigma \in \mathbb{R}$ we define

$$\widetilde{L}_h = \begin{bmatrix} L_h & \mathbb{1} \\ \mathbb{1}^{\mathsf{T}} & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}, \quad \widetilde{u}_h = \begin{bmatrix} u_h \\ \lambda \end{bmatrix} \in \mathbb{R}^{N+1}, \quad \widetilde{f}_h = \begin{bmatrix} f_h \\ \sigma \end{bmatrix} \in \mathbb{R}^{N+1}.$$

Here, σ can be chosen arbitrarily (usually $\sigma = 0$). In this construction $\widetilde{L}_h$ is invertible and thus, $\widetilde{L}_h \widetilde{u}_h = \widetilde{f}_h$ has a unique solution. Two cases are possible:

a) $\lambda = 0$: Then it holds $L_h u_h = f_h$ and u_h is the solution which satisfies $\langle \mathbb{1}, u_h \rangle = \sigma$.

b) $\lambda \neq 0$: Then it holds $L_h u_h = f_h - \lambda\mathbb{1}$. Then u_h can be interpreted as an approximation of the problem

$$\begin{cases} -u''(x) = f(x) - \lambda, \\ u'(0) = g_0,\ u'(1) = g_1. \end{cases}$$

The solution u is again normalized by the condition $\langle \mathbb{1}, u_h \rangle = \sigma$. In fact, λ is a correction of the discrete problem, such that the discrete solvability condition (II.31) is satisfied for $f_h - \lambda\mathbb{1}$.

Remark II.26: a) The orders of consistency and convergence w. r. t. $\|\cdot\|_\infty$ are both equal to 1 due to the use of non-central differences at the boundary.

b) It is possible to obtain order 2 if the central differences are used in the following way. Let

$$u'(x) \approx \frac{u(x+h) - u(x-h)}{2h}$$

at the boundary $x \in \Gamma$. Note that $u(-h)$, $u(1+h)$ lie outside the domain Ω. This is "unphysical" for the exact problem, but we can do that for the numerical solution.

By extending $\overline{\Omega}_h$ with the points $\{-h, 1+h\}$, the central differences yield

$$\begin{aligned} u'(0) \approx D^0 u_h(0) &= \frac{u_h(h) - u_h(-h)}{2h} = g_0, \\ u'(1) \approx D^0 u_h(1) &= \frac{u_h(1+h) - u_h(1-h)}{2h} = g_1, \end{aligned}$$

then the new grid points are eliminated

$$\begin{aligned} u_h(-h) &= u_h(h) - 2hg_0, \\ u_h(1+h) &= u_h(1-h) + 2hg_1. \end{aligned}$$

That is,

$$\begin{aligned} f(0) &= \frac{1}{h^2}\left(-u_h(-h) + 2u_h(0) - u_h(h)\right) \\ &= \frac{1}{h^2}\left(+2hg_0 + 2u_h(0) - 2u_h(h)\right), \end{aligned}$$

and

$$\begin{aligned} f(1) &= \frac{1}{h^2}\left(-u_h(1-h) + 2u_h(1) - u_h(1+h)\right) \\ &= \frac{1}{h^2}\left(-2u_h(1-h) + 2u_h(1) - 2hg_1\right), \end{aligned}$$

which we both divide by two. Note that for this we have to extend the function f to the boundary in this case. Overall, this leads to the matrix

$$L_h^c u_h = -\frac{1}{h^2}\begin{bmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix}\begin{bmatrix} u_h(0) \\ u_h(h) \\ \vdots \\ u_h(1-h) \\ u_h(1) \end{bmatrix} = \begin{bmatrix} \frac{1}{2}f(0) - \frac{1}{h}g_0 \\ f(h) \\ \vdots \\ f(1-h) \\ \frac{1}{2}f(1) + \frac{1}{h}g_1 \end{bmatrix}$$

where $L_h^c \in \mathbb{R}^{(N+2)\times(N+2)}$. In the last step we divided the first and last row of L_h^c by two to make the system of equation symmetric. However, this system of equation still has a zero eigenvalue. Interestingly, we obtain a slightly modified solvability condition

$$\frac{h}{2}\big(f(0) + f(1)\big) + h\sum_{n=1}^{N} f(x_n) = -(g_1 - g_0), \tag{II.32}$$

which by midpoint rule is a second-order approximation of the continuous solvability condition (II.25). Again, in order to solve this discrete equation we introduce an additional condition, for instance $\langle \mathbb{1}, u_h \rangle = \sigma$ as discussed above. The scheme will be convergent of order 2 w. r. t. $\|\cdot\|_\infty$ provided the exact solution is sufficiently regular.

Remark II.27 (Extensions to Robin and Higher Dimensions)**:** Apart from the details in the manipulation of the difference stencil near the boundary, all these steps carry over to the higher-dimensional case analogously. It is also clear from the arguments above, that mixed bounary conditions $\alpha u(0) + \beta u'(0) = g_0$ should be treated analogously. For instance, in one spatial dimension a central difference of the form

$$\alpha u_h(0) + \frac{\beta}{2h}(u_h(h) - u_h(-h)) = g(0), \tag{II.33}$$

should be used to obtain a second order accurate scheme, where the degree of freedom at $u_h(-h)$ will be eliminated as illustrated before. In general, as soon as $\alpha > 0$ on any part of the boundary, the resulting operator is invertible without any further modification.

Remark II.28 (Stability with Neumann boundary conditions)**:** Without the modification the problem is ill-posed and in particular $\|\!|L_h^{-1}|\!\| = \infty$. However, after the modification one can show

$$\max_{\Omega_h} |u_h| \leq C_f \max_{\Omega_h} |f| + C_g \max_{\Gamma_h} |g|,$$

e.g. see [Hac92].

II.6.2. Periodic Boundary conditions

For simplicity, we consider the one-dimensional Poisson problem with periodic boundary conditions. The statement is

$$\begin{cases} -u'' = f & \text{in } \Omega = (0,1), \\ u(0) = u(1), \\ u'(0) = u'(1), \end{cases} \tag{PP}$$

which using integration over the domain again yields the solvability condition

$$\int_0^1 f(x)\,\mathrm{d}x = 0. \tag{II.34}$$

Similarly as before we define the modified problem

$$\begin{cases} -u'' + \lambda = f & \text{in } \Omega = (0,1), \\ u(0) = u(1), \\ u'(0) = u'(1), \\ \int_0^1 u(x)\,\mathrm{d}x = 0, \end{cases} \tag{MPP}$$

for $u : \overline{\Omega} \to \mathbb{R}$ and Lagrange multiplier $\lambda \in \mathbb{R}$. If f satisfies the solvability condition, then $\lambda = 0$ and u is the unique solution of the modified problem, or $\lambda \neq 0$ and u is the unique periodic solution of $-u'' = \bar{f}$ with $\bar{f} = f - \lambda$ satisfying the solvability condition.

Concerning the numerical discretization we proceed with the usual compact 3-point stencil

$$\frac{-1}{h^2}\begin{pmatrix}1 & -2 & 1\end{pmatrix} u_h(x_n) = f(x_n) \tag{II.35}$$

for $x_n = nh$ for $n = 0, ..., N$ and $h = 1/(N+1)$. However, we identify $u_{N+1} = u_0$ and $u_N = u_{-1}$ which produces the linear system of (sparse) equations

$$L_h^p u_h = -\frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{bmatrix}\begin{bmatrix} u_h(0) \\ u_h(h) \\ \vdots \\ u_h(1-2h) \\ u_h(1-h) \end{bmatrix} = \begin{bmatrix} f(0) \\ f(h) \\ \vdots \\ f(1-2h) \\ f(1-h) \end{bmatrix},$$

to be solved for $u_h \in \mathbb{R}^{N+1}$ with compatibility condition

$$\sum_{i=0}^{N} f(x_i)h = 0. \tag{II.36}$$

This equation again must be modified by finding the unique solution (u_h, λ) of

$$\widetilde{L}_h = \begin{bmatrix} L_h^p & \mathbb{1} \\ \mathbb{1}^\mathsf{T} & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \quad \widetilde{u}_h = \begin{bmatrix} u_h \\ \lambda \end{bmatrix} \in \mathbb{R}^{N+2}, \quad \widetilde{f}_h = \begin{bmatrix} f_h \\ \sigma \end{bmatrix} \in \mathbb{R}^{N+2}.$$

for some $\sigma \in \mathbb{R}$ (usually $\sigma = 0$). The advantage of periodic boundary conditions is that one can easily implement high-order stencils which can be easily implemented also near the boundary. However, the matrices are not tridiagonal anymore. The same procedure can be generalized to higher dimensions. However, periodic boundary conditions are usually restricted to box-shaped domains. In some cases it can be useful to impose periodic boundary conditions just for a part of the boundary.

II.6.3. Boundary Conditions for Arbitrary Domains

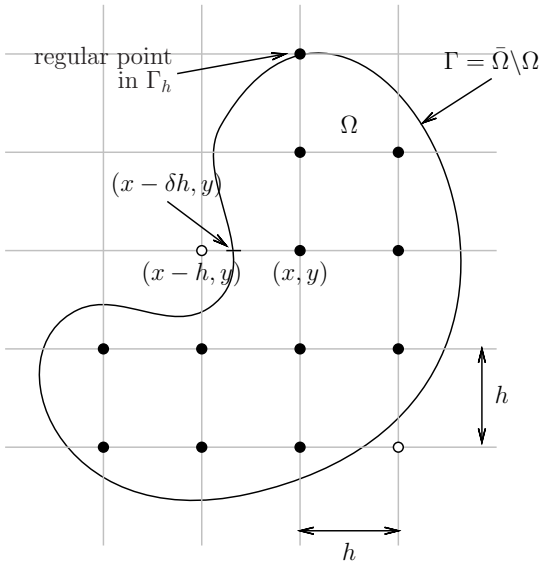
Consider again

$$\begin{cases} -\Delta u = f & \text{in } \Omega \subset \mathbb{R}^2, \\ u = g & \text{on } \partial\Omega = \Gamma. \end{cases}$$

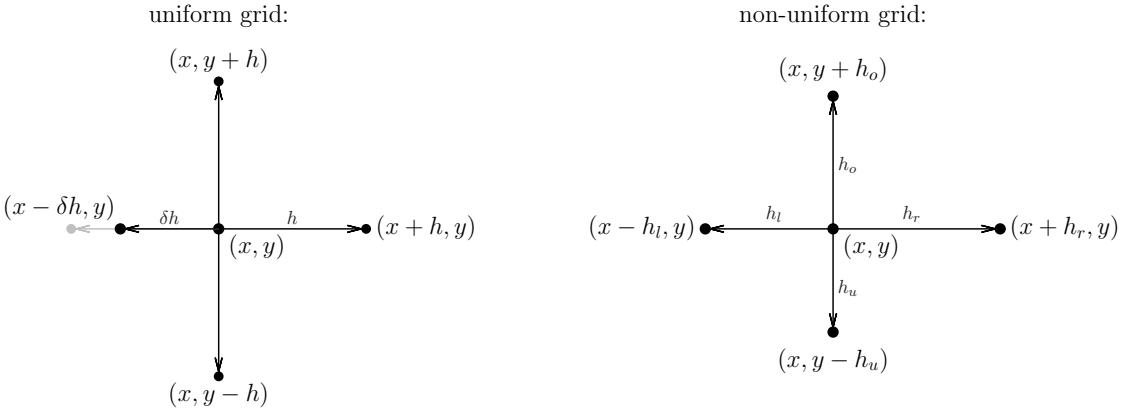
Here, Ω is not necessarily a "nice" domain like the unit square. One way to define an equidistant mesh is

$$\Omega_h = \left\{ (x, y) \in \Omega \;\middle|\; \frac{x}{h}, \frac{y}{h} \in \mathbb{Z} \right\}.$$

Further, we have a set of boundary mesh points Γ_h: If $(x, y) \in \Omega_h$, but $(x - h, y) \notin \Omega$. Then there exists a minimal value for $\delta \in (0, 1]$ such that $(x - \delta h, y) \in \Gamma$. We collect these points in Γ_h (and do the same for the right, lower, and upper neighbors). In the case $\delta = 1$, that is $(x - h, y) \in \Gamma_h$, we say that $(x - h, y)$ is a *regular point on the boundary*. In addition, we say that $(x, y) \in \Omega_h$ is *close to the boundary* if at least one of its neighbors is in Γ_h.



For the five-point difference stencil we use again the center point and the four neighbors. If $(x, y) \in \Omega_h$ is close to the boundary, then the five-point difference stencil becomes non-uniform:



In this situation, the five-point difference stencil for the approximation of the Laplace operator is given by the weights (see also exercise sheet 3)

$$\begin{array}{ccccc} & & \frac{1}{h^2} & & \\ \frac{2}{\delta(1+\delta)h^2} & — & -\left(\frac{2}{\delta h^2}+\frac{2}{h^2}\right) & — & \frac{2}{(1+\delta)h^2}. \\ & & \frac{1}{h^2} & & \end{array}$$

Note that this simplifies to the standard five-point difference stencil if $\delta = 1$ (for regular points on the boundary). Moreover, the signs of the weights change if we want to approximate $-\Delta u$ instead of Δu.

In general, we obtain the "Shortley-Weller difference stencil" for the approximation of the Laplace operator on a general non-uniform grid

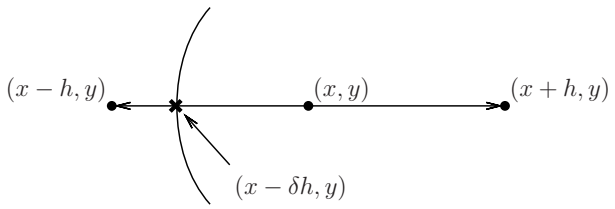
$$\begin{array}{ccccc} & & \frac{2}{h_o(h_u+h_o)} & & \\ & & | & & \\ \frac{2}{h_l(h_l+h_r)} & — & -\left(\frac{2}{h_l h_r}+\frac{2}{h_o h_u}\right) & — & \frac{2}{h_r(h_l+h_r)}. \\ & & | & & \\ & & \frac{2}{h_u(h_u+h_o)} & & \end{array}$$

Remark II.29: a) The derivation relies on the Taylor expansion.

b) The order of consistency is 1 in case the grid is non-uniform everywhere, provided $u \in C^3(\overline{\Omega})$.

c) The resulting matrix L_h is non-symmetric. However, L_h is still an M-matrix and the method is stable.

An alternative is to use an interpolation approach. This yields a symmetric matrix L_h: Let $(x, y) \in \Omega_h$ be close to the boundary such that $(x - h, y) \notin \Omega$ and $(x - \delta h, y) \in \Gamma_h$, for example



By linear interpolation between $u_h(x - \delta h, y)$ and $u_h(x + h, y)$ we get

$$\begin{aligned} u_h(x,y) &= u_h(x-\delta h,y)\cdot\frac{(x+h)-x}{(x+h)-(x-\delta h)} + u_h(x+h,y)\cdot\frac{x-(x-\delta h)}{(x+h)-(x-\delta h)} \\ &= u_h(x-\delta h,y)\frac{1}{1+\delta} + u_h(x+h,y)\frac{\delta}{1+\delta}. \end{aligned}$$

In general, if all neighbors are possibly on the boundary, then with $h_o = \delta_o h$, $h_l = \delta_l h$, $h_r = \delta_r h$, $h_u = \delta_u h$ we obtain

$$\begin{aligned} \frac{1}{h^2}\Bigg(-\frac{u_h(x-\delta_l h,y)}{\delta_l} - \frac{u_h(x+\delta_r h,y)}{\delta_r} + \left(\frac{\delta_l+\delta_r}{\delta_l\delta_r}+\frac{\delta_o+\delta_u}{\delta_o\delta_u}\right)u_h(x,y) \\ -\frac{u_h(x,y+\delta_o h)}{\delta_o} - \frac{u_h(x,y-\delta_u h)}{\delta_u}\Bigg) = 0. \end{aligned}$$

For points which are not close to the boundary we apply the usual five-point difference stencil. After eliminating the boundary points in Γ_h we obtain a symmetric linear system $L_h u_h = f_h$.

Remark II.30: The consistency is the same as above (order 1).

Theorem II.31: Let Ω be a bounded domain. If the exact solution $u \in C^4(\overline{\Omega})$, then the FDM based on the Shortley-Weller difference stencil or on the interpolation approach are convergent of order 2, more precisely it holds

$$\|u_h - R_h u\|_\infty \leq Ch^2 \|u\|_{C^4(\overline{\Omega})},$$

where

$$\|u\|_{C^4(\overline{\Omega})} := \max \left\{ \left\| \frac{\partial^{\nu_1+\nu_2} u}{\partial x^{\nu_1} \partial y^{\nu_2}} \right\|_\infty \;\middle|\; \nu_1 + \nu_2 \leq 4 \right\},$$

where for $v \in C(\overline{\Omega})$,

$$\|v\|_\infty := \max_{(x,y)\in\overline{\Omega}} |v(x,y)|.$$

Proof. See [Hac92]. □

Note that a bit surprisingly the order of convergence is higher than the order of consistency.

II.7. Eigenvalue Problem for Elliptic Operators

In the following we assume L is a general linear, second-order, elliptic operator and we consider the associated eigenvalue problem: Find the set of eigenfunctions $v_k : \overline{\Omega} \to \mathbb{R}$ and corresponding eigenvalues $\lambda_k \in \mathbb{R}$ such that

$$Lv_k = \lambda_k v_k \qquad \text{in } \Omega \tag{II.37}$$

to be supplemented with suitable boundary conditions. Possible choices are homogeneous Dirichlet boundary conditions $u = 0$, homogeneous Neumann boundary conditions $\nu \cdot \nabla u = 0$, homogeneous Robin boundary conditions $\alpha u + \beta \nu \cdot \nabla u = 0$, or periodic boundary conditions. We also allow for combinations of these boundary conditions of different parts of the boundary $\partial\Omega$.

Example II.32 (Laplace operator in 1D)**:** For example, consider $\Omega = (0, 1)$ and the problem

$$-v_k''(x) = \lambda_k v_k(x), \qquad \text{in } (0, 1). \tag{II.38}$$

a) Homogeneous Dirichlet boundary conditions: We have eigenfunctions $v_k(x) = \sin(k\pi x)$ and eigenvalues $\lambda = k^2\pi^2$ for $k = 1, 2, 3, \ldots \in \mathbb{N}$.

b) Homogeneous Neumann boundary conditions: We have eigenfunctions $v_k(x) = \cos(k\pi x)$ and eigenvalues $\lambda = k^2\pi^2$ for $k = 0, 1, 2, \ldots \in \mathbb{N}_0$.

c) Periodic boundary conditions: We have eigenfunctions $v_k^1 = \sin(2k\pi x)$ for $k = 1, 2, 3, \ldots \in N$ and $v_k^2 = \cos(2k\pi x)$ for $k = 0, 1, 2, \ldots \in N_0$ and eigenvalues $\lambda = 4k^2\pi^2$, which for $k \in \mathbb{N}$ have multiplicity two.

d) Mixture of conditions: $u(0) = 0$ and $u'(1) = 0$ gives $v_k(x) = \sin((k - \frac{1}{2})\pi x)$ for $k \in N$.

In particular note that the smallest eigenvalue $\lambda_{\min}$ is different in each of these cases, i.e., we have a) $\lambda_1 = \pi^2$, b) $\lambda_0 = 0$, c) $\lambda_0 = 0$, d) $\lambda_1 = \pi^2/4$.

There are a couple of reasons why the elliptic eigenvalue problem (EVP) are particularly interesting. Assume we want to solve the parabolic initial boundary value problem

$$\partial_t u + Lu = 0, \qquad \text{in } Q_T = (0, \infty) \times \Omega, u(t = 0, \cdot) = u_0, \qquad \text{in } \Omega, \tag{II.39}$$

with one of the above mentioned homogeneous boundary conditions and we can decompose the initial data in terms of the eigenfunctions of L, i.e.,

$$u_0(x) = \sum_k a_k v_k(x), \tag{II.40}$$

then we find the solution

$$u(t, x) = \sum_k a_k v_k(x) \exp(-\lambda_k t). \tag{II.41}$$

For inhomogeneous boundary conditions we simply find

$$u(t,x) = u_{\text{hom}}(x) + \sum_k \bar{a}_k v_k(x) \exp(-\lambda_k t), \qquad u_0 - u_{\text{hom}} = \sum_k \bar{a}_k v_k(x), \tag{II.42}$$

where we solve the homogeneous elliptic problem $Lu_{\text{hom}} = 0$ with inhomogeneous boundary conditions.

These arguments directly carry over to the discrete eigenvalue problem $L_h v_h^k = \lambda_h^k v_h^k$ for $L_h \in \mathbb{R}^{N\times N}$, where we focus on the symmetric reduced problem. Using the symmetry of L_h with respect to the scalar product defined as

$$(u_h, v_h)_{2,h} = h \sum_{i=1}^{N} u_h(x_i) v_h(x_i), \tag{II.43}$$

which is associated to the discrete L_2 norm via $\|u_h\|_{2,h} = (u_h, u_h)_{2,h}$ and using the abbreviation $\|\cdot\| = \|\cdot\|_{2,h}$ we attempt to verify the stability

$$\left\|\left|L_h^{-1}\right|\right\|^2 = \sup_{v\in\mathbb{R}^N, \|v\|=1} \|L_h^{-1} v\|^2.$$

We use that any vector v has a representation $v = \sum a_k v_h^k$ for some coefficients $a_i \in \mathbb{R}$ and that, due to the symmetry of L_h, we have $(v_h^k, v_h^l) = \delta_{k,l}$ and $\|v_h^k\| = 1$ for all k, l. Then the normalization condition becomes $\|v\|^2 = (v,v)_{2,h} = \sum_k a_k^2 = 1$. Using $L_h^{-1} v = \sum a_k/\lambda_h^k v_h^k$ we get

$$\left\|\left|L_h^{-1}\right|\right\|^2 = \sup_{\{a_k\}} \left(\sum_{k=1}^{N} \frac{a_k^2}{\lambda_k^2} \right) = \frac{1}{(\min_k |\lambda_k|)^2}.$$

which we get by using $a_k = \delta_{ki}$ for i corresponding to the smallest (in absolute value) eigenvalue. In this sense the smallest eigenvalues provided in Example II.32 provide stability/continuity constants for the elliptic boundary value problem. In case b,c) we have $\left\|\left|L_h^{-1}\right|\right\| = \infty$. However, we discussed in the numerical discretization we need to consider a modified problem, which also results in a generalized eigenvalue problem of the form

$$\begin{bmatrix} L_h & \mathbb{1} \\ \mathbb{1}^\top & 0 \end{bmatrix} \bar{v}_h^k = \lambda_k \begin{bmatrix} \mathbb{I}_N & 0 \\ 0 & 0 \end{bmatrix} \bar{v}_h^k \tag{II.44}$$

where the last component ℓ^k in $\bar{v}_h^k = (v_h^k, \ell^k) \in \mathbb{R}^{N+1}$ is the multiplier of the extended problem. Note that this is, in principle, a very simple method to show stability of the operator L_h. But explicit information about the eigenvalues of the discrete operator will only be available in a few simple cases, i.e., for simple hyperrectangular domain shapes. In this respect, the stability bound in the max-norm is much more flexible, since it works in any dimension and for arbitrary domain shapes. Our proof was restricted to Dirichlet boundary conditions. Next, we compute the eigenvalues explicitly in one spatial dimension.

Example II.33 (Eigenvalues of the discrete Laplace operator)**:** We now consider the eigenvalue problem for the discrete Laplace operator in 1D. Let $\overline{\Omega}_h = \{0, h, \dots, 1\}$ with $h = 1/(N+1)$ as usual and

$$-\frac{1}{h^2}(\ 1, -2,\ 1)v_k = \lambda_k v_k \qquad \text{in } \Omega_h \tag{II.45}$$

and $u_h(0) = u_h(1) = 0$. We make the ansatz $w(x_n) = e^{i\mu nh}$ and plug this into the finite difference quotient

$$\begin{aligned} L_h w(x_n) &= -\frac{1}{h^2}(e^{i\mu h(n-1)} + e^{i\mu h(n+1)} - 2e^{i\mu hn}) = \frac{-1}{h^2}(e^{-i\mu} + e^{+i\mu} - 2)e^{i\mu hn} \\ &= \frac{1}{h^2}(2 - (e^{-ih\mu} + e^{+hi\mu}))w(x_n) = \lambda w(x_n) \end{aligned}$$

where $\lambda = h^{-2}(2 - (e^{-ih\mu} + e^{+ih\mu}) = h^{-2}(2 - 2\cos(h\mu)) = 4h^{-2}\sin^2(\frac{h\mu}{2})$. In order to satisfy the boundary conditions we use the imaginary part of w and find $\mu = \pi k$ for $k \in \mathbb{N}$. Hence we found the solution of the discrete eigenvalue

$$v_k(x_n) = \sin(\pi k n h), \qquad \lambda_k = 4h^{-2}\sin^2(\tfrac{h\pi k}{2}). \tag{II.46}$$

We can expand the eigenvalues for $h \to 0$ and fixed k as

$$\lambda_k = 4h^{-2}\sin^2(\tfrac{h\pi k}{2}) = \pi^2 k^2 + \mathcal{O}(h^2), \tag{II.47}$$

which coincides with the exact eigenvalues to leading order as $h \to 0$.

Example II.34 (Eigenvalues on the disc)**:** As a nontrivial example consider the following elliptic eigenvalue problem. Let $\Omega = \{x \in \mathbb{R}^2 : \|x\| < 1\}$ and find $v : \overline{\Omega} \to \mathbb{R}$ and $\lambda \in \mathbb{R}$ such that

$$-\Delta v = \lambda v, \qquad \text{in } \Omega, \tag{II.48}$$

$$v = 0, \qquad \text{on } \partial\Omega. \tag{II.49}$$

As before we make a separation ansatz $v = R(r)F(\varphi)$ in polar coordinates and use the corresponding representation of the Laplacian. We obtain

$$-(R''F + r^{-1}R'F + r^{-2}RF'') = \lambda RF, \tag{II.50}$$

or equivalently

$$\frac{r^2(R'' + r^{-1}R' - \lambda R)}{R} = -\frac{F''}{F} = \sigma \in \mathbb{R}, \tag{II.51}$$

where $\sigma \geq 0$ such that $F(\varphi) = a_n\cos(n\varphi) + b_n\sin(n\varphi)$ for $n \in \mathbb{N}_0$ and $\sigma = n^2$.

$$r^2R'' + rR' + (r^2\lambda - n^2)R = 0 \tag{II.52}$$

Transforming $R(r) = a(\xi)$ with $\xi = \sqrt{\lambda} r$ we get

$$\xi^2 a'' + \xi a' + (\xi^2 - n^2) a = 0, \tag{II.53}$$

which is Bessel's differential equation. Since we want the solution to be finite at the origin $\xi = 0$, we have $a(\xi) = J_n(\xi)$ where J_n are Bessel functions of the first kind as shown in Fig. II.12. The first few eigenfunctions for the eigenvalue problem are shown in Fig. II.13. The Bessel functions K_n of second kind, which in principle also solve Bessel's differential equation, have been neglected because they are singular at the origin. In order to satisfy the homogeneous Dirichlet boundary conditions we need to satisfy $R(1) = a(\sqrt{\lambda}) = 0$, which implies that $\sqrt{\lambda}$ corresponds to the zeros of the Bessel function J_n. The kth zero of the Bessel function J_n is denoted by $K_{n,k}$ and the first few zeros are shown in Table II.2.

Note that the multiplicity of each eigenvalue for J_0 is one, while all other eigenvalues for $n \geq 1$ appear twice with $F(\varphi) = \sin n\varphi$ and $F(\varphi) = \cos n\varphi$. The corresponding eigenfunctions are orthogonal since $J_n(r\sqrt{\lambda}) \int_0^{2\pi} \sin n\varphi \cos n\varphi \, \mathrm{d}\varphi = 0$. We do not discuss normalization of the eigenfunctions.

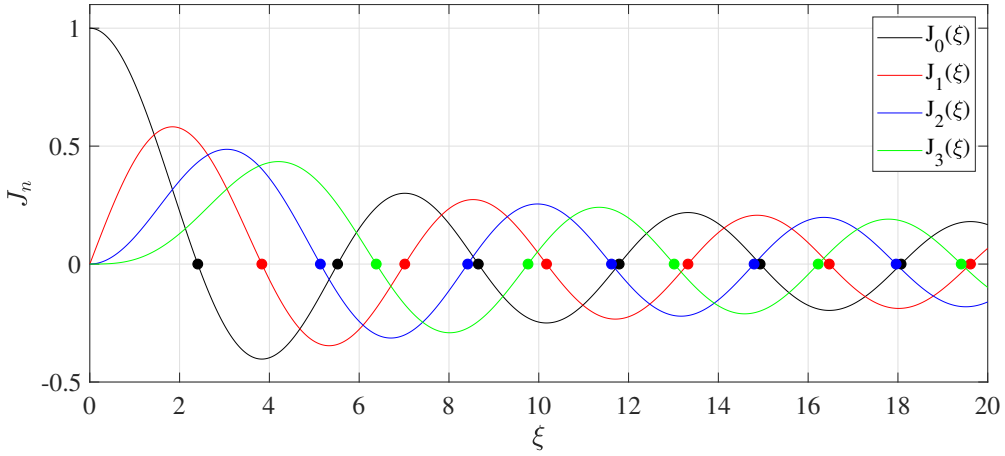


Figure II.12.: Bessel function of first kind $J_n(\xi)$ for $n = 0, \ldots, 3$ and its zeros.

Bessel function	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
J_0	2.40483	5.52008	8.65373	11.79153	14.93092
J_1	3.83171	7.01559	10.17347	13.32369	16.47063
J_2	5.13562	8.41724	11.61984	14.79595	17.95982
J_3	6.38016	9.76102	13.01520	16.22347	19.40942

Table II.2.: k-th zeros $K_{n,k}$ of the Bessel function J_n.

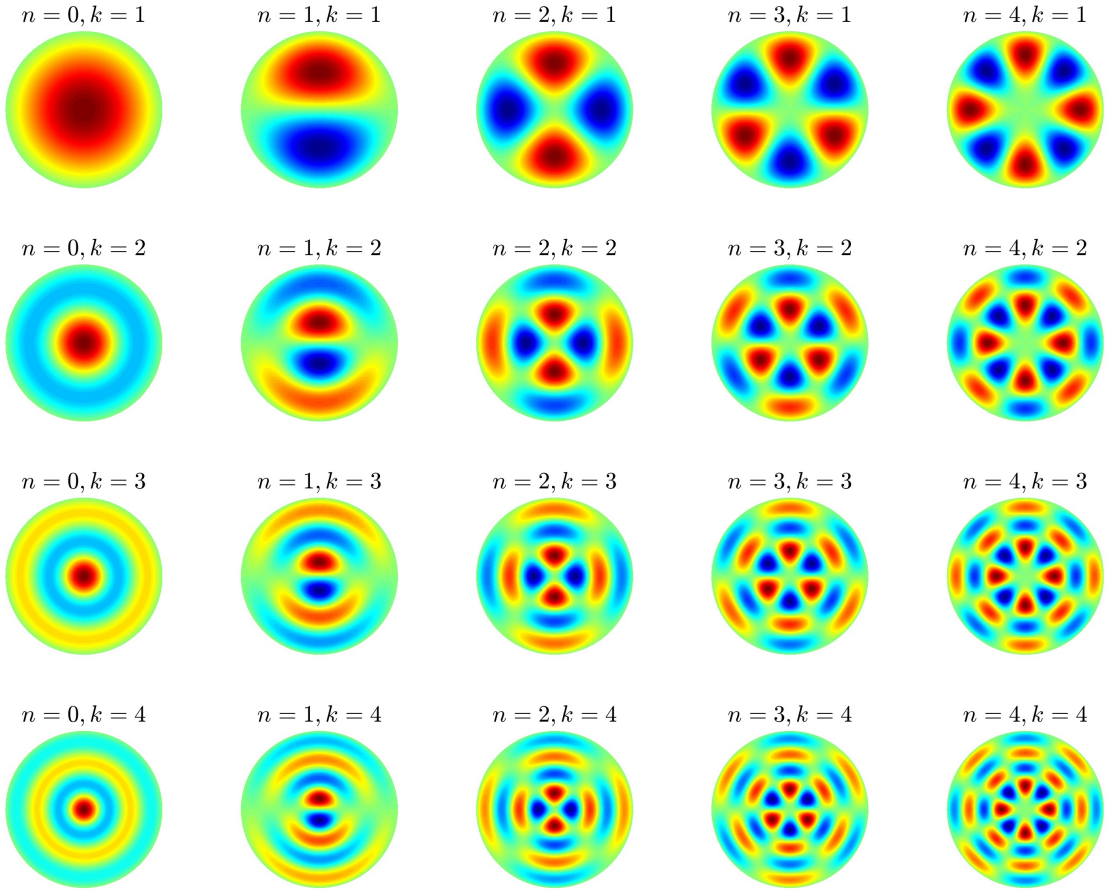


Figure II.13.: Eigenfunctions $v = \cos(n\varphi)J_n(rK_{n,k})$ of Laplace operator where $K_{n,k} = \sqrt{\lambda}$ is the k-th zero of the Bessel function J_n (Note: multiplicity).

These eigenfunctions can also be solved to solve the hyperbolic wave equation, where the eigenfunctions then give rise to the vibrations of a circular drum, see for instance https://en.wikipedia.org/wiki/Vibrations_of_a_circular_membrane. For this, it is straightforward to verify that

$$u(t, x, y) = (A\cos(c\lambda t) + B\sin(c\lambda t))v(x, y) \tag{II.54}$$

solves the wave equation

$$\frac{\partial^2 u}{\partial t^2} + c^2 Lu = 0 \qquad \text{in } Q_T = (0, T) \times \Omega, \tag{II.55}$$

for (λ, v) eigenvalue and eigenfunction of the Laplace operator $Lu = -\Delta u$ on the disc Ω with homogeneous Dirichlet boundary conditions $u(t, x, y) = 0$ for $(x, y) \in \partial\Omega$. If we expand the given initial data in terms of eigenfunctions, then the superposition of eigenmodes in (II.54) gives the general solution of the homogeneous wave equation. This is related to the interesting question: "Can one hear the shape of a drum?", see also https://en.wikipedia.org/wiki/Hearing_the_shape_of_a_drum.

II.8. Finite Differences for Parabolic IBVP

II.8.1. Introduction

In the following let $u : Q_T = (0,T) \times \Omega \to \mathbb{R}$ a scalar function and $\Omega \subset \mathbb{R}^n$. As usual, the function $u(t,x)$ depends on time and space. Let L be an elliptic operator acting on the spatial part of u, then we can write a parabolic PDE equivalently to our previous classification as

$$\partial_t u + Lu = f, \qquad \text{in } Q_T \tag{II.56}$$

$$u(t=0,\cdot) = u_0, \qquad \text{in } \Omega \tag{II.57}$$

using initial data $u_0 : \Omega \to \mathbb{R}$. In accordance with our previous considerations, the operator L can be supplied with Dirichlet, Neumann, Robin or periodic boundary conditions on the boundary $\partial\Omega$. The main difference in the treatment of elliptic and parabolic problems can be explained using the concept of the domain of dependence as shown in Fig. II.14.

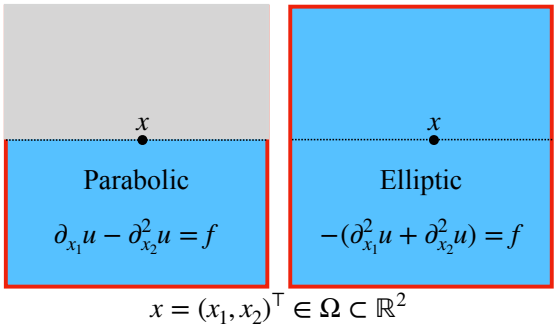


Figure II.14.: **Domain of dependence:** The solution $u(x)$ at a point $x = (x_1, x_2)^\top$ depends on the data provided on the boundary marked in red. While for elliptic problems we need to specify conditions on the entire boundary $\partial\Omega$, for a parabolic problem we only need to know the initial data at $t = 0$ and the boundary conditions in x_2 for previous times $0 < t < x_1$.

This allows us to propose the following strategy: Similar as in the previous sections, we discretize the spatial part PDE L, but we will also discretize the time derivative using finite differences. As opposed to elliptic problems, we are not going (to try) to solve the problem on the whole domain Q_T or a corresponding discrete version of it. Instead, we integrate in time: This means that starting from $u_h(0,x)$ we advance the grid function for a small time-increment $\tau = T/M$ and obtain $u_h(\tau,x)$ by solving a linear system of equations and then iterate this process to obtain $u_h(0,x) \to u_h(\tau,x) \to \ldots \to u_h(T = M\tau, x)$ as a sequence of sparse linear problems. In the following we will propose strategies and show their convergence to the solution of the parabolic PDE.

II.8.2. Time-discretization

Let $\Omega = (0, L)$ find $u : (0, T) \times \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{aligned} \partial_t u(t,x) - a\partial_x^2 u(t,x) &= f(t,x), && \text{for } t = (0,T), x \in \Omega, \\ u(t=0,x) &= u_0(x), && \text{for } x \in \Omega \\ u(t,0) &= \alpha(t), && \text{for } t = (0,T), \\ u(t,L) &= \beta(t), && \text{for } t = (0,T), \end{aligned}$$

where we again use a uniform spatial mesh $\overline{\Omega}_h = \{0, h, \dots, (N+1)h = L\}$ with grid spacing $h = L/(N+1)$. As before, we denote $x_n = nh$. We also introduce a constant time-discretization $\tau = T/M$ and $t^k = k\tau$ for $k \in \mathbb{N}_0$. As before we seek to approximate the exact solution $u(t^k, x_n)$, which we will denote by u_n^k (for brevity we omit an extra index h or τ). As before we will discretize the second spatial derivatives using

$$D_x^+ D_x^- u_n^k = \frac{1}{h^2}(u_{n-1}^k - 2u_n^k + u_{n+1}^k) \tag{II.58}$$

and first time-derivatives using

$$D_t^+ u_n^k = \frac{1}{\tau}(u_n^{k+1} - u_n^k). \tag{II.59}$$

Definition II.35 (Theta scheme)**:** For a real-parameter $0 \le \theta \le 1$ this allows to define the following discretization of the parabolic problem

$$D_t^+ u_n^k + \left[\theta L_h u_n^{k+1} + (1-\theta) L_h u_n^k\right] = \bar{f}_n^k,$$

for $k = 1 \dots M$ and $n = 1 \dots N$ and a suitable approximation $\bar{f}_n^k$. We call this the θ-scheme. For $k = 0$ we replace $u_n^0 = u_0(x_n)$ and for the spatial part we include the boundary conditions as discussed for the elliptic problem. Specifically with $L_h = -D^+D^-$, the θ-scheme has three well-known special cases:

i) **Explicit/Forward Euler scheme** $\theta = 0$:

$$u_n^{k+1} = (1 - 2\gamma)u_n^k + \gamma(u_{n+1}^k + u_{n-1}^k) + \tau f(t^k, x_n) \tag{II.60}$$

ii) **Implicit/Backward Euler scheme** $\theta = 1$:

$$(1 + 2\gamma)u_n^{k+1} - \gamma(u_{n+1}^{k+1} + u_{n-1}^{k+1}) = u_n^k + \tau f(t^{k+1}, x_n) \tag{II.61}$$

iii) **Crank-Nicolson scheme** $\theta = 1/2$:

$$(1 + \gamma)u_n^{k+1} - \tfrac{\gamma}{2}(u_{n+1}^{k+1} + u_{n-1}^{k+1}) = (1 - \gamma)u_n^k + \tfrac{\gamma}{2}(u_{n+1}^k + u_{n-1}^k) + \tau \bar{f}_n^k \tag{II.62}$$

where $\bar{f}_n^k = f(t^k + \tau/2, x_n)$.

In all three cases we used the abbreviation $\gamma = \tau/h^2$. All of these methods work similarly well in higher spatial dimensions and with general elliptic operators L_h.

Only the explicit Euler scheme can be readily solved without solving a system of equations, whereas the implicit Euler and the Crank-Nicolson scheme lead to tridiagonal systems of sparse equations. These schemes differ in the numerical effort to solve, the precision/consistency error, and the stability.

II.8.3. Convergence of solutions

Let us begin investigating the convergence of the θ-scheme in the maximum norm using two useful estimates for $M \sim L_h$, where L_h is the discretization of the elliptic operator.

Theorem II.36 (Useful Estimates)**:** Let $M \in \mathbb{R}^{N\times N}$ weakly diagonal dominant with $M_{ii} > 0$ and $M_{ij} \leq 0$ for $j \neq i$ and $\sum_j M_{ij} \geq 0$ for all $i, j = 1, ..., N$.

a) Then we have the estimate $\|v\|_\infty \leq \|(\mathbb{I} + M)v\|_\infty$.

b) Furthermore assume $0 < M_{ii} \leq 1$. Then we have $\|(\mathbb{I} - M)v\|_\infty \leq \|v\|_\infty$.

Proof. a) Let $k = 1, ..., N$ the index such that $\|v\|_\infty = |v_k| \neq 0$.

$$\begin{aligned}\|(\mathbb{I} + M)v\|_\infty \geq |(\mathbb{I} + M)v|_k = |\sum_j (\mathbb{I} + M)_{kj} v_j| = |1 + \sum_j M_{kj}\frac{v_j}{v_k}| \cdot |v_k| \\ \geq |1 + M_{kk} + \sum_{j\neq k} M_{jk}| \cdot |v_k| \geq |v_k| = \|v\|_\infty.\end{aligned}$$

b) $\|(\mathbb{I} - M)v\|_\infty \leq \max_i \left[(1 - M_{ii})\|v\|_\infty + \sum_{j\neq i}(-M_{ij})\|v\|_\infty\right] \leq \|v\|_\infty$. □

Theorem II.37 (Consistency of θ-scheme)**:** The θ-scheme defined in II.35 has the following consistency error in the maximum norm:

a) assuming $u \in C^{2,4}(\bar{Q}_T)$ we have $\mathcal{O}(\tau + h^2)$,

b) assuming $u \in C^{3,4}(\bar{Q}_T)$ we have $\mathcal{O}(\tau^2 + h^2)$,

where $C^{\alpha,\beta}(Q_T)$ denotes functions, which are α-times in time and β-times in space continuously differentiable.

Proof. The proof works as usual using Taylor's theorem. Let $u_n^k = u(t^k, x_n)$ with u solving the exact problem $\partial_t u + Lu = f$ and u_h^k the corresponding vector of spatial components at time t^k. Then

$$u_h^{k+1} = u_h^k + R_h\left[\dot{u}\tau + \ddot{u}\tfrac{\tau^2}{2}\right] + \mathcal{O}(\tau^3), \quad f(t + \tfrac{\tau}{2}, x) = f(t, x) + \left[\dot{f}(t, x)\right]\tfrac{\tau}{2} + \mathcal{O}(\tau^2),$$

where the dot indicates the time-derivative. Inserting this into the discretization gives

$$\frac{u_h^{k+1} - u_h^k}{\tau} = R_h\left(\dot{u} + \ddot{u}\frac{\tau}{2}\right) + \mathcal{O}(\tau^2), \qquad L_h u_h^{k+1} = L_h R_h(u + \tau\dot{u}) + \mathcal{O}(\tau^2),$$

so that using consistency of the elliptic operator such that $\|w\|_{h,\infty} = \mathcal{O}(h^2)$ with $w = L_h R_h u - f = L_h R_h u - R_h L u$ for fixed time we get

$$\begin{aligned}0 &= \frac{u_h^{k+1} - u_h^k}{\tau} + \theta L_h u_h^{k+1} + (1 - \theta)L_h u_h^{k+1} - \bar{f}_n^k \\ &= R_h\left(\dot{u} + Lu - f + \frac{\tau}{2}\left[\ddot{u} + 2\theta L\dot{u} - \dot{f}\right]\right) + \mathcal{O}(\tau^2 + h^2).\end{aligned}$$

The second bracket vanishes for $\theta = 1/2$, because then the term inside is the time-derivative of the parabolic PDE. Otherwise the consistency is $\mathcal{O}(\tau + h^2)$. □

In the following we will consider the stability of the θ-scheme, where the approach differs slightly depending on the considered norm. We start considering the discrete maximum norm. We can rewrite the θ-scheme as

$$\begin{aligned} &-\gamma\theta(u_{n-1}^{k+1}+u_{n+1}^{k+1})+(2\theta\gamma+1)u_n^{k+1}=F_n^k,\\ &\quad\text{where}\quad F_n^k=(1-\theta)\gamma(u_{n-1}^k+u_{n+1}^k)+(1-2(1-\theta)\gamma)u_n^k+\tau\bar{f}_n^k. \end{aligned}$$

The diagonal dominance of the first equation implies

$$\max_n|u_n^{k+1}|\le\max_n|F_n^k|,$$

where we used Theorem II.36. Using $0\le\theta\le 1$ and $(1-2r)\ge 0$ with $r=(1-\theta)\gamma$ gives

$$\begin{aligned} \max_n|u_n^{k+1}|&\le\max_n|F_n^k|\le 2r\max_n|u_n^k|+(1-2r)\max_n|u_n^k|+\tau\max|\bar{f}_n^k|\\ &=\max_n|u_n^k|+\tau\max|\bar{f}_n^k|. \end{aligned}$$

Iterating this argument over time-step produces the equivalent of the stability for time-dependent problem

$$\max_{k,n}|u_n^{k+1}|=\max_n|u_0(x_n)|+\tau\sum_{k=1}^{M}\max_n|\bar{f}_n^k|. \tag{II.63}$$

This argument can be extended to a wider class of discrete operators L_h which are diagonally dominant, but will suffice for the moment for $L_h=-D^+D^-$. Interestingly, the restriction $(1-2r)\ge 0$ is translated into a restriction of the time-step size

$$(1-\theta)\tau\le\tfrac{1}{2}h^2, \tag{II.64}$$

very much in the spirit of the Courant–Friedrichs–Lewy (CFL) condition for hyperbolic PDEs. The condition is only trivially satisfied for the implicit Euler method $\theta=1$. This analysis allows us to make the following statement about convergence of solutions.

Theorem II.38 (Convergence of solutions)**:** Let $(1-\theta)\tau\le\frac{1}{2}h^2$ and $u\in C^{2,4}(\bar{Q}_T)$ and $\bar{f}_n^k=f(t^k,x_n)$. Then we have

$$\max_{k,n}|u_n^k-u(t^k,x_n)|=\mathcal{O}(h^2+\tau). \tag{II.65}$$

For the Crank-Nicolson scheme with $\tau\le h^2$ we have

$$\max_{k,n}|u_n^k-u(t^k,x_n)|=\mathcal{O}(h^2+\tau^2). \tag{II.66}$$

given that $u\in C^{3,4}(\bar{Q}_T)$.

Proof. Let $w_n^k=u_n^k-u(t^k,x_n)$ solving the discrete problem with zero initial and boundary data and the given consistency error ε. Then

$$\max_{k,n}|w_n^k|=\underbrace{\max_n|w_n^0|}_{0}+\tau\sum_{k=1}^{M}\max_n|\varepsilon|\le T|\varepsilon|. \qquad \square$$

II.8.4. Generalization of θ-scheme

Lets consider the θ-scheme with a general discrete elliptic operator L_h. We have

$$[\mathbb{I} + \theta\tau L_h]u_n^{k+1} = F_n^k = \tau \bar{f}_n^k + [\mathbb{I} - \tau(1-\theta)L_h]u_n^k,$$

Theorem II.39: If L_h is weakly diagonal domiant, then the discrete parabolic IVP is stable in the max-norm $\|\|\|_\infty$.

Proof. Use first estimate in Theorem II.36 with $M = \theta\tau L_h$ and then the second estimate using the CFL-type condition $0 \leq \tau(1-\theta)(L_h)_{ii} \leq 1$ to get stability in the max-norm. □

Remark II.40: In particular when solving higher-order parabolic PDEs, e.g., the Cahn-Hilliard equation is a fourth-order parabolic equation, then $(L_h)_{ii} \sim h^{-4}$ in the stability condition of the explicit scheme leads to severe restriction for the time-step size as $h \to 0$.

Example II.41 (Stability of parabolic problem with constant coefficients)**:** Consider the following discretizations of the problem with constant coefficients with $a > 0$ and $c \geq 0$. Furthermore we assume $b > 0$. Then

i) $L_h = \frac{-a}{h^2}(1,-2,1) + \frac{b}{2h}(-1,0,1) + c(0,1,0)$ is weakly diagonally dominant if $bh < 2a$ and $(L_h)_{ii} = 2a/h^2 + c$.

ii) $L_h = \frac{-a}{h^2}(1,-2,1) + \frac{b}{h}(-1,1,0) + c(0,1,0)$ is always weakly diagonally dominant and $(L_h)_{ii} = 2a/h^2 + c + \frac{b}{h}$.

iii) $L_h = \frac{-a}{h^2}(1,-2,1) + \frac{b}{h}(0,-1,1) + c(0,1,0)$ is not weakly diagonally dominant.

With the other sign for b the role of ii) and iii) exchange.

Example II.42 (General Laplace operator)**:** The general Laplace operator with the standard 2n+1−stencil is weakly diagonally dominant and $(L_h)_{ii} = 2nh^{-2}$.

Example II.43 (Stability of Implicit Euler scheme)**:** The implicit Euler scheme reads

$$(\mathbb{I} + \tau L_h)u^{k+1} = \tau f^k + u^k \tag{II.67}$$

where we assume L_h has positive eigenvalues. Then the eigenvalues of $\mathbb{I} + \tau L_h$ are larger than one and $\left\|\left|(\mathbb{I} + \tau L_h)^{-1}\right|\right\| \leq 1$. Hence we obtain a stability inequality

$$\|u^{k+1}\| \leq \left\|\left|(\mathbb{I} + \tau L_h)^{-1}\right|\right\| (\tau\|\bar{f}^k\| + \|u^k\|) \leq \tau\|\bar{f}^k\| + \|u^k\| \tag{II.68}$$

where we used the triangle inequality. By iterating this condition we can again obtain an estimate of $\|u^k\|$ in terms of the data in an appropriate norm with respect to time.

II.8.5. Stability in the discrete L_2 norm

In the following we study the stability of the θ-scheme in the discrete L_2 norm. Our method is closely related to the von Neumann stability analysis. For this investigation we consider the homogeneous one-dimensional problem on $\Omega = (0,1)$ with homogeneous Dirichlet boundary conditions. We already observed that the elliptic Poisson operator has discrete eigenfunctions $v_m(x_n) = \sin(\pi m x_n)$ with $x_n = nh$ as usual. Since $L_h = -D^+D^-$ is symmetric, the eigenvalues are real and eigenfunctions form an orthnormal basis. Hence, we can expand the discrete solution in terms of eigenfunctions using

$$u_h(t^k, x_n) = u_n^k = \sum_m \omega_m^k v_m(x_n), \tag{II.69}$$

which plugging into the discretized system and using the fact that v_m is an eigenfunction gives

$$\frac{\omega_m^{k+1} - \omega_m^k}{\tau} = -\lambda_m(\theta\omega_m^{k+1} + (1-\theta)\omega_n^k), \tag{II.70}$$

for all eigenfunctions m. The initial coefficients can be obtained via $\omega_m^0 = (u_0, v_m)_h$. We can write (II.70) in the form

$$\omega_m^{k+1} = q(\tau\lambda_m)\omega_m^k, \qquad \text{where} \qquad q(s) = \frac{1-(1-\theta)s}{1+\theta s} \tag{II.71}$$

with the special cases

$$q(s) = 1 - s, \qquad \text{explicit Euler } \theta = 0, \tag{II.72}$$

$$q(s) = \frac{1}{1+s}, \qquad \text{implicit Euler } \theta = 1, \tag{II.73}$$

$$q(s) = \frac{1-\frac{1}{2}s}{1+\frac{1}{2}s}, \qquad \text{Crank-Nicolson } \theta = \tfrac{1}{2}. \tag{II.74}$$

If we want the error of the method to be bounded, a sufficient condition would be $|q(\tau\lambda_m)| \leq 1$ for all eigenvalues λ_m. This would allow us to conclude $\|u^k\|_2 \leq \|u_0\|_2$ for all $k \geq 0$ in the discrete L_2 norm. From our previous considerations we know $\lambda_m = 4h^{-2}\sin(h\pi m/2)$, which gives for the explicit Euler method $2\tau \leq h^2$, the implicit Euler and the Crank-Nicolson scheme are unconditionally stable.

Remark II.44: While the von Neumann stability analysis is remarkably simple, it is restricted to linear problems and simple domains, where we can obtain explicit expressions for the eigenfunctions. In this sense the stability based on diagonal dominance is much more robust and applicable to a wider range of problems and geometries.

II.9. Concluding Remarks

Let us briefly summarize the results of this chapter:

a) The general idea of the finite difference method is to replace derivatives by difference quotients and functions by grid functions. Thus, the derivation of the schemes themselves is usually straight forward.

b) The PDE will be converted into a high dimensional $\sim \mathbb{R}^{N^n}$ system of linear equations $L_h u_h = f_h$.

c) We have discussed the FDM for elliptic BVPs and parabolic IBVP in 1D and 2D with Dirichlet and Neumann boundary conditions.

d) We have conducted an error analysis and discussed concepts such as stability (using M-matrices, discrete maximum principle), consistency (using Taylor expansions), and convergence (by using consistency and stability).

On the other hand, the FDM has quite some short-comings:

a) Convergence proofs require strong assumption for the exact solution, e.g. $u \in C^4(\overline{\Omega})$, which is often not true even in simple cases.

Moreover, we have seen that the inviscid Burgers equation

$$u_t + \partial_x(u^2) = 0 \quad \text{for } (t,x) \in (0,T) \times (0,1)$$

even admits solutions that are discontinuous in the interior of the domain.

b) For Neumann boundary conditions and for general non-box-shaped domains $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$), the order of convergence can be reduced.

c) It is not so easy to adjust/refine the mesh locally. This would lead to a non-uniform mesh and thus potentially a reduction of the order of convergence.

So we would like to have a spatial discretization method that

- can easily treat general domains,
- can handle non-smooth (non-classical) solutions and data,
- can handle all types of boundary conditions (Dirichlet, Neumann, Robin),
- allows a local refinement of the mesh,
- and allows for realistic theoretical statements about the well-posedness of the PDE and convergence of numerical solutions.

We will see in the next chapter that the finite element method meets these criteria.

III. Finite Element Method

III.1. Introduction

The finite element method (FEM) is a widely used numerical method to solve partial differential equations that emerge in engineering and physics problems.

There are a lot of free finite element libraries available, for example

- FEniCS https://fenicsproject.org/
- deal.ii https://www.dealii.org/
- DUNE https://www.dune-project.org/

or also commercial software packages such as

- COMSOL Multiphysics https://www.comsol.com/
- ANSYS https://www.ansys.com/
- JCMsuite https://jcmwave.com/jcmsuite.

For a more extensive list https://en.wikipedia.org/wiki/List_of_finite_element_software_packages. For related tasks there exists free software, e.g.,

- Gmsh (mesh generation) http://gmsh.info/
- Triangle (2D mesh generation) https://www.cs.cmu.edu/~quake/triangle
- TetGen (3D mesh generation) http://www.tetgen.org/
- ParaView (visualization) https://www.paraview.org/
- VisIt (visualization) https://visit.llnl.gov/.

The main idea of the FEM is to divide a domain into a set of smaller subdomains, the so called elements, on which we are going to define functions to expand the exact solution with. The foundation of the finite element method has its origin in the works of Euler and Lagrange and the variational calculus developed since then in the 17th and 18th century. For a detailed historical account on the developments since then we refer to [Ste14]. Since then, engineers have been a major driving force in advancing the theoretical understanding and the practical use of the finite element method. The method has the major advantage that it:

- is designed to handle non-smooth solutions (compared to FDM),
- can handle complicated domains,
- can easily resolve local effects using non-uniform meshes,
- and the theory is based on realistic assumptions about solutions.

Previously we claimed numerous times, that the assumption underlying the construction of the finite difference method are unrealistic. Now we show an example which underlines this fact.

Example III.1 (Elliptic problem for two materials)**:** In the last chapter we extensively considered elliptic equations such as the Poisson problem. In reality, often we find this problem in a domain $\Omega \subset \mathbb{R}^n$, which connects different materials with different material properties. For instance, consider the one-dimensional Poisson problem

$$-(a(x)u'(x))' = f(x) \qquad \text{for } x \in (0,2), \tag{III.1}$$

with Dirichlet boundary conditions $u(0) = u_0, u(2) = u_2$. In electrostatics a is the dielectric constant, u is the electric potential, f is a charge distribution. We assume

$$a(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 2 & 1 < x \le 2 \end{cases},$$

due to the fact that the material in $0 < x < 1$ has different dielectric properties compared to the material in $1 < x < 2$. Without electric charges, the homogeneous problem has the solution

$$u(x) = u_0 + \begin{cases} \alpha x & 0 \le x < 1 \\ \alpha + \beta(x-1) & 1 \le x \le 2 \end{cases},$$

where $\alpha + \beta = u_2 - u_0$ and $\alpha - 2\beta = 0$ determine the coefficients. For $u_0 = 0$ and $u_2 = 1$ the solution has $\alpha = 2/3$ and $\beta = 1/3$ and is shown in Fig. III.1. Not only is this solution not twice continuously differentiable, already the first derivative is not continuous. However, we still need to explain how we computed the solution and where the conditions for $\alpha, \beta \in \mathbb{R}$ are coming from. Now we will develop a new solution concept to overcome this conceptual difficulty.

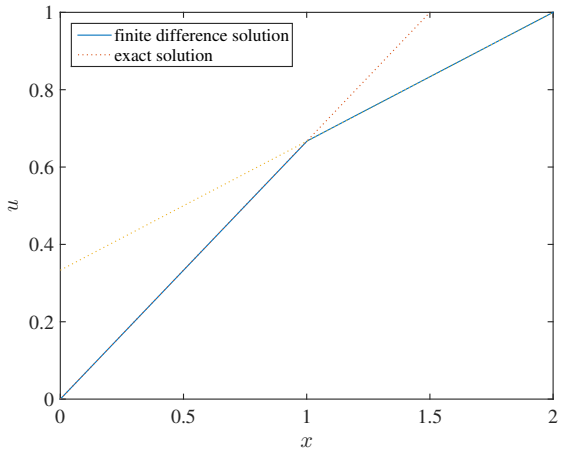


Figure III.1.: Solution of (III.1) with jumping coefficients.

III.2. Weak Solutions and Variational Problems

Consider again the 1D elliptic boundary value problem (BVP)

$$\begin{cases} -u'' = f, & \text{in } \Omega = (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{III.2}$$

Let $v \in C^1(\overline{\Omega})$ be an arbitrary continuously differentiable function. Then by multiplying (III.2) with the function v and integrating over the domain Ω we obtain

$$-\int_\Omega u''(x)v(x)\,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x.$$

Note that this relation is actually true for all continuous $v \in C^0(\overline{\Omega})$. Next, if $v \in C^1(\overline{\Omega})$ (and thus, in particular differentiable), according to the product rule we obtain

$$-u''v = u'v' - (u'v)'.$$

After integrating over Ω we therefore obtain

$$\int_\Omega u'(x)v'(x)\,\mathrm{d}x - [u'(x)v(x)]_0^1 = -\int_\Omega u''(x)v(x)\,\mathrm{d}x \overset{\text{(III.2)}}{=} \int_\Omega f(x)v(x)\,\mathrm{d}x. \tag{III.3}$$

This is again true for all $v \in C^1(\overline{\Omega})$.

Now set $C_0^1(\overline{\Omega}) := \{w \in C^1(\overline{\Omega}) \mid w(0) = w(1) = 0\}$. Then it holds true that $C_0^1(\overline{\Omega}) \subset C^1(\overline{\Omega})$. Moreover, for every $v \in C_0^1(\overline{\Omega})$ we obtain from (III.3)

$$\int_\Omega u'(x)v'(x)\,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x \tag{III.4}$$

This motivates to study the following *variational problem* or *weak problem*:

Find a function $u \in C_0^1(\overline{\Omega})$, such that (III.4) holds true for all $v \in C_0^1(\overline{\Omega})$.

This problem is *weaker* than the classical problem (III.2) in the sense that we search for a solution in the larger set $C_0^1(\overline{\Omega})$ instead of $C_0^2(\overline{\Omega})$ in (III.2).

Observation III.2: As we have seen above, the classical solution u to (III.2) solves the variational problem (III.4).

On the other hand, if a mapping $u \in C_0^2(\overline{\Omega})$ solves the variational problem (III.4), then an application of integration by parts shows that

$$\begin{aligned} \int_\Omega -u''(x)v(x)\,\mathrm{d}x = \int_\Omega u'(x)v'(x)\,\mathrm{d}x &\overset{\text{(III.4)}}{=} \int_\Omega f(x)v(x)\,\mathrm{d}x \\ \Leftrightarrow \int_\Omega (-u''(x) - f(x))v(x)\,\mathrm{d}x = 0 \quad &\forall v \in C_0^1(\overline{\Omega}). \end{aligned}$$

For the next result we need to define the following set of functions. Let $h\colon (a,b) \to \mathbb{R}$. Then

$$\operatorname{supp}(h) = \overline{\{x \in (a,b) \mid h(x) \neq 0\}}$$

is called the *support* of h. With this we define the set of *smooth functions with compact support* which are defined by

$$C_c^\infty(a,b) := \{h : (a,b) \to \mathbb{R} \mid h \text{ is infinitely often differentiable and } \operatorname{supp}(h) \subset (a,b) \text{ is compact}\}.$$

Recall that in $\mathbb{R}^n$, a set is compact, if it is closed and bounded. Since (a,b) is bounded and $\operatorname{supp}(h)$ is closed, the compactness is automatically fulfilled (but it has to be imposed if the domain is unbounded). Typical examples of smooth functions with compact support are so-called "bump functions". For example, $h : (-2,2) \to \mathbb{R}$ with

$$h(x) = \begin{cases} \exp\left(-\frac{1}{1-x^2}\right) & \text{for } x \in (-1,1), \\ 0 & \text{otherwise} \end{cases} \tag{III.5}$$

is such a bump function. This function is smooth, since $\exp(\cdot)$ and 0 are both smooth and since $\lim_{x\to\pm 1} \frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp\left(-\frac{1}{1-x^2}\right) = 0$ for $k = 0,\ 1,\ 2,\ \ldots$. Moreover, $\operatorname{supp}(h) = [-1,1]$ and thus $h \in C_c^\infty(-2,2)$.

Now we are ready to formulate the following lemma.

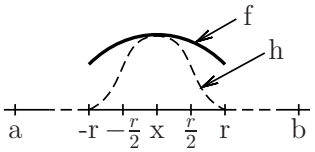
Lemma III.3 (Fundamental Lemma of Calculus of Variations)**:** Let $(a,b) \subset \mathbb{R}$ be an open interval and let $f\colon (a,b) \to \mathbb{R}$ be continuous. If it holds true that $\int_a^b f(x)h(x)\,\mathrm{d}x = 0$ for all $h \in C_c^\infty(a,b)$, then f is identically zero.

Proof. Assume without loss of generality that $f(x) > 0$ (or vice versa) for some $x \in \Omega$. Since f is continuous we find $B_r(x)$ such that $f(y) > 0$ for all $y \in B_r(x)$, $r > 0$. Here, $B_r(x)$ denotes the open ball of radius r around x, that is

$$B_r(x) := \big\{y \in (a,b) \mid |y - x| < r\big\}.$$

Next, choose $h \in C_c^\infty(a,b)$ with

$$\begin{cases} h(y) > 0, & \text{for all } y \in B_{\frac{r}{2}}(x), \\ h(y) \geq 0, & \text{for all } y \in B_r(x) \backslash B_{\frac{r}{2}}(x), \\ h(y) = 0, & \text{for all } y \text{ outside } B_r(x). \end{cases}$$



by continuity: $f(x) > 0$, $f(x-\delta) \neq 0\ \forall \delta \in (0,r)$

Such a smooth mapping h always exists (think of an appropriately scaled and shifted bump function)!

Then, it follows from the properties of f and h that

$$\begin{aligned}\int_a^b f(\widetilde{x})h(\widetilde{x})\,\mathrm{d}\widetilde{x} &= \int_{B_r(x)} f(\widetilde{x})h(\widetilde{x})\,\mathrm{d}\widetilde{x} + \int_{(a,b)\setminus B_r(x)} f(\widetilde{x})\underbrace{h(\widetilde{x})}_{=0}\,\mathrm{d}\widetilde{x}\\ &= \underbrace{\int_{B_r(x)\setminus B_{\frac{r}{2}}(x)} \overbrace{f(\widetilde{x})}^{>0}\overbrace{h(\widetilde{x})}^{\geq 0}\,\mathrm{d}\widetilde{x}}_{\geq 0} + \int_{B_{\frac{r}{2}}(x)} f(\widetilde{x})h(\widetilde{x})\,\mathrm{d}\widetilde{x}\\ &\geq 0 + \int_{B_{\frac{r}{2}}(x)} \min_{y\in B_{\frac{r}{2}}(x)} (f(y)h(y))\,\mathrm{d}\widetilde{x}\\ &= \underbrace{\min_{y\in B_{\frac{r}{2}}(x)} (f(y)h(y))}_{>0}\cdot \underbrace{\int_{B_{\frac{r}{2}}(x)} 1\,\mathrm{d}\widetilde{x}}_{=r} > 0.\end{aligned}$$

This contradicts with our assumption that $\int_a^b f(\widetilde{x})h(\widetilde{x})\,\mathrm{d}\widetilde{x} = 0$ for all $h \in C_c^\infty(a,b)$. Thus, there is no $x \in (a,b)$ with $f(x) \neq 0$. □

Remark III.4: In the above lemma, the functions $h \in C_c^\infty(a,b)$ are called *test functions*, since the function f is "tested" with h. Note, that the above lemma is still true, if we assume $h \in C_0^1(a,b) := \{f \in C^1(a,b) \mid \lim_{x\to a-} f(x) = \lim_{x\to b+} f(x) = 0\}$, since $C_0^\infty(a,b) \subset C_0^1(a,b)$.

Then, an application of this lemma leads to the following observation:

Observation III.5: If u is the solution to (III.4) and if $u \in C_0^2(\overline{\Omega})$, then u is already a "classical solution" to (III.2), since every $u \in C_0^2(\overline{\Omega})$ automatically fulfills the boundary conditions and

$$\int_\Omega (-u''(x) - f(x))v(x)\,\mathrm{d}x = 0 \quad \forall v \in C_0^1(\Omega) \supset C_c^\infty(\Omega)$$

yields $-u''(x) = f(x)$ for every $x \in \Omega$.

Remark III.6: Traditionally, variational problems are formulated as optimization problems on function spaces. For example, consider the functional $J : C_0^1(\overline{\Omega}) \to \mathbb{R}$ given by

$$J(u) = \int_\Omega \frac{1}{2}(u'(x))^2 - f(x)u(x)\,\mathrm{d}x.$$

We already saw an example of a PDE derived from an optimization argument, when we derived the minimal surface PDE in the example I.3. The *variational problem* consists of finding $u_{\min} \in C_0^1(\overline{\Omega})$ such that

$$J(u_{\min}) \leq J(u) \quad \forall\, u \in C_0^1(\overline{\Omega}). \tag{III.6}$$

Assume now that a solution $u_{\min} \in C_0^1(\overline{\Omega})$ exists. Let $v \in C_0^1(\overline{\Omega})$ be an arbitrary function and let $\alpha \in \mathbb{R}$ be a parameter. If the mapping $\alpha \mapsto J(u_{\min} + \alpha v)$ is differentiable, then the first order condition for a minimum reads

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} J(u_{\min} + \alpha v)\Big|_{\alpha=0} = 0.$$

This implies

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left[\int_\Omega \frac{1}{2}(u'_{\min}(x) + \alpha v'(x))^2 - f(x)(u_{\min}(x) + \alpha v(x))\,\mathrm{d}x\right] = \int_\Omega (u'_{\min}(x) + \alpha v'(x))v'(x) - f(x)v(x)\,\mathrm{d}x.$$

Since $\alpha = 0$ is a (local) minimum we get

$$\int_\Omega u'_{\min}(x)v'(x) - f(x)v(x)\,\mathrm{d}x \overset{!}{=} 0.$$

This is true for all $v \in C_0^1(\overline{\Omega})$. Observe that the minimization in (III.6) coincides with the weak problem (III.4).

The interpretation in the Poisson problem is that minimizing the function J corresponds to the "principle of minimal potential energy" that is often found to be valid for mechanical problems. For the minimal surface PDE the reasoning is obvious.

Remark III.7: Note that (III.4) only requires $u \in C_0^1(\overline{\Omega})$ while (III.2) requires $u \in C_0^2(\overline{\Omega})$. Sometimes the problem (III.2) does not have a classical solution $u \in C_0^2(\overline{\Omega})$, but there still exists a solution $u \in C_0^1(\overline{\Omega})$ to (III.4). In this case we call u a *weak solution* to (III.2) (or (III.4) respectively). We say that (III.4) is the *weak formulation* (= variational formulation) of (III.2).

Let us now have a look at the Poisson equation in 2D: We want to find $u : \Omega \to \mathbb{R}$ with $\Omega \subset \mathbb{R}^2$ with

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_1, \text{ (Dirichlet boundary conditions)} \\ \frac{\partial}{\partial \nu} u = 0 & \text{on } \Gamma_2, \text{ (Neumann boundary conditions)} \end{cases} \tag{III.7}$$

where $\Gamma_1 \cup \Gamma_2 = \Gamma = \partial\Omega$ with $\Gamma_1 \cap \Gamma_2 = \emptyset$.

Let $V_0 = \{w \in C^1(\overline{\Omega}) \mid w(x) = 0 \text{ on } \Gamma_1\}$, called the *trial space*, in which we search for a solution for the problem we consider now. The space of test functions is called the *test space*, and we will see, that this space is also V_0 in our example. However, in general the trial and test spaces can be different.

If u is a solution to (III.7), then, as in (III.3), we get

$$-\int_\Omega \Delta u \cdot v\,\mathrm{d}x = \int_\Omega f \cdot v\,\mathrm{d}x \quad \forall\, v \in V_0.$$

(From now on, we will mostly drop the arguments in the integrands to avoid excessive notation.) As in 1D we want to transform the integral on the left hand side such that second order derivatives disappear: The 2D version of the product rule reads

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

This implies

$$-\int_\Omega \Delta u \cdot v \,\mathrm{d}x = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \int_\Omega \operatorname{div}(v\nabla u)\,\mathrm{d}x$$

Now we apply the theorem of Gauss which yields

$$\int_\Omega \operatorname{div} w \,\mathrm{d}x = \int_\Gamma \nu \cdot w \,\mathrm{d}A,$$

where $\int_\Gamma \cdots \mathrm{d}A$ denotes the integral over the boundary and as usual, ν denotes the outward normal vector. Altogether we have

$$-\int_\Omega \Delta u \cdot v \,\mathrm{d}x = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \int_\Gamma \nu \cdot (v\nabla u)\,\mathrm{d}A \quad \forall\, v \in V_0.$$

Now recall that $\frac{\partial}{\partial \nu} u = \nu \cdot \nabla u \overset{!}{=} 0$ on Γ_2. Therefore, we get

$$\begin{aligned}\int_\Gamma \nu \cdot (v\nabla u)\,\mathrm{d}A &= \int_\Gamma v \frac{\partial}{\partial \nu} u \,\mathrm{d}A \\ &= \int_{\Gamma_1} v \frac{\partial}{\partial \nu} u \,\mathrm{d}A + \int_{\Gamma_2} v \frac{\partial}{\partial \nu} u \,\mathrm{d}A \\ &= \int_{\Gamma_1} v \frac{\partial}{\partial \nu} u \,\mathrm{d}A = 0,\end{aligned}$$

where the latter equality follows from $v \in V_0$ (and thus $v \equiv 0$) on Γ_1. Hence, we arrive at the variational formulation of (III.7), which is

$$\text{Find } u \in V_0 \text{ such that } \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \forall v \in V_0. \tag{III.8}$$

Remark III.8: a) In order to go from (III.7) to (III.8) we have to assume that Ω satisfies the conditions of the theorem of Gauss (e. g. it has to be a bounded set with piecewise smooth boundaries).

b) The Dirichlet boundary conditions are called *essential boundary conditions*, since they affect the choice of the underlying space V_0. The Neumann boundary conditions are called *natural boundary conditions*, since they do not affect the space V_0. Robin boundary conditions are also natural boundary conditions since $\frac{\partial u}{\partial \nu} + \sigma u = 0$ on Γ_3 leads to

$$\int_{\Gamma_3} v \frac{\partial}{\partial \nu} u \,\mathrm{d}A = -\int_{\Gamma_3} v\sigma u \,\mathrm{d}A.$$

The right-hand side then enters the variational formulation which is:

$$\text{Find } u \in V_0 \text{ such that } \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x + \int_{\Gamma_3} v\sigma u \,\mathrm{d}A = \int_\Omega f v \,\mathrm{d}x \quad \forall\, v \in V_0.$$

Next, we introduce the concept of *weak derivatives:*

Definition III.9 (Square integrable function)**:** Let $\Omega \subset \mathbb{R}^d$ be a domain and define the space of all square integrable functions on Ω

$$L^2(\Omega) := \left\{ v : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega |v(x)|^2 \,\mathrm{d}x < \infty \right\},$$

which becomes a complete vector space with the inner product

$$(u, v)_{L^2} = \int_\Omega u(x)v(x) \,\mathrm{d}x, \tag{III.9}$$

and the corresponding induced norm $\|u\|_{L^2} = \sqrt{(u,u)_{L^2}}$. Strictly speaking, the space L^2 consists of equivalence classes of functions, which only differ on sets of measure zero and $\|u\|_{L^2} = 0$ if $u(x) = 0$ almost everywhere. In particular, elements of $L^2(\Omega)$ do not need to be continuous functions, e.g. step functions.

Now, let $\Omega = (-1, 1)$. Consider $u : \Omega \to \mathbb{R}$, $x \mapsto |x|$. Obviously, we have $u \in L^2(\Omega)$, since

$$\int_\Omega |x|^2 \,\mathrm{d}x = \int_{-1}^{1} x^2 \,\mathrm{d}x = \frac{2}{3}.$$

But u is not differentiable at 0. We will show that u is weakly differentiable in the following sense.

Definition III.10: Let $u \in L^2(\Omega)$, where $\Omega \subset \mathbb{R}$ is an open interval. We say that $v \in L^2(\Omega)$ is the *weak derivative* of u if

$$\int_\Omega u\varphi' \,\mathrm{d}x = -\int_\Omega v\varphi \,\mathrm{d}x \quad \forall\, \varphi \in C_c^\infty(\Omega).$$

In this case we write $v =: u'$.

Example III.11: For the absolute value functions $u : \Omega \to \mathbb{R}$, $x \mapsto |x|$ we have

$$v(x) = u'(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Then it holds $v \in L^2(\Omega)$ and

$$\begin{aligned}\int_\Omega u(x)\varphi'(x)\,\mathrm{d}x &= \int_{-1}^0 (-x)\varphi'(x)\,\mathrm{d}x + \int_0^1 x\varphi'(x)\,\mathrm{d}x \\ &= [-x\varphi(x)]_{-1}^0 - \int_{-1}^0 (-1)\varphi(x)\,\mathrm{d}x + [x\varphi(x)]_0^1 - \int_0^1 \varphi(x)\,\mathrm{d}x \\ &= 0 - \int_{-1}^0 v(x)\varphi(x)\,\mathrm{d}x + 0 - \int_0^1 v(x)\varphi(x)\,\mathrm{d}x \\ &= -\int_\Omega v(x)\varphi(x)\,\mathrm{d}x \quad \forall\, \varphi \in C_c^\infty(\Omega).\end{aligned}$$

Remark III.12: a) Note that we can choose the value of v at $x = 0$ differently without changing the value of the integrals, so except for a set of measure zero, the weak derivative is uniquely determined.

b) If $u \in C^1(\Omega)$, then the weak derivative and the classical derivative coincide.

In order to introduce the weak derivative in $\mathbb{R}^d$, we use the previously defined derivative $D^\alpha u$ based on multi-indices.

Definition III.13 (Weak derivative)**:** Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with piecewise smooth boundary. Then $v \in L^2(\Omega)$ is the *weak derivative of u of α-order* if

$$\int_\Omega v\varphi\,\mathrm{d}x = (-1)^{|\alpha|}\int_\Omega uD^\alpha\varphi\,\mathrm{d}x \quad \forall\, \varphi \in C_c^\infty(\Omega).$$

Definition III.14: The set

$$H^1(\Omega) = \left\{ u : \Omega \to \mathbb{R} \;\middle|\; u \in L^2(\Omega),\ D^\alpha u \in L^2(\Omega) \text{ for } |\alpha| = 1 \right\},$$

i. e., the space of all square integrable functions with existing first order partial weak derivatives in $L^2(\Omega)$, is called (first-order) *Sobolev space*.

Let us return to (III.8): Find $u : \Omega \to \mathbb{R}$ such that $\Omega \subset \mathbb{R}^2$

$$\underbrace{\int_\Omega \nabla u \cdot \nabla v\,\mathrm{d}x}_{=a(u,v) \text{ bilinear form}} = \underbrace{\int_\Omega fv\,\mathrm{d}x}_{=F(v) \text{ linear form}}$$

Note that it is enough to consider weak derivatives in the bilinear form $a(\cdot,\cdot)$. Redefine

$$V_0 := \{u \in H^1(\Omega) \mid u = 0 \text{ on } \Gamma_1\}.$$

Then (III.8) reads

$$\text{Find } u \in V_0 \text{ such that } a(u,v) = F(v) \text{ for all } v \in V_0. \tag{III.10}$$

Definition III.15 (Bilinear/linear form)**:** Let V a vector space over $\mathbb{R}$. Then $a : V \times V \to \mathbb{R}$ is called a bilinear form if

$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w),$$
$$a(w, \alpha u + \beta v) = \alpha a(w, u) + \beta a(w, v),$$

for all $u, v, w \in V$ and all $\alpha, \beta \in \mathbb{R}$. Correspondingly, $f : V \to \mathbb{R}$ is called a linear form if

$$f(\alpha u + \beta v) = \alpha f(u) + \beta f(v),$$

for all $u, v \in V$ and all $\alpha, \beta \in \mathbb{R}$.

Remark III.16 (Reduction to homogeneous Dirichlet boundary conditions)**:** If we have inhomogeneous Dirichlet boundary conditions (i. e. $u = \varphi \neq 0$ on Γ_1), then the following helps to reduce the problem to the homogeneous case:

a) Find any $u_\varphi \in H^1(\Omega)$ such that $u_\varphi(x) = \varphi(x)$ on Γ_1. It can be shown that $u_\varphi \in H^1(\Omega)$ can be chosen continuous and hence it can be extended to the boundary Γ.

b) Find $u_0 \in V_0$ such that (with zero boundary conditions on Γ_1)

$$a(u_0, v) = \underbrace{F(v) - a(u_\varphi, v)}_{=:\widetilde{F}(v)} \quad \forall\, v \in V_0.$$

Then $u := u_0 + u_\varphi \in H^1(\Omega)$, $u(x) = u_\varphi(x) = \varphi(x)$ on Γ_1, and therefore,

$$a(u, v) = a(u_0 + u_\varphi, v) = F(v) \quad \forall\, v \in V_0.$$

Note that the trial space is $H^1(\Omega)$, while the test space is V_0.

III.3. Galerkin Methods

Galerkin methods are a general strategy in numerical analysis to convert a continuous (variational) problem into a finite-dimensional system of linear equations. The general strategy is described next.

a) **Get the variational/weak formulation:** Let V be an arbitrary vector space. Then the weak formulation is:

$$\text{Find } u \in V \text{ such that } a(u, v) = F(v) \quad \forall\, v \in V. \tag{III.11}$$

The space in which we search for a solution is called *trial space*. The space containing the test functions v is called the *test space*. In Galerkin methods, the test space and the trial space often coincide.

b) **Perform a Galerkin dimension reduction:** For each $n \in \mathbb{N}$ let $V_n \subset V$ be a subspace of V with $\dim(V_n) = n$. Then we solve:

$$\text{Find } u_n \in V_n \text{ such that } a(u_n, v_n) = F(v_n) \quad \forall\, v_n \in V_n \tag{III.12}$$

The problem (III.12) is called the *Galerkin equation* to (III.11) and may be interpreted as a projection of (III.11) onto V_n.

c) **Derive a linear system of equations:** Since V_n is of finite dimension there exists a basis $\{\varphi_1, \ldots, \varphi_n\}$ of V_n. Thus the solution u_n of (III.12) can be written as

$$u_n = \sum_{i=1}^{n} \alpha_i \varphi_i.$$

Using the bilinearity of a we get

$$a(u_n, v_n) = \sum_{i=1}^{n} \alpha_i a(\varphi_i, v_n) \quad \overset{v_n = \varphi_j}{\Longrightarrow} \quad a(u_n, \varphi_j) = \sum_{i=1}^{n} \alpha_i a(\varphi_i, \varphi_j).$$

Instead of general v_n test with the basis functions only, then we obtain

$$\text{Find } u_n \in V_n \text{ such that } a(u_n, \varphi_i) = F(\varphi_i) \quad \forall\, i \in \{1, \ldots, n\}. \tag{III.13}$$

Using the (bi-)linearity of a and F as above we can translate the problem (III.13) into a matrix-vector equation that can be solved on a computer:

Aim: Compute $\alpha = \left[\alpha_1, \ldots, \alpha_n\right]^{\mathsf{T}} \in \mathbb{R}^n$ from the matrix-vector equation $A_n \alpha = f_n$, where

$$A_n \alpha = \begin{bmatrix} a(\varphi_1, \varphi_1) & \cdots & a(\varphi_n, \varphi_1) \\ \vdots & \ddots & \vdots \\ a(\varphi_1, \varphi_n) & \cdots & a(\varphi_n, \varphi_n) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} F(\varphi_1) \\ \vdots \\ F(\varphi_n) \end{bmatrix} =: f_n \in \mathbb{R}^n. \tag{III.14}$$

We call A_n the *Galerkin matrix* associated to the Galerkin problem.

Observe, that the two problems (III.12) and (III.13) are indeed equivalent. In fact, it is enough to only test with basis functions by the (bi-)linearity of F and a: Assume we have determined u_n by solving (III.13) or (III.14), then by bilinearity of a we have

$$a(u_n, v_n) = \sum_{i=1}^{n} \alpha_i a(\varphi_i, v_n)$$

for every test function $v_n \in V_n$. Assume now that an arbitrary given test function has the representation

$$v_n = \sum_{j=1}^{n} \beta_j \varphi_j \text{ with coefficients } \beta = \left[\beta_1, \ldots, \beta_n\right]^{\mathsf{T}} \in \mathbb{R}^n.$$

Inserting the above representation yields

$$\begin{aligned} a(u_n, v_n) &= \sum_{i=1}^{n} \alpha_i a(\varphi_i, v_n) = \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{n} \beta_j a(\varphi_i, \varphi_j) \right) \\ &= \sum_{j=1}^{n} \beta_j \left(\sum_{i=1}^{n} \alpha_i a(\varphi_i, \varphi_j) \right) \\ &= \sum_{j=1}^{n} \beta_j a(u_n, \varphi_j) \\ &\overset{\text{(III.13)}}{=} \sum_{j=1}^{n} \beta_j F(\varphi_j) \overset{F \text{ linear}}{=} F \underbrace{\left(\sum_{j=1}^{n} \beta_j \varphi_j \right)}_{=v_n} = F(v_n), \end{aligned}$$

Therefore, the mapping $u_n \in V_n$ determined by (III.13) satisfies (III.12) for all test functions $v_n \in V_n$.

Next we introduce the concept of *Galerkin orthogonality*. Let u be the exact solution of (III.11) and let $u_n \in V_n$ be the solution of (III.12). Since $V_n \subset V$ we can "test" u also with elements from V_n. Thus we have

$$a(u_n, v_n) \overset{\text{(III.12)}}{=} F(v_n) \overset{\text{(III.11)}}{=} a(u, v_n).$$

This gives the property that

$$a(u_n - u, v_n) = 0, \quad \forall\, v_n \in V_n.$$

which is called *Galerkin orthogonality*. Recall, that two vectors v_1, v_2 in an inner-product space V with the inner product $\langle \cdot, \cdot \rangle$ are called orthogonal, if and only if

$$\langle v_1, v_2 \rangle = 0.$$

Moreover, we say that v_1 is *orthogonal* to a subspace $\widetilde{V} \subset V$ if

$$\langle v_1, w \rangle = 0 \quad \forall w \in \widetilde{V}.$$

What does that mean in our situation? If a admits an inner product on V, then the error $e_n = u - u_n$ is orthogonal to V_n with respect to the inner product induced by a. This shows that $u_n \in V_n$ is the *best approximation* of the exact solution u in terms of the metric/norm induced by a which is given by

$$\|w\|_a = \sqrt{a(w, w)}.$$

Then we get

$$\|e_n\|_a = \|u - u_n\|_a = \inf_{v_n \in V_n} \|u - v_n\|_a .$$

Remark III.17: a) If the discrete (finite dimensional) test space does not coincide with the discrete trial space, we call the method a *Petrov-Galerkin-method*.

b) A solution of (III.12) exists, if and only if the matrix A_n is invertible.

c) The solution u_n is not a grid function (as for the FDM) but u_n is a function defined on Ω. The dimension of V_n is often called the *degree of freedom of* u_n.

Definition III.18: Let V be an $\mathbb{R}$-vector space and $a : V \times V \to \mathbb{R}$ be a bilinear form. Then we call a

- *symmetric*, if $a(u, v) = a(v, u) \quad \forall\, u,\, v \in V$,
- *positive definite*, if $a(u, u) \geq 0\ \forall\, u \in V$ and $a(u, u) = 0$, if and only if $u = 0$,
- *negative definite*, if $-a$ is positive definite.

This directly translates into the corresponding properties for the matrix A_n, i. e., A is called positive definite, if and only if $x^{\mathsf{T}} A_n x \geq 0\ \forall\, x \in \mathbb{R}^n$ and $x^{\mathsf{T}} A_n x = 0$ is equivalent to $x = 0$. In particular, the positive definiteness of a implies that A_n is invertible.

Lemma III.19: Let $a : V \times V \to \mathbb{R}$ be a positive definite bilinear form of the variational problem (III.11). If a solution u exists, then it is unique.

Proof. Let u_1 and u_2 be two solutions to (III.11). Then we obtain

$$a(u_1, v) = F(v) \quad \forall\, v \in V \quad \text{and} \quad a(u_2, v) = F(v) \quad \forall\, v \in V.$$

This yields

$$a(u_1 - u_2, v) = a(u_1, v) - a(u_2, v) = 0 \quad \forall\, v \in V.$$

In particular, since $u_1 - u_2 \in V$ we have

$$a(u_1 - u_2, u_1 - u_2) = 0.$$

Since a is positive definite we get $u_1 - u_2 = 0$ and thus $u_1 = u_2$. □

The lemma stays true if a is negative definite.

Example III.20: Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with piecewise smooth boundary. The mapping $a : V \times V \to \mathbb{R}$ defined by

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x$$

is a symmetric bilinear form. If

$$V = \left\{ v \in H^1(\Omega) \;\middle|\; \frac{\partial}{\partial \nu} v = 0 \text{ on } \partial\Omega \right\}$$

then V contains all constant functions. But if $v \equiv k \in \mathbb{R}$ with $k \neq 0$, then it still holds $\nabla v \equiv 0$ and thus $a(v,v) = 0$, but $v \neq 0$! Therefore, a is not positive definite for this choice of V.

However, if the bilinear form a is restricted to the space

$$V_0 = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega\},$$

then the only constant function in V_0 is the zero function. In fact, $a : V_0 \times V_0 \to \mathbb{R}$ is positive definite, since

$$a(v,v) = \int_\Omega \|\nabla v\|^2 \,\mathrm{d}x = 0$$

implies $\|\nabla v\| = 0$ and so v is constant (up to a set of measure zero). In this example we see that the choice of the space V is not trivial and must be done with care.

Theorem III.21: Let $a : V \times V \to \mathbb{R}$ be a positive definite and symmetric bilinear form of the variational problem

$$\text{Find } u \in V \text{ such that } a(u,v) = F(v) \text{ for all } v \in V. \tag{V}$$

Define $J : V \to \mathbb{R}$ by $J(w) = \frac{1}{2}a(w,w) - F(w)$. Then the minimization problem

$$\text{Find } u \in V \text{ such that } J(u) \leq J(v) \text{ for all } v \in V. \tag{M}$$

has the same solution (if it exists) as (V).

Proof. Let $u \in V$ be a solution to (V). Let $v \in V$ be arbitrary. Set $w := v - u \in V \Leftrightarrow v = w + u$. Then we have

$$\begin{aligned} J(v) = J(u+w) &= \frac{1}{2}a(u+w,u+w) - F(u+w) \\ &= \frac{1}{2}(a(u,u) + a(w,u) + a(u,w) + a(w,w)) - F(u) - F(w) \\ &= \underbrace{\frac{1}{2}a(u,u) - F(u)}_{=J(u)} + \underbrace{a(u,w) - F(w)}_{=0 \text{ by (V)}} + \underbrace{\frac{1}{2}a(w,w)}_{\geq 0} \\ &\geq J(u) \end{aligned}$$

Therefore, if u solves (V), then it also solves (M).

Now assume that u is a solution of (M). We have to show that u also solves (V). Let $v_0 \in V$ and $\varepsilon \in \mathbb{R}$ be arbitrary. Define $K : \mathbb{R} \to \mathbb{R}$, $K(\varepsilon) := J(u + \varepsilon v_0)$. Then it holds

$$K(0) = J(u) \leq J(u + \varepsilon v_0) = K(\varepsilon) \quad \forall\, \varepsilon \in \mathbb{R}.$$

By definition of J we then get

$$\begin{aligned} K(\varepsilon) &= \frac{1}{2}a(u,u) - F(u) + \varepsilon a(u,v_0) - \varepsilon F(v_0) + \frac{1}{2}a(\varepsilon v_0, \varepsilon v_0) \\ &= h + g\varepsilon + f\varepsilon^2 \end{aligned}$$

Therefore, K is differentiable with respect to ε since it is a polynomial. Since K has a minimum at 0, we get $K'(0) = 0$ and furthermore,

$$K'(\varepsilon) = a(u, v_0) - F(v_0) + a(v_0, v_0)\varepsilon$$
$$K'(0) = a(u, v_0) - F(v_0) \stackrel{!}{=} 0.$$

This shows that u solves (V), since v_0 is arbitrary. □

Example III.22: Consider our usual elliptic BVP $-u''(x) = f(x)$ for $x \in (0,1)$ with homogeneous Dirichlet boundary conditions $u(0) = u(1) = 0$ leading to the bilinear form $a(u,v) = \int_0^1 u'v' \,\mathrm{d}x$ and $f(v) = \int fv \,\mathrm{d}x$. Using $\varphi_j(x) = \sin(j\pi x)$ for $j \in \mathbb{N}$ we define the finite-dimensional subspace

$$V_n = \operatorname{span}\{\varphi_j(x)\}_{1 \le j \le n},$$

which leads to

$$a(\varphi_i, \varphi_j) = \int_0^1 \pi^2 ij \cos(i\pi x)\cos(j\pi x)\,\mathrm{d}x = \begin{cases} \frac{j^2}{2}\pi^2 & i = j \\ 0 & \text{else} \end{cases},$$

and leads to a diagonal Galerkin matrix and $(f_n)_j = \int_0^1 \sin(j\pi x) f(x)\,\mathrm{d}x$, which can be readily solved using $\alpha_j = 2(f_n)_j/(j^2\pi^2)$ and $u_n(x) = \sum_j \alpha_j \sin(j\pi x)$. This is an exceptionally simple example. Usually we will not be able to solve the Galerkin equations explicitly due to the lack of an orthogonal basis of eigenfunctions.

Remark III.23: a) The usefulness of the Galerkin methods depend on the finite dimensional spaces $(V_n)_{n\in\mathbb{N}}$ with $V_n \subset V$ for each $n \in \mathbb{N}$. The quality of the approximation depends on the best-approximation

$$\inf_{v_n \in V_n} \|u - v_n\|_V \quad \text{(which needs to be small)}.$$

This is the minimal error, that is caused by replacing the infinite dimensional space V by the finite dimensional subspace V_n.

b) Usually we require $V_n \subset V$, i. e., all elements in V_n have to satisfy the same regularity and essential boundary conditions as the exact solution.

c) The "art" of the Galerkin method is to identify $(V_n)_{n\in\mathbb{N}}$ such that the resulting linear problem $A_n\alpha = f_n$ is "easy" to solve and the Galerkin matrix "easy" to compute.

d) In general we have to choose n very large to get a good approximation, so we have to solve high-dimensional linear systems (same as for the FDM). This means we better find a construction of V_n which produces a sparse Galerkin matrix.

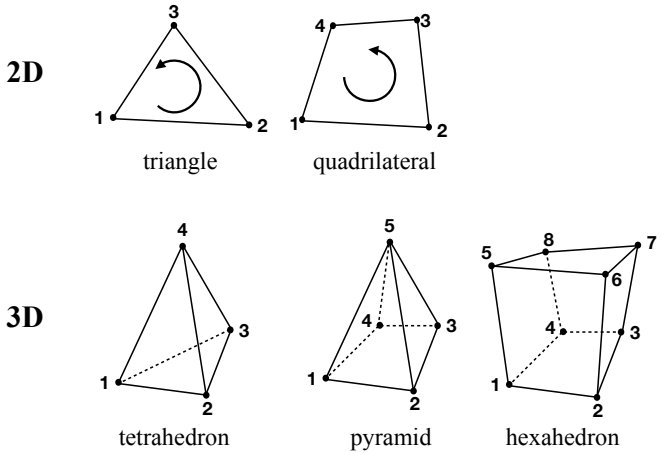
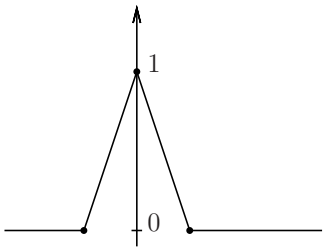


Figure III.2.: Common elements $\overline{\Omega}_i$ used to decompose domains in 2D and 3D.

In the Galerkin finite element method we will see further below that the finite dimensional spaces V_n are spanned, for instance, by basis functions of the following shape:

$$\varphi(x) = \begin{cases} 1-x, & \text{if } x \in [0,1], \\ x+1, & \text{if } x \in [-1,0], \\ 0, & \text{else.} \end{cases}$$



Observe that this function is differentiable in the weak sense, but not in the classical sense.

III.4. Finite Elements

In the previous section we have seen that the choice of the basis functions $(\varphi_i)_{i=1}^n \subset V_n$ determines the usefulness of the Galerkin method. Typical examples are polynomials (monomials) over Ω, or spectral Galerkin method (eigenfunctions of the differential operators).

In this section we introduce the finite element basis functions. We choose $(\varphi_i)_{i=1}^n \subset V_n$ in such a way that $\varphi_i \neq 0$ holds only in a small region of the domain Ω. Therefore, $a(\varphi_i, \varphi_j) = 0$ for "most" i and j, which implies that the matrix A_n becomes sparse.

From this we derive the following strategy:

a) Decompose $\overline{\Omega}$ into finitely many "nice" parts $\overline{\Omega} = \bigcup_{i=1}^{n_e} \overline{\Omega}_i$ in an *admissible way*. Each of the $\overline{\Omega}_i$ is called an "element" of the decomposition. Different types of commonly used elements are shown in Fig. III.2.

b) On each element Ω_i we only allow for a certain (finite dimensional) space of functions

supported on the element. Typically, these are polynomials up to a certain degree. An element together with a class of functions is called a *finite element*.

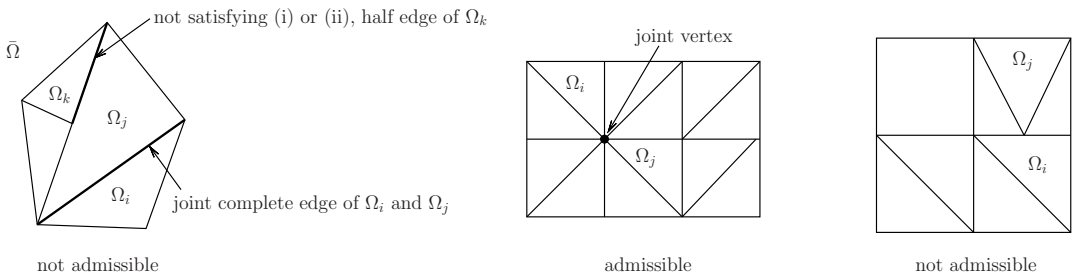
However, in order to be able to construct a basis and build the Galerkin matrix, the decomposition will have to satisfy certain properties.

Definition III.24 (Admissible decomposition)**:** Assume a given domain $\Omega \subset \mathbb{R}^d$ with polygonal boundary and let

$$\overline{\Omega} = \bigcup_{i=1}^{n_e} \overline{\Omega}_i,$$

be a decomposition of $\overline{\Omega}$ into elements $\{\overline{\Omega}_i\}_{i=1,..,n_e}$.

- $d = 1$: If $\Omega = (a, b) \subset \mathbb{R}$ and $a = x_0 < x_1 < \ldots < x_N < x_{N+1} = b$, then an admissible decomposition is given by a union of non-overlapping but adjacent intervals, i.e., $\Omega_i = (x_{i-1}, x_i)$, $i = 1, \ldots, n_e = N + 1$.
- $d = 2$: A decomposition of Ω into triangles (or convex quadrilaterals etc) $\{\Omega_i\}_{i=1}^{n_e}$ is admissible, if for each i and j exactly one of the following cases is true:
  i) $\Omega_i = \Omega_j$;
  ii) $\overline{\Omega}_i \cap \overline{\Omega}_j$ is a joint complete edge of both Ω_i and Ω_j;
  iii) $\overline{\Omega}_i \cap \overline{\Omega}_j$ is a joint vertex of Ω_i and Ω_j;
  iv) $\overline{\Omega}_i \cap \overline{\Omega}_j = \emptyset$.



- $d = 3$: A decomposition into tetrahedrons (or hexahedrons, pyramids, etc) $\{\Omega_i\}_{i=1}^{n_e}$ is called admissible, if for each i and j exactly one of the following cases is true:
  i) $\Omega_i = \Omega_j$;
  ii) $\overline{\Omega}_i \cap \overline{\Omega}_j$ is a joint complete face of both Ω_i and Ω_j (triangle, quadrilateral, ...);
  iii) $\overline{\Omega}_i \cap \overline{\Omega}_j$ is a joint complete edge of both Ω_i and Ω_j;
  iv) $\overline{\Omega}_i \cap \overline{\Omega}_j$ is a joint vertex of Ω_i and Ω_j;
  v) $\overline{\Omega}_i \cap \overline{\Omega}_j = \emptyset$.

Examples for two-dimensional admissible decompositions are shown in Fig. III.3.

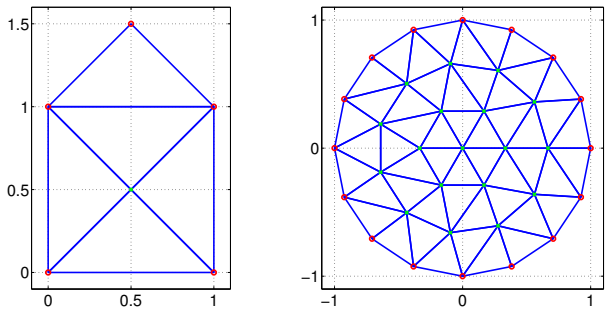


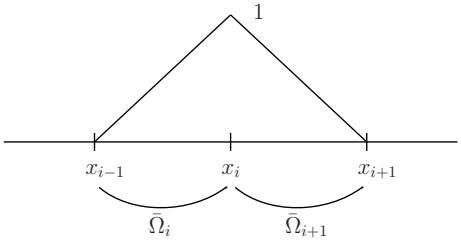
Figure III.3.: Admissible decompositions **(left)** of a house and **(right)** of a disc.

Example III.25 (Linear finite elements in 1D)**:** Consider the BVP

$$-u'' = f \qquad \text{in } \Omega = (0,1),$$

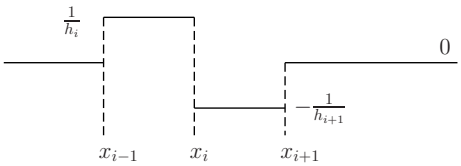
with $u(0) = u(1) = 0$. We choose an admissible decomposition of Ω as $\overline{\Omega} = \bigcup_{i=1}^{n_e} \overline{\Omega}_i$ where $\Omega_i = (x_{i-1}, x_i)$ and $0 = x_0 < x_1 < \ldots < x_N < x_{N+1} = 1$. We abbreviate $h_i = x_i - x_{i-1}$ for $i = 1, \ldots, N+1$ and set $n_e = N+1$. We define the linear finite element basis functions

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h_i}, & \text{if } x \in \overline{\Omega}_i, \\ \frac{x_{i+1} - x}{h_{i+1}}, & \text{if } x \in \overline{\Omega}_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$



for all $i = 1, ..., N$. In other words, φ_i is piecewise linear in Ω with $\varphi_i(x_j) = \delta_{ij}$. We have $\varphi_i \in C^0(\overline{\Omega})$ for $i = 1, \ldots, N$. The Galerkin methods then use the spaces $V_n = \text{span}\{\varphi_i \mid i = 1, \ldots, n = N\}$. Note that φ_i is not differentiable in the classical sense, but it is weakly differentiable with weak derivative

$$\varphi_i'(x) = \begin{cases} \frac{1}{h_i}, & \text{if } x \in \Omega_i, \\ -\frac{1}{h_{i+1}}, & \text{if } x \in \Omega_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$



The Galerkin matrix $A_n = [a(\varphi_i, \varphi_j)]_{i,\,j=1}^n$ has the entries

$$\begin{aligned}
a(\varphi_i, \varphi_j) &= \int_\Omega \varphi_i'(x)\varphi_j'(x)\,\mathrm{d}x \\
&= \begin{cases} \int_{\Omega_i \cup \Omega_{i+1}} (\varphi_i'(x))^2\,\mathrm{d}x, & \text{if } i = j, \\ \int_{\Omega_{i+1}} \varphi_i'(x)\varphi_{i+1}'(x)\,\mathrm{d}x, & \text{if } j = i+1, \\ \int_{\Omega_i} \varphi_i'(x)\varphi_{i-1}'(x)\,\mathrm{d}x, & \text{if } j = i-1, \\ 0, & \text{otherwise,} \end{cases} \\
&= \begin{cases} h_i \frac{1}{h_i^2} + h_{i+1}\frac{1}{h_{i+1}^2}, & \text{if } i = j, \\ -\frac{1}{h_{i+1}}\frac{1}{h_{i+1}} h_{i+1}, & \text{if } j = i+1, \\ -\frac{1}{h_i}\frac{1}{h_i} h_i, & \text{if } j = i-1, \\ 0, & \text{otherwise,} \end{cases} \\
&= \begin{cases} \frac{1}{h_i} + \frac{1}{h_{i+1}}, & \text{if } i = j, \\ -\frac{1}{h_{i+1}}, & \text{if } j = i+1, \\ -\frac{1}{h_i}, & \text{if } j = i-1, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

Altogether, this yields

$$A_n = \begin{bmatrix} \frac{1}{h_1} + \frac{1}{h_2} & -\frac{1}{h_2} & & \\ -\frac{1}{h_2} & \frac{1}{h_2} + \frac{1}{h_3} & \ddots & \\ & \ddots & \ddots & -\frac{1}{h_n} \\ & & -\frac{1}{h_n} & \frac{1}{h_n} + \frac{1}{h_{n+1}} \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Note that if we have a uniform mesh, i. e., $h = h_i = h_{i+1}$ for $i = 1, \ldots, n$, then

$$A_n = \frac{1}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Up to a scaling with $\frac{1}{h}$ this is the same matrix as for the finite difference method. However, the vector on the right hand side is, in general, different. For the right-hand side we obtain

$$f_n = \begin{bmatrix} F(\varphi_1) \\ \vdots \\ F(\varphi_n) \end{bmatrix} = \begin{bmatrix} \int_\Omega f(x)\varphi_1(x)\,\mathrm{d}x \\ \vdots \\ \int_\Omega f(x)\varphi_n(x)\,\mathrm{d}x \end{bmatrix}.$$

In practice, the vector $f_n \in \mathbb{R}^n$ is computed by quadrature methods. Then the finite element approximation can be obtained by solving the linear system $A_n\alpha = f_n$ on a computer.

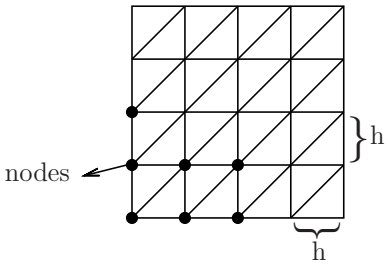
Example III.26 (Linear finite elements in 2D)**:** Consider

$$\begin{cases} -\Delta u = f & \text{in } \Omega = (0,1) \times (0,1), \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

The domain Ω is decomposed as follows:

a) Decompose $\overline{\Omega}$ into squares with edge length $h = \frac{1}{N+1}$, $N \in \mathbb{N}$. This gives $(N+1)^2$ squares.

b) Decompose each square along one of the diagonals into two triangles. This decomposition into $n_e = 2(N+1)^2$ elements is admissible.



Each element is a triangle of the form  up to rotation.

The vertices of the triangles are also called *nodes*.

In this example the nodes are $x_{lm} = (lh, mh)$, l, $m = 0, \ldots, N+1$. There are $(N+2)^2$ nodes (and N^2 inner nodes). In light of the Dirichlet boundary condition, the inner nodes determine the dimensions of the finite element space, if we use Dirichlet boundary condition on the whole boundary. The dimension of the finite element space V_n are also called the *degrees of freedom* of the global FEM approximation. More formally, we have

$$V_n = \operatorname{span}\left\{ \varphi_{lm} \in C^0(\overline{\Omega}) \;\middle|\; \varphi_{lm}(x_{\widetilde{l}\widetilde{m}}) = \begin{cases} 1, & \text{if } \widetilde{l} = l \text{ and } \widetilde{m} = m, \\ 0, & \text{otherwise,} \end{cases} \right.$$
$$\left. \varphi_{lm} \text{ is piecewise linear on each element} \right\}.$$

These basis functions are called *pyramid/tent/hat functions*. On each element they have the form

$$\varphi_{lm}(x_1, x_2) = a_0^{ilm} + a_1^{ilm} x_1 + a_2^{ilm} x_2, \qquad \text{for all } (x_1, x_2) \in \Omega_i.$$

Figure III.4 shows a typical basis function. It is only nonzero on the shaded triangles. Observe that the values on the nodes completely determines the basis function on $\overline{\Omega}$.

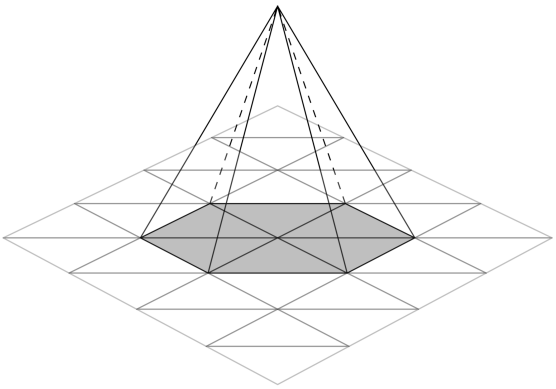


Figure III.4.: A typical pyramid basis function

III.5. FEM – Mesh Generation

The process of generating a decomposition of a domain into simpler elements is called mesh or grid generation. We distinguish *structured* and *unstructured meshes*. When we solved PDEs using the finite difference method, we already considered structured meshes that were identified using points/nodes $x_{i_1..i_n}$ for some indices $i_l \in \mathbb{N}$, where usually adjacency/connectivity information could be derived from the values of the indices $i_1, ..., i_n$. For unstructured meshes such an information is usually not given and adjacency/connectivity can be very irregular. The element shapes that are used most often are edges (1D), triangles and quadrilaterals (2D), tetrahedrons and hexahedrons (3D). Below we introduce some simple and practical concepts useful when considering mesh generation.

Definition III.27 (k-simplex)**:** Let $x_k \in \mathbb{R}^n$ be a set of $k+1$ distinct points so that $x_r - x_0$ for $r = 1, ..., k$ are linearly independent. Then the convex hull of points

$$C(x_0, ..., x_k) = \{\xi_0 x_0 + \xi_1 x_1 + ... + \xi_k x_k : \sum_{i=0}^{k} \xi_i = 1 \text{ and } \xi_i \geq 0\} \subset \mathbb{R}^n, \qquad \text{(III.15)}$$

is a k-simplex. For a point $x \in C$ we call $(\xi_0, ...\xi_k)$ its barycentric coordinates. A 0-simplex is a vertex, a 1-simplex is a line, a 2-simplex is a triangle, a 3-simplex is a tetrahedron and linear independency requires $n \geq k$. By removing the ith vertex each k-simplex is composed of $k+1$ $(k-1)$-simplices $C(x_0, ..., \widehat{x_{i+1}}, ..., x_k)$ for $i = 0, ..., k$. E.g., each tetrahedron (3-simplex) is composed of 4 triangles (2-simplex), which are the faces/boundaries of the tetrahedron, each triangle (2-simplex) is composed of 3 lines (1-simplex), which are the edges/boundaries of the triangle. Each line (1-simplex) is composed of 2 vertices (0-simplex), which are the boundaries of the line.

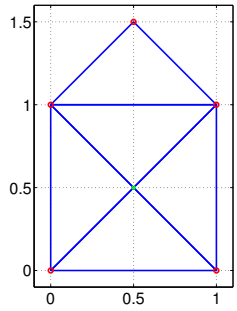
Definition III.28 (Mesh representation with simplices)**:** Let $x_k \in \mathbb{R}^n$ for $k = 1, ..., n_p$ be a set of $n_p \in \mathbb{N}$ distinct points and $e_k \in \mathbb{N}^{n_e \times (k+1)}$ such that

$$\Omega_l = C(x_{e_k(l,1)}, ..., x_{e_k(l,k+1)}),$$

using (III.15) is a k-simplex for all $l = 1, ..., n_e$. We say n_p is the *number of points* and n_e is the *number of elements*. As before we require the set Ω_l to form an admissible decomposition of $\overline{\Omega} = \bigcup_{l=1}^{n_e} \overline{\Omega}_l$. The mesh is represented using the vertex positions $x_k \in \mathbb{R}^n$ and using the element description encoded in $e_k \in \mathbb{N}^{n_e \times (k+1)}$.

Definition III.29 (Global mesh refinement)**:** Splitting each k-simplex of a mesh into $(k+1)$ congruent k-simplices of half the size is called *global mesh refinement*. Global mesh refinement does not change the quality of the mesh due to the congruence.

Example III.30 (House mesh)**:** The mesh shown below can be represented as follows:



$$x_1 = (0,0)^\top, \quad x_2 = (1,0)^\top, \quad x_3 = (0.5,0.5)^\top,$$
$$x_4 = (0,1)^\top, \quad x_5 = (1,1)^\top, \quad x_6 = (0.5,1.5)^\top,$$
$$e_2 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 4 \\ 2 & 5 & 3 \\ 3 & 5 & 4 \\ 4 & 5 & 6 \end{pmatrix} \in \mathbb{N}^{5\times 3}, \quad n_p = 6, \quad n_e = 5, \quad k = 2.$$

The domain consists of $n_p = 6$ vertices, 10 lines, $n_e = 5$ triangles.

Remark III.31: While for example $C(x_1, x_2, x_3)$ and $C(x_3, x_2, x_1)$ describe the same 2-simplex, we will sometimes pick a particular order of points in the corresponding e_2 in order to obtain a certain orientation of the simplex. In the example above all triangles in e_2 have the same orientation in the sense that points are enumerated anticlockwise.

III.5.1. Structured Meshes

The advantage of structured meshes is that they are usually easy and fast to generate and efficient to store. In many cases, due to the regular structure vertex positions and connectivity can be even computed for the given shape. They offer a lot of control for resolution and using block-wise structured meshes also allows for realistic geometries.

1D mesh: Interval and ring

The simplest possible structured mesh is the interval $\Omega = (0, L) \subset \mathbb{R}^1$, where $x_i = (i-1)h$ with $i = 1, ..., N+2$ and $h = L/(N+1)$. Here we have

$$e_1 = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \\ .. & .. \\ N+1 & N+2 \end{pmatrix} \in \mathbb{N}^{N+1\times 2}, \quad n_p = N+2, \quad n_e = N+1, \quad k = 1. \tag{III.16}$$

Another simple structured mesh of intervals is a (piecewise linear) ring $x_i = (\cos\varphi_i, \sin\varphi_i)^\top \in \mathbb{R}^2$ with $\varphi_i = 2\pi(i-1)/(N+1)$ with $i = 1, ..., N+1$ with

$$e_1 = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \\ .. & .. \\ N+1 & 1 \end{pmatrix} \in \mathbb{N}^{N+1\times 2}, \quad n_p = N+1, \quad n_e = N+1, \quad k = 1. \tag{III.17}$$

2D mesh: Tensor mesh

Let $0 = \bar{x}_1 < \bar{x}_2 < ... < \bar{x}_k = L_x$ and $0 = \bar{y}_1 < \bar{y}_2 < ... < \bar{y}_l = L_y$ increasing x and y coordinates. Then we obtain a tensor mesh with $n_p = k \cdot l$ vertices via $x_{ij} = (\bar{x}_i, \bar{y}_j)$ for $1 \le i \le k$ and $1 \le j \le l$. The resulting mesh consists of $n_e = 2(k-1)(l-1)$ triangles built from $(x_{ij}, x_{(i+1)j}, x_{i(j+1)})$ and $(x_{(i+1)(j+1)}, x_{(i+1)j}, x_{i(j+1)})$ for $i = 1, .., k-1$ and $j = 1, .., l-1$ using the lexicographical order. In Listing III.1 the MATLAB code that generates the tensor mesh based on a lexicographical ordering and with nonuniform element size in y-direction is presented and the corresponding simplex mesh of triangles e_2 with vertices x, y is plotted using the MATLAB command `triplot` and shown in Figure III.5.

Listing III.1: MATLAB code tensor mesh generation

```
clear all
close all

% generate points
k = 10; l = 20;
xx = linspace(0,1,k);
yy = linspace(0,1,l).^(1.3);
[x,y]=meshgrid(xx,yy);
x = x';y = y';

% lexicographical ordering
ix  = reshape(1:(k*l),[k l]);
ix1 = ix(1:k-1,1:l-1); ix1 = ix1(:);
ix2 = ix(2:k  ,1:l-1); ix2 = ix2(:);
ix3 = ix(1:k-1,2:l  ); ix3 = ix3(:);
ix4 = ix(2:k  ,2:l  ); ix4 = ix4(:);

% connectivity (2 triangles per quad)
e1 = [ix1 ix2 ix3]; % lower triangles
e2 = [ix2 ix4 ix3]; % upper triangles
v = [x(:),y(:)];    % vertices

patch('Faces',e1,'Vertices',v,'FaceColor'
    ,'r')
patch('Faces',e2,'Vertices',v,'FaceColor'
    ,'b')
e = [e1 ; e2];
```

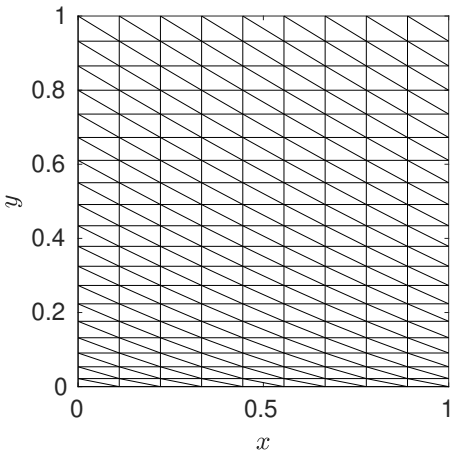


Figure III.5.: Structured tensor mesh.

2D mesh: Deformed ring and disc

We transform the rectangular structured mesh shown in the left panel of Figure III.6 via $(x, y) = x_2(\cos x_1, \sin x_1)$ into the ring $\Omega = \{x \in \mathbb{R}^2 : 1 < \|x\| < 2\}$ shown in the right panel of Figure III.6. When identifying the degrees of freedom for $x_1 = 0$ and $x_1 = 2\pi$ in the lexicographical ordering, the resulting decomposition is a suitable admissible decomposition of a ring.

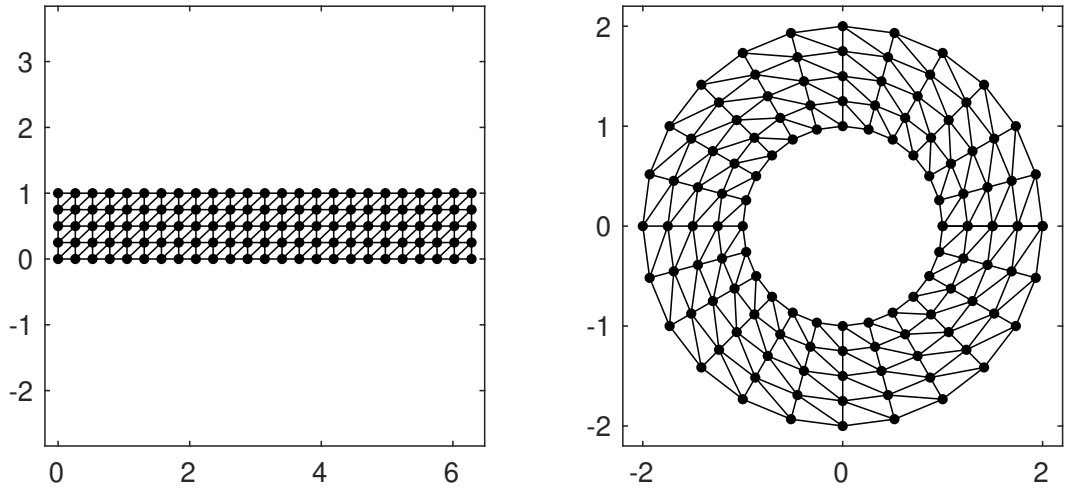


Figure III.6.: Structured rectangular mesh transformed into ring.

Attempting a similar transformation of a cube $(-1, 1)^2$ into a disc is possible in principle and is shown in Figure III.7 for the resulting cells of quadrilateral shape (to be divided into triangles). However, the resulting almost singular aspect ratio of triangles in the corners make this particular mapping that generates the deformed mesh infeasible for solving PDEs.

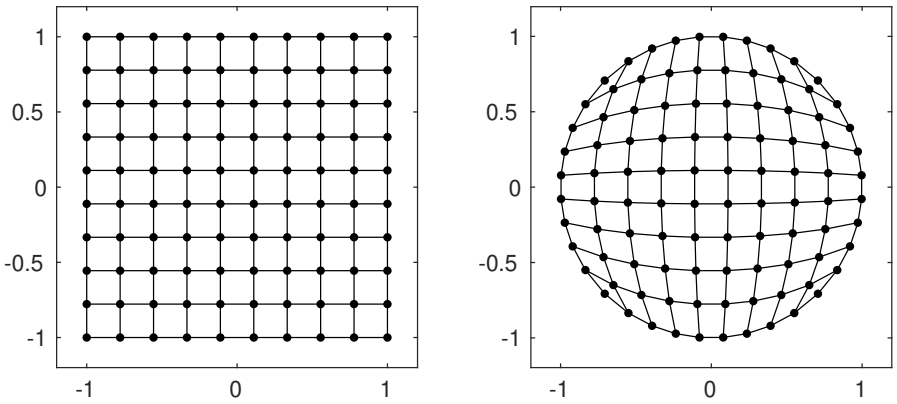


Figure III.7.: Structured rectangular mesh transformed into disc.

However, combining different structured meshes as shown in Figure III.8 allows to obtain a somewhat better result in terms of mesh quality with a different mapping.

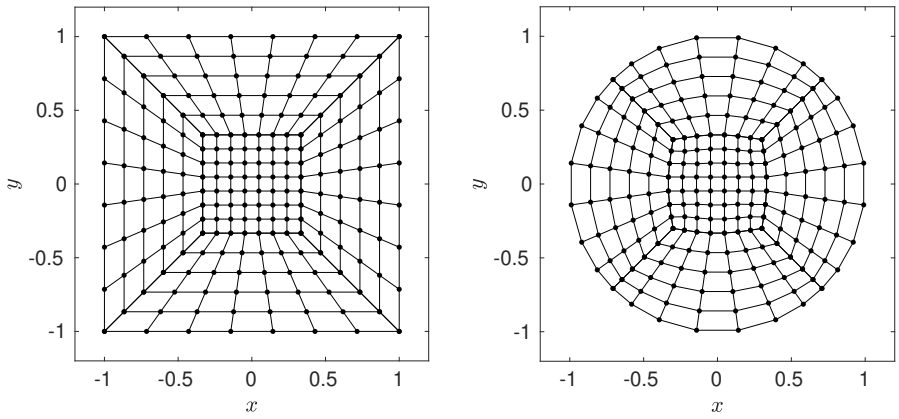


Figure III.8.: Union of 5 quadrilateral based structured meshes transformed into disc.

2D mesh: Deformed domain

In general, if the shape/boundary of the domain is parametrized by a function $y = f(x)$ for $-2 \leq x \leq 2$, then it might be possible to use a simple domain $(\bar{x}, \bar{y}) \in (-2, 2) \times (-1, 1)$ such as a reactangle and deform it using $(x, y) = (\bar{x}, \bar{y} f(\bar{x}))$ as shown in Figure III.9. As long as the deformation is not too strong, the resulting mesh will have a reasonable quality and the original connectivity will not be affected.

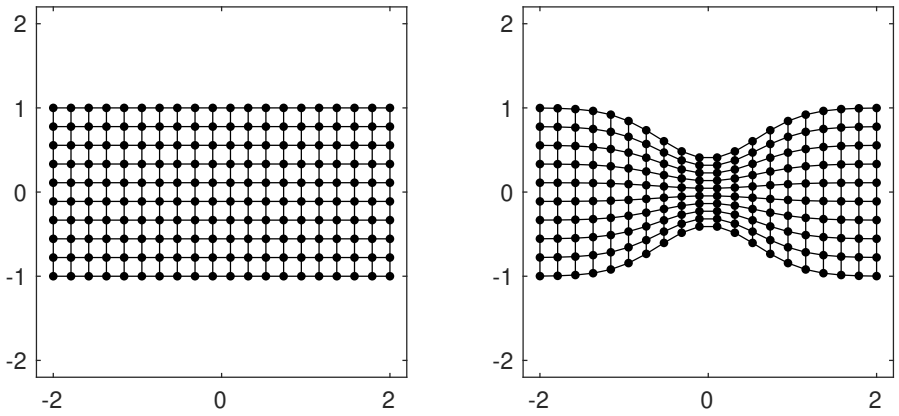


Figure III.9.: Rectangular mesh deformed using a generic function.

III.5.2. Unstructured Meshes

Unstructured grids are decompositions of a domain into simple shapes such as simplices, which have no apparent or an irregular connectivity. These meshes are usually not

generated 'by hand' as the structured meshes above but using sophisticated algorithms. Unstructured meshes are generated starting from a description of the boundary of the domain, where the algorithm then generates additional points and their connectivity. The representation of unstructured meshes is as pointed out in Definition III.28.

Most common algorithms to create high-quality meshes are based on Delaunay triangulation and constrained Delaunay triangulation.

Definition III.32 (Delaunay triangulation)**:** A mesh $x_i \in \mathbb{R}^n$ for $i = 1, ..., n_p$ and $e_k \in \mathbb{N}^{n_e \times (k+1)}$ is a *Delaunay triangulation*, if no point x_i is inside any of the n_e circumhyperspheres of the k-simplices in e_k. In particular for $n = 2$ this means that no point x_i is in the circumsphere of any triangle in e_k.

Definition III.33 (Convex set)**:** A subset $\Omega \subset \mathbb{R}^n$ is called *convex*, if for every $x, y \in \Omega$ also the convex combination lies in the subset, i.e., $tx + (1-t)y \in \Omega$ for all $t \in [0, 1]$.

As an example, the Delaunay triangulation for a set of 20 random points is shown in the left panel of Figure III.10. The domain $\overline{\Omega} = \bigcup_i \overline{\Omega}_i$ is the convex-hull of these 20 points, i.e., the smallest convex set containing all given points. Due to this property Delaunay triangulations would be of little use for non-convex domains. However, constrained Delaunay triangulations come to rescue in this case.

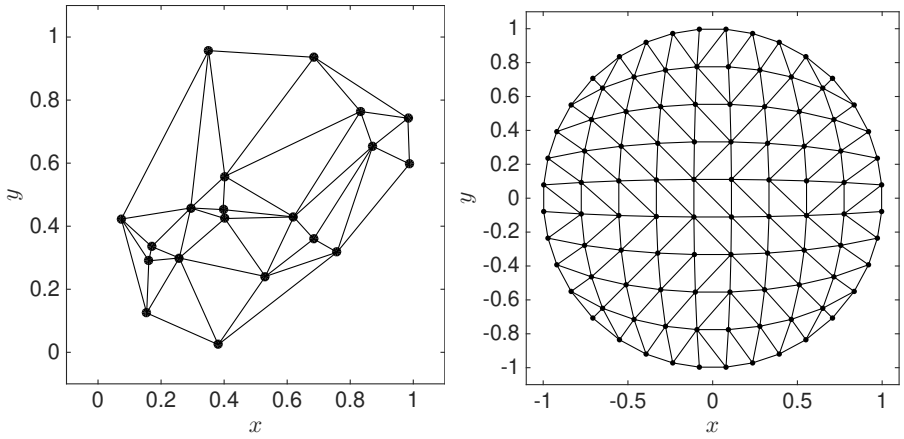


Figure III.10.: Delaunay triangulation **(left)** of a set of 20 random points and **(right)** of a disc based on a point set of a structured mesh.

If one is able to describe the (convex) domain given a generated point set (point set can be obtained by transformation as before for structured meshes), then the connectivity can be obtained using a Delaunay trianguation as is shown in the right panel of Figure III.10 for the example of the point set describing a disc of radius one.

For more complex domains as the annulus we discussed before, we can use the Delaunay triangulation to generate the connectivity based on a generated point set and then remove the extra elements/simplices (from e_2 using a slice) as shown in Figure III.11.

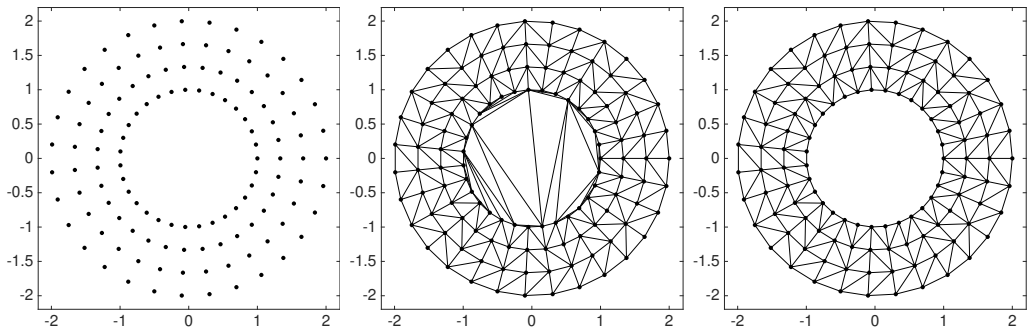


Figure III.11.: **(left)** point list **(middle)** Delaunay triangulation of point list with extra elements for $\|r\| < 1$ **(right)** triangulation with extra elements removed.

However, this procedure will not always work due to the effect shown in the left panel of Figure III.12, where an edge of the constructed Delaunay triangulation intersects with the intended domain boundary shown as a dashed blue line. Selecting the corresonding edge and computing a constrained Delaunay triangulation resolves the issue as shown in the corresonding middle panel. The right panel shows the final mesh with the extra elements removed.

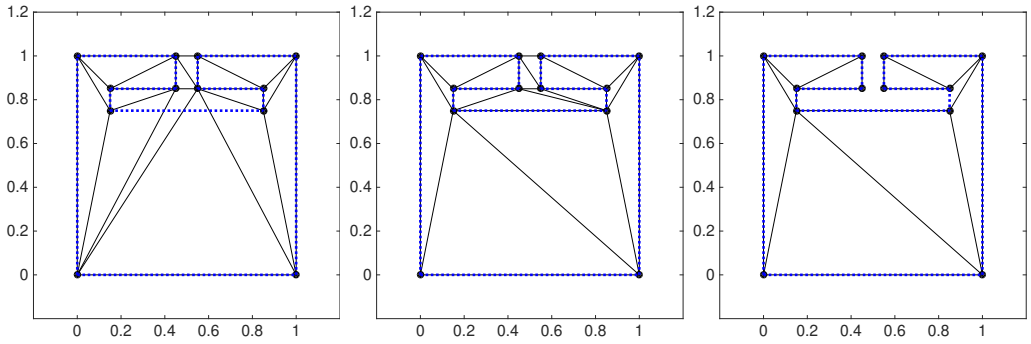


Figure III.12.: **(left)** Delaunay triangulation of point list intersects domain boundary (blue dotted) **(middle)** constrained Delaunay avoids intersection **(right)** constrained Delaunay with extra elements removed.

III.5.3. Mesh Generation Software

In the course we will use the 2D mesh generation tool Triangle, which is freely available in version 1.6 from http://www.netlib.org/voronoi/triangle.zip and copyrighted by the author Jonathan Shewchuk. This program creates Delaunay triangulations, constrained Delaunay triangulations and triangle meshes of high quality. Below we point out some hints and examples concerning usage. After downloading the program, compile using your favorite C-compiler, e.g.,

```
cc triangle.c -o triangle
```

or using `make` as explained in the `README` file to produce an executable. Then running `./triangle` gives a short help message, whereas `./triangle -h` gives a very detailed information about Triangle, input files, output files, command line switches. In particular, the detailed help (`./triangle -h`) describes the content of a file `box.poly` that when called via `./triangle -pqc box.poly` generates the mesh shown in the left panel of Figure III.14 and when called again via `./triangle -pqca0.05 box.1.poly` generates the mesh shown in the right panel of Figure III.14.

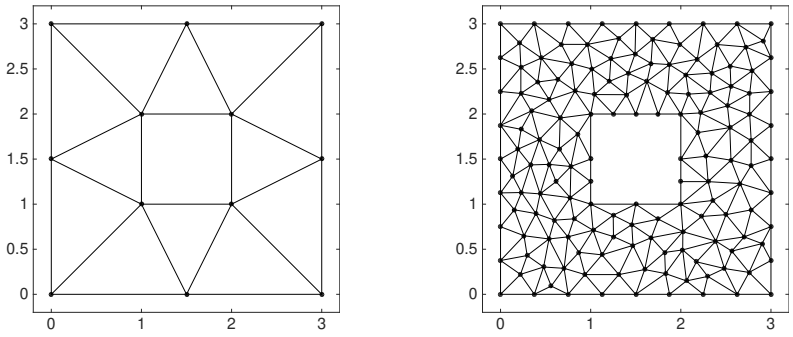


Figure III.13.: **(left)** Delaunay triangulation `box.1.ele/node/poly` and **(right)** `box.2.ele/node/poly` generated by Triangle.

Triangle automatically adds vertices to a given domain specified by points and edges describing its boundaries. It is guaranteed to create quality meshes in the sense that the smallest angle of any triangle is larger then a given lower bound (28.6 degrees but in practice larger). For 3-simplices in 3D the software TetGen has a similar functionality and syntax as Triangle.

Starting with a quality mesh, then uniform mesh refinment is a viable option to obtain finer meshes without regenerating them. Also, if a PDE solution has regions requiring higher resolution, e.g. boundary layers, then this can be accounted for during the generation of the mesh or using *adaptive mesh refinement*, which leads to interesting research.

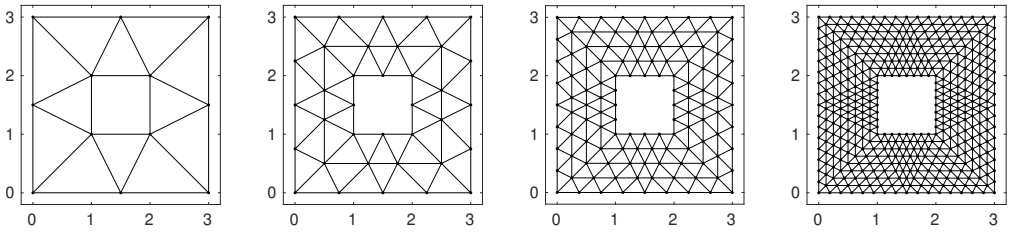


Figure III.14.: Increasing level of uniform refinement of triangulation from left to right.

III.6. FEM – Matrix Assembly

Now we will dive into the details of the construction and different variants to construct the Galerkin matrix. We will restrict our considerations to conformal finite elements for the space $V = H^1(\Omega)$, and in particular to Lagrange elements. Let us summarize the *general recipe of the Galerkin FEM:* Recall the Galerkin method works as follows:

PDE	$Lu = f$ in Ω + b.c.
$\downarrow$	
Variational problem:	Find $u \in V$ s.t. $a(u, v) = F(v)\ \forall v \in V$
$\downarrow$	
Galerkin projection:	Find $u_n \in V_n$ s.t. $a(u_n, v_n) = F(v_n)\ \forall v_n \in V_n$
$\downarrow$	
Assemble the linear system:	Solve $A_n\alpha = f_n$

Then the finite element method tells you how to construct $V_n \subset V$:
Step 1: Find an admissible decomposition of Ω by

$$\overline{\Omega} = \bigcup_{l=1}^{n_e} \overline{\Omega}_l \quad \text{with } n_e \text{ “elements" } \Omega_l \subset \Omega.$$

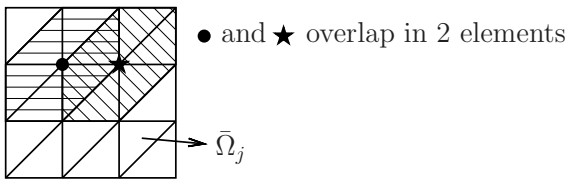
Step 2: Determine suitable basis of functions $(\varphi_i)_{i=1}^n$ for V_n such that

i) Basis functions are continuous $\varphi_i \in C^0(\overline{\Omega})$ for $i = 1, \ldots, n$ so that $V_n \subset V = H^1(\Omega)$;

ii) $\varphi_i|_{\Omega_l}$ is a polynomial for every $i = 1, \ldots, n$ (basis function), $l = 1, \ldots, n_e$ (element);

iii) There exists a set of “nodes" $\{z_i\}_{i=1}^n \subset \Omega$ (e. g. the vertices of the triangles), such that we have the defining relation for Lagrange elements $\varphi_i(z_j) = \delta_{ij}$.

iv) If $z_i \notin \overline{\Omega}_j$, then $\varphi_i|_{\overline{\Omega}_j} \equiv 0$. This ensures that the stiffness matrix is sparse.



Step 3: Construct V_n: Let Γ_D denote the part $\partial\Omega$ with (zero) Dirichlet boundary conditions. Then we set

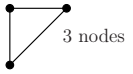
$$V_n = \operatorname{span}\{\varphi_i \mid z_i \notin \Gamma_D\}.$$

The space V_n is the linear span of all basis functions whose associated node is not in Γ_D. Altogether this gives the *Galerkin-FEM*.

Some standard examples of Lagrange finite elements in 2D:

- *linear finite elements* on a triangle

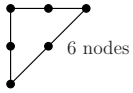
$$\varphi_i|_{\Omega_j}(x_1, x_2) = a_0^j + a_1^j x_1 + a_2^j x_2$$



  degree of freedom of one element: 3

- *quadratic finite elements* on a triangle

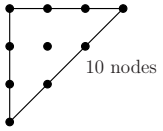
$$\varphi_i|_{\Omega_j}(x_1, x_2) = a_0^j + a_1^j x_1 + a_2^j x_2 + a_3^j x_1 x_2 + a_4^j x_1^2 + a_5^j x_2^2$$



  degree of freedom of one element: 6

- *cubic finite elements* on a triangle

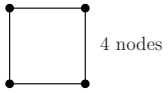
$$\varphi_i|_{\Omega_j}(x_1, x_2) = a_0^j + a_1^j x_1 + \ldots + a_5^j x_2^2 + a_6^j x_1^3 + a_7^j x_1^2 x_2 + a_8^j x_1 x_2^2 + a_9^j x_2^3$$



  degree of freedom of one element: 10

- *bilinear finite elements* on a quadrilateral: fix one coordinate as a coefficient $\rightsquigarrow$ bilinear form w.r.t. the other coordinate

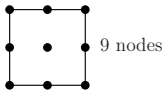
$$\varphi_i|_{\Omega_j}(x_1, x_2) = a_0^j + a_1^j x_1 + a_2^j x_2 + a_3^j x_1 x_2$$



  degree of freedom of one element: 4

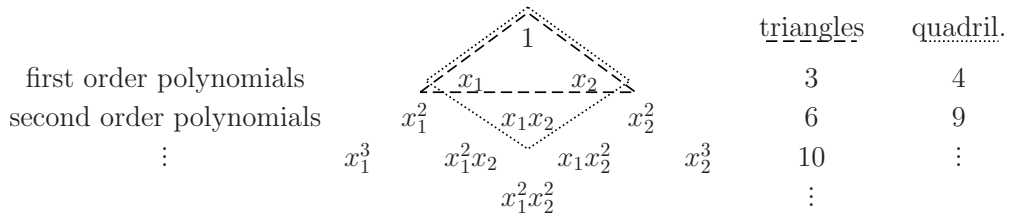
- *biquadratic finite elements* on a quadrilateral: fix one coordinate $\rightsquigarrow$ quadratical function w.r.t. the other coordinate

$$\varphi_i|_{\Omega_j}(x_1, x_2) = a_0^j + a_1^j x_1 + a_2^j x_2 + a_3^j x_1 x_2 + a_4^j x_1^2 + a_5^j x_2^2 + a_6^j x_1^2 x_2 + a_7^j x_1 x_2^2 + a_8^j x_1^2 x_2^2$$



  degree of freedom of one element: 9

Schematic:



The class of all these basis functions are called *Lagrange* basis functions. The name comes from polynomial interpolation, in which the function values at certain positions p_k indicated by the black dots in the elements above are given.

Remark III.34: A further class of basis functions is given by *Hermite elements*. For these elements, not only the function values but also the values of some derivatives are specified. This class is used for example if we need that $\varphi_i \in C^1(\overline{\Omega})$ (not just $C^0(\overline{\Omega})$).

Remark III.35: How do we assemble the Galerkin matrix (stiffness matrix) in two spatial dimensions? Recall the Galerkin matrix is

$$A_n = [a(\varphi_j, \varphi_i)]_{i,j=1}^n ,$$

which we would like to compute for the Poisson problem, as an example. The actual computation of the coefficients $a_{ij} = a(\varphi_j, \varphi_i)$ is usually done as follows. We have

$$a_{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x = \sum_{l=1}^{n_e} \int_{\Omega_l} \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x.$$

where the sum follows from the fact that $\Omega = \bigcup_l \Omega_l$. Thus it is enough to compute $a_{ij}^l := \int_{\Omega_l} \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x$. By construction, a_{ij}^l is zero if $z_i \notin \overline{\Omega}_l$ or $z_j \notin \overline{\Omega}_l$. For the remaining cases we go back to a reference triangle.

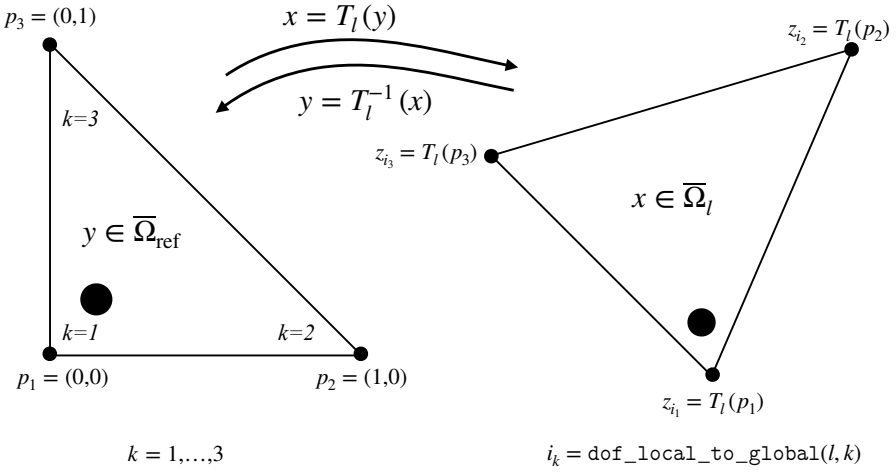


Figure III.15.: Mapping T_l from the reference element $\overline{\Omega}_{\text{ref}}$ to the element Ω_l, where the black disc indicates the location of the first point $k = 1$ or i_1 and reference points p_k and mapped points z_{i_k} are indicated for $k = 1, 2, 3$.

Here $T_l : \overline{\Omega}_{\text{ref}} \to \overline{\Omega}_l$ is an affine linear mapping on triangles with

$$T_l\left(p_1 = \begin{bmatrix}0\\0\end{bmatrix}\right) = z_i, \quad T_l\left(p_2 = \begin{bmatrix}1\\0\end{bmatrix}\right) = z_k, \quad T_l\left(p_3 = \begin{bmatrix}0\\1\end{bmatrix}\right) = z_j.$$

We want to map a point $y = \left[\begin{smallmatrix}y_1\\y_2\end{smallmatrix}\right] \in \overline{\Omega}_{\text{ref}}$ on the *reference triangle* (bijectively) to a point $x = \left[\begin{smallmatrix}x_1\\x_2\end{smallmatrix}\right] \in \overline{\Omega}$.

A short computation shows that $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] = T_l\left(\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]\right) = z_i + (z_j - z_i)y_1 + (z_k - z_i)y_2$ realizes such a mapping by an affine linear transformation. In the end, we use the changed coordinates to compute a_{ij}^l on the reference triangle as follows: With

$$z_i = \begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix}, \quad z_j = \begin{bmatrix} z_{j1} \\ z_{j2} \end{bmatrix}, \quad z_k = \begin{bmatrix} z_{k1} \\ z_{k2} \end{bmatrix},$$

we get

$$\underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = T_l(y) = \underbrace{\begin{bmatrix} z_{j1} - z_{i1} & z_{k1} - z_{i1} \\ z_{j2} - z_{i2} & z_{k2} - z_{i2} \end{bmatrix}}_{=:F_l} \underbrace{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_{y} + \underbrace{\begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix}}_{z_i}. \tag{III.18}$$

Assume that there are piecewise polynomial ansatz functions $\widehat{\varphi}_k$ for $k = 1, ..., m$ defined on the reference triangle Ω_{ref}. Then we get

$$\varphi_i(x) = \varphi_i(T_l(y)) = \widehat{\varphi}_k(y), \qquad \text{for } k = 1, ..., m, \quad i = \texttt{dof\_local\_to\_global}(l, k),$$

where $\texttt{dof\_local\_to\_global}(l, k)$ maps the index k of the *local* reference test function on the element l to the corresponding *global* index i of the degree of freedom (dof). As the book-keeping behind this mapping is absolutely essential for the matrix assembly, we explain the concept for two examples. Piecewise linear and piecewise quadratic finite element basis functions in 2D. The generalization to other basis functions or other spatial dimensions will be more or less straighforward to do. In Figure III.17 the enumeration of elements and corresponding global basis functions is shown. On the reference triangle $\overline{\Omega}_{\text{ref}} = \{(y_1, y_2) \in \mathbb{R}^2 : y_1, y_2 \geq 0, y_1 + y_2 \leq 1\}$ we have the three basis functions

$$\begin{aligned} \widehat{\varphi}_1(y_1, y_2) &= (1 - y_1 - y_2), & p_1 &= (0, 0), \\ \widehat{\varphi}_2(y_1, y_2) &= y_1, & p_2 &= (1, 0), \\ \widehat{\varphi}_3(y_1, y_2) &= y_2, & p_3 &= (0, 1), \end{aligned}$$

so that $\widehat{\varphi}_k(p_s) = \delta_{ks}$ for the corresponding local points $p_s \in \overline{\Omega}_{\text{ref}}$ and $k, s = 1, ..., 3$. This relates local points to the global points $z_i = T_l(p_k)$ for i=$\texttt{dof\_local\_to\_global}(l, k)$ and local basis functions to global basis functions as $\varphi_i(z_j) = \delta_{ij}$ for the global points associated to the global degrees of freedom. How this mapping of degrees of freedom works for P_1 finite elements, where degrees of freedom are associated with z_i being all vertices of the mesh is shown in Figure III.17 and Table III.1. Note, the first (local) basis function on each element is associated with the vertex and global basis function denoted (arbitrarily) with the dot in the element. Second and third (local) basis function then follow in anticlockwise direction for each element.

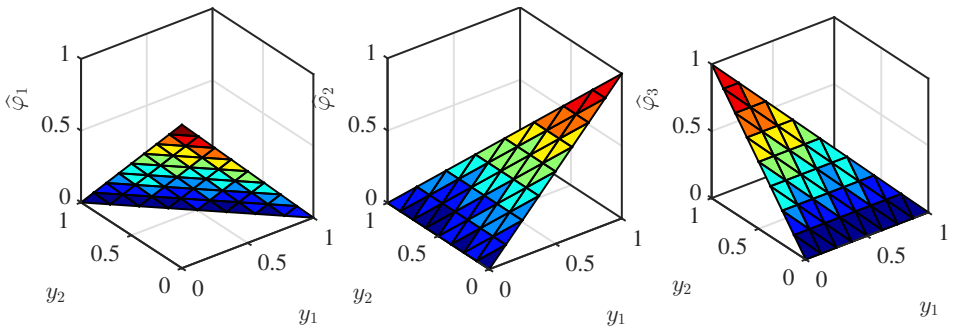


Figure III.16.: P_1 basis functions $\widehat{\varphi}_k$ on the reference triangle.

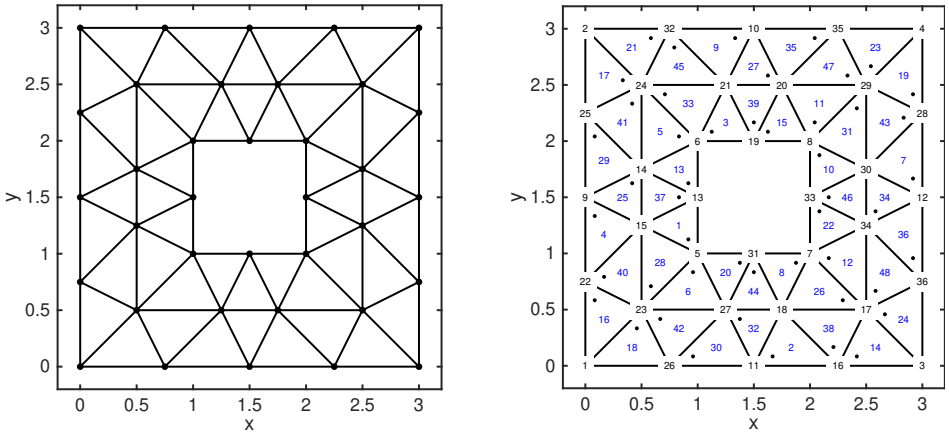


Figure III.17.: **(left)** Delaunay mesh **(right)** with enumeration of elements (black) and of vertices $\sim$ degrees of freedom (blue). The dot in each element shows the first local basis function, the other basis functions are obtained by cycling through vertices in anticlockwise direction.

`dof_local_to_global`(l,k)	$k=1$	$k=2$	$k=3$
$l=1$	5	13	15
$l=2$	11	16	18
$l=3$	6	19	21
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$l=46$	33	34	30
$l=47$	29	35	20
$l=48$	17	36	34

Table III.1.: Mapping of the 3 local reference basis functions $\widehat{\varphi}_k(y)$ of the lth element to global basis function $i =$`dof_local_to_global`(l,k) for P_1 finite elements for the 2D mesh shown in Figure III.17.

For P_2 functions we have the six local basis functions on the reference

$$\begin{aligned}
\widehat{\varphi}_1(y_1, y_2) &= (1 - y_1 - y_2)(1 - 2y_1 - 2y_2), & p_1 &= (0, 0) \\
\widehat{\varphi}_2(y_1, y_2) &= y_1(2y_1 - 1), & p_2 &= (1, 0) \\
\widehat{\varphi}_3(y_1, y_2) &= y_2(2y_2 - 1), & p_3 &= (0, 1) \\
\widehat{\varphi}_4(y_1, y_2) &= 4y_1(1 - y_1 - y_2), & p_4 &= (1/2, 0) \\
\widehat{\varphi}_5(y_1, y_2) &= 4y_1 y_2, & p_5 &= (1/2, 1/2) \\
\widehat{\varphi}_6(y_1, y_2) &= 4y_2(1 - y_1 - y_2), & p_6 &= (0, 1/2)
\end{aligned}$$

giving exactly the same relation $\widehat{\varphi}_s(p_k) = \delta_{sk}$ for the local points $p_k \in \overline{\Omega}_{\text{ref}}$ and $s, k = 1, ..., 6$ as given above. The corresponding mapping and the support points of the degrees of freedom are shown in Figure III.19. In Figure III.20 we show how this produces a global distribution of degrees of freedom for a full 2D Delaunay mesh. Finally, in Table III.2 the corresponding mapping `dof_local_to_global` from local to global degrees of freedom is shown for certain selected elements.

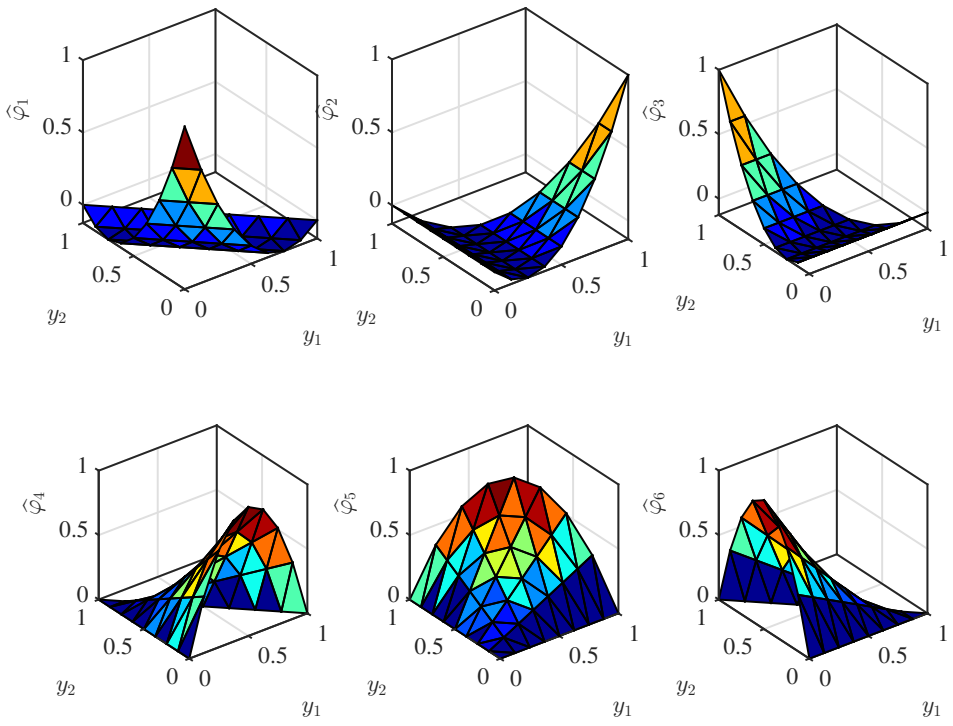


Figure III.18.: P_2 basis functions $\widehat{\varphi}_k$ on the reference triangle.

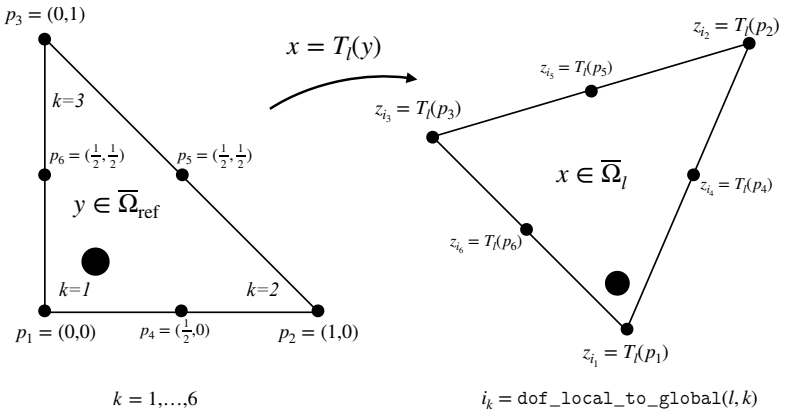


Figure III.19.: Mapping T_l from the reference element $\overline{\Omega}_{\text{ref}}$ to the element Ω_l, where the black disc indicates the location of the first point $k = 1$ or i_1 and reference points p_k and mapped points z_{i_k} are indicated for $k = 1, ..., 6$.

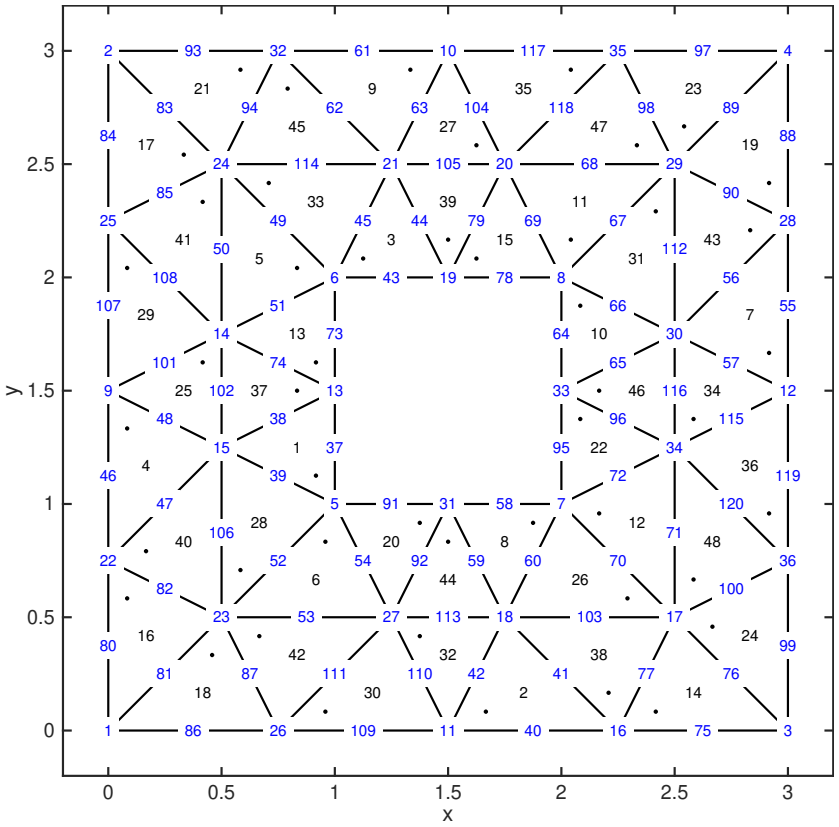


Figure III.20.: 2D Delaunay mesh with enumeration of elements (black) and of edges + vertices ∼ global basis functions ∼ degrees of freedom (blue).

dof_local_to_global(l,k)	$k_\bullet = 1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
$l=1$	5	13	15	37	38	39
$l=2$	11	16	18	40	41	42
$l=3$	6	19	21	43	44	45
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$l=46$	33	34	30	96	116	65
$l=47$	29	35	20	98	118	68
$l=48$	17	36	34	100	120	71

Table III.2.: Mapping of the 6 local reference basis functions $\widehat{\varphi}_k(y)$ of the lth element to global basis function $i =$dof_local_to_global(l,k) for P_2 finite elements for the 2D mesh shown in Figure III.20.

Now we want to do the change of coordinates in the integrals, in particular, we must change the coordinates when forming the gradients. According to the chain rule in multivariate differentiation, we obtain

$$\begin{aligned}\nabla_x\varphi(x) := \begin{bmatrix}\frac{\partial\varphi(x)}{\partial x_1}\\ \frac{\partial\varphi(x)}{\partial x_2}\end{bmatrix} &= \begin{bmatrix}\frac{\partial\varphi(T_l(y))}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial\varphi(T_l(y))}{\partial y_2}\frac{\partial y_2}{\partial x_1}\\ \frac{\partial\varphi(T_l(y))}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial\varphi(T_l(y))}{\partial y_2}\frac{\partial y_2}{\partial x_2}\end{bmatrix}\\ &= \begin{bmatrix}\frac{\partial\widehat{\varphi}(y)}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial\widehat{\varphi}(y)}{\partial y_2}\frac{\partial y_2}{\partial x_1}\\ \frac{\partial\widehat{\varphi}(y)}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial\widehat{\varphi}(y)}{\partial y_2}\frac{\partial y_2}{\partial x_2}\end{bmatrix}\\ &= \underbrace{\begin{bmatrix}\frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1}\\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2}\end{bmatrix}}_{=:(\nabla_y T_l)^{-T}}\underbrace{\begin{bmatrix}\frac{\partial\widehat{\varphi}(y)}{\partial y_1}\\ \frac{\partial\widehat{\varphi}(y)}{\partial y_2}\end{bmatrix}}_{=:\nabla_y\widehat{\varphi}(y)}.\end{aligned}$$

for each $\widehat{\varphi}(y) = \widehat{\varphi}_k(y)$. Here we introduced the Jacobian F_l of the transformation

$$\nabla_y T_l(y) = \begin{bmatrix}\frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2}\\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2}\end{bmatrix} =: F_l \overset{\text{(III.18)}}{=} \begin{bmatrix}z_{j1}-z_{i1} & z_{k1}-z_{i1}\\ z_{j2}-z_{i2} & z_{k2}-z_{i2}\end{bmatrix}. \tag{III.19}$$

Since the transformation is piecewise linear $x = T_l(y) = F_l y + z_i$, the Jacobian is constant on $\overline{\Omega}_{\text{ref}}$. Furthermore, since $y = F_l^{-1}(x - z_i)$, we see that using (III.18) we have

$$(\nabla_y T_l)^{-T} = \begin{bmatrix}\frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1}\\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2}\end{bmatrix} = F_l^{-\mathsf{T}} = \frac{1}{\det F_l}\begin{bmatrix}z_{k2}-z_{i2} & -(z_{j2}-z_{i2})\\ -(z_{k1}-z_{i1}) & z_{j1}-z_{i1}\end{bmatrix}$$

with $\det F_l = (z_{j1}-z_{i1})(z_{k2}-z_{i2}) - (z_{k1}-z_{i1})(z_{j2}-z_{i2})$. For the gradient of a function φ in the reference this gives

$$\nabla_x\varphi(x) = F_l^{-\mathsf{T}}\nabla_y\widehat{\varphi}(y). \tag{III.20}$$

This expression can be generalized to nonlinear transformations $T_l(y)$, in which case $F_l(y) = \nabla_y T_l(y) \in \mathbb{R}^{n\times n}$ will in general still depend on $y \in \overline{\Omega}_{\text{ref}}$. Nonlinear (or higher

order polynomial) transformations can be useful, if we want to approximate domains with smooth curved boundaries more accurately. The transformation formula for coordinate changes (III.20) in the integration gives us

$$a_{ij}^l = \int_{\Omega_l} \nabla_x \varphi_i(x) \cdot \nabla_x \varphi_j(x) \,\mathrm{d}x = \int_{\Omega_{\mathrm{ref}}} \left(F_l^{-\mathsf{T}} \nabla_y \widehat{\varphi}_{k_i}(y)\right) \cdot \left(F_l^{-\mathsf{T}} \nabla_y \widehat{\varphi}_{k_j}(y)\right) |\det F_l| \,\mathrm{d}y.$$

Remark III.36 (Simplified assembly)**:** For piecewise linear mappings and simple bilinear forms, most parts of the matrix a_{ij}^l can be precomputed and constructed in an even simpler way. Using the constant symmetric matrix

$$B_l = \begin{bmatrix} b_{11}^l & b_{12}^l \\ b_{12}^l & b_{22}^l \end{bmatrix} := F_l^{-1} F_l^{-\mathsf{T}} \in \mathbb{R}^{2\times 2}$$

we can rewrite the matrix a_{ij}^l as

$$\begin{aligned} a_{ij}^l &= \int_{\Omega} \nabla_x \varphi_i(x) \cdot \nabla_x \varphi_j(x) \,\mathrm{d}x = \int_{\Omega_{\mathrm{ref}}} \left(F_l^{-\mathsf{T}} \nabla_y \widehat{\varphi}_{k_i}(y)\right) \cdot \left(F_l^{-\mathsf{T}} \nabla_y \widehat{\varphi}_{k_j}(y)\right) |\det F_l| \,\mathrm{d}y \\ &= \int_{\Omega_{\mathrm{ref}}} \left(\nabla_y \widehat{\varphi}_{k_i}(y)\right)^{\mathsf{T}} \underbrace{F_l^{-1} F_l^{-\mathsf{T}}}_{B_l} \left(\nabla_y \widehat{\varphi}_{k_j}(y)\right) |\det F_l| \,\mathrm{d}y \\ &= |\det F_l| \Big(b_{11}^l \int_{\Omega_{\mathrm{ref}}} \frac{\partial \widehat{\varphi}_{k_i}(y)}{\partial y_1} \frac{\partial \widehat{\varphi}_{k_j}(y)}{\partial y_1} \,\mathrm{d}y + b_{22}^l \int_{\Omega_{\mathrm{ref}}} \frac{\partial \widehat{\varphi}_{k_i}(y)}{\partial y_2} \frac{\partial \widehat{\varphi}_{k_j}(y)}{\partial y_2} \,\mathrm{d}y \quad + \\ &\qquad b_{12}^l \int_{\Omega_{\mathrm{ref}}} \frac{\partial \widehat{\varphi}_{k_i}(y)}{\partial y_1} \frac{\partial \widehat{\varphi}_{k_j}(y)}{\partial y_2} + \frac{\partial \widehat{\varphi}_{k_i}(y)}{\partial y_2} \frac{\partial \widehat{\varphi}_{k_j}(y)}{\partial y_1} \,\mathrm{d}y \Big). \end{aligned}$$

Note, as before the mapping of local k_i to global i degrees of freedom is facilitated using $i =$`dof_local_to_global`(l, k_i) for each element l and basis function $k_i = 1, ..., n_\varphi$ in the reference domain Ω_{ref}. Each of the integrals above on the reference triangle needs to be computed once and only the values $\det F_l$, b_{11}^l, b_{12}^l, and b_{22}^l need to be computed for each element separately. This computation requires the knowledge of the vertices and the connectivity of the elements Ω_l. For linear finite elements in 1D/2D/3D we have $n_\varphi = 2/3/4$, for quadratic finite elements in 1D/2D/3D we have $n_\varphi = 3/6/10$.

However, in practise these integrals are often computed using numerical quadrature such as Gauss quadrature on intervals, triangles, tetrahedrons. This is particularly true

- for right sides f of the equation,
- for problems with nonlinear transformations $T_l(y)$,
- for nonlinear PDEs,
- for PDEs with space-dependent coefficients,

where such a simplification might not be possible or result in cluttered code.

We indicate a few simple quadrature rules for triangles in two dimensions. First of all, for a given function on the reference triangle $f : \Omega_{\text{ref}} \to \mathbb{R}$ Gauss integration is about approximating the integral over the domain by

$$\int_{\Omega_{\text{ref}}} f(y)\, \mathrm{d}y = A \sum_{q=1}^{n_q} w_q f(y^q) \tag{III.21}$$

where $A = \frac{1}{2}$ is the area of the reference triangle, w_q are the Gaussian weights, and y_q are the locations of the corresponding Gaussian points for $q = 1, \ldots, n_q$. Such an approximation is of order p, if it is exact for polynomials to order p. Then the strategy is to store the values of $\widehat{\varphi}_k(y)$ and $\partial\widehat{\varphi}_k(y)/\partial y_k$ at the corresponding Gaussian points $y^q = (y_1^q, y_2^q) \in \Omega_{\text{ref}}$. In the Table III.3 some example for Gauss quadrature rules on triangles are given. Examples for Gauss quadrature in one spatial dimensions are easily accessible at https://en.wikipedia.org/wiki/Gaussian_quadrature. Two typical examples for integrals in a finite element program are the so-called stiffness matrix

$$\begin{aligned} a_{ij}^l &= \int_{\Omega_l} \nabla_x \varphi_i(x) \cdot \nabla_x \varphi_j(x)\, \mathrm{d}x \\ &= \int_{\Omega_{\text{ref}}} \left(F_l^{-\mathsf{T}} \nabla_y \widehat{\varphi}_{k_i}(y)\right) \cdot \left(F_l^{-\mathsf{T}} \nabla_y \widehat{\varphi}_{k_j}(y)\right) |\det F_l|\, \mathrm{d}y \\ &= \sum_{q=1}^{n_q} \left(F_l^{-\mathsf{T}} \nabla_y \widehat{\varphi}_{k_i}(y^q)\right) \cdot \left(F_l^{-\mathsf{T}} \nabla_y \widehat{\varphi}_{k_j}(y^q)\right) |\det F_l(y^q)|\, \frac{w_q}{2}, \end{aligned} \tag{III.22}$$

and the mass matrix

$$\begin{aligned} m_{ij}^l &= \int_{\Omega_l} \varphi_i(x)\varphi_j(x)\, \mathrm{d}x = \int_{\Omega_{\text{ref}}} \widehat{\varphi}_{k_i}(y)\widehat{\varphi}_{k_j}(y) |\det F_l|\, \mathrm{d}y \\ &= \sum_{q=1}^{n_q} \widehat{\varphi}_{k_i}(y^q)\widehat{\varphi}_{k_j}(y^q)\, |\det F_l(y^q)|\, \frac{w_q}{2}. \end{aligned} \tag{III.23}$$

Higher dimensional Gauss integration is entirely analogous to the procedure in two dimensions. Often, the Gauss points are rather specified using their barycentric coordinates in the triangle. Also, the generalization to space dependent coefficients just requires the insertion of another function into the integral, i.e., $a_{ij}^l = \int_{\Omega_l} a(x) \nabla\varphi_i \cdot \nabla\varphi_j(x)\, \mathrm{d}x$. Boundary integrals, i.e., integrals over edges or in general faces, are computed analogously using Gauss integration but require a little extra attention concerning the mapping of local to global degrees of freedom. For convection-diffusion problem we require an extra term $\int_{\Omega_l} \varphi_i \mathbf{u} \cdot \nabla\varphi_j\, \mathrm{d}x$ in the bilinear form, which is integrated analogously. As opposed to finite difference methods, one would not assemble boundary conditions explicitly into the Galerkin matrix. In order to treat inhomogeneous Dirichlet boundary conditions, one assumes that an arbitrary function $u_0 \in V$ satisfies the inhomogeneous boundary conditions and then solve for $u = u_0 + \bar{u}$ using the problem

$$a(\bar{u}, v) = \widehat{f}(v), \qquad \widehat{f}(v) = f(v) - a(u_0), \tag{III.24}$$

where $\bar{u}, v$ satisfy homogeneous Dirichlet boundary conditions. One way to assemble these matrices is to build the Galerkin matrix for all degrees of freedom (also on the Dirichlet boundary) first, then compute $\widehat{f}$ and reduce the matrices to the *active* degrees of freedom by removing all degrees of freedom associated to the Dirichlet boundary.

n_q	1	3			4			
q	1	1	2	3	1	2	3	4
y_1^q	1/3	1/6	2/3	1/6	1/3	1/5	1/5	3/5
y_2^q	1/3	1/6	1/6	2/3	1/3	3/5	1/5	1/5
w_q	1	1/3	1/3	1/3	-27/48	25/96	25/96	25/96

n_q	6					
q	1	2	3	4	5	6
y_1^q	r_1	r_1	$1-r_1$	r_2	r_2	$1-2r_2$
y_2^q	r_1	$1-2r_1$	r_1	r_2	$1-2r_2$	r_2
w_q	w_1	w_1	w_1	w_2	w_2	w_2

n_q	7						
q	1	2	3	4	5	6	7
y_1^q	ξ_1	ξ_2	ξ_3	ξ_2	ξ_4	ξ_5	ξ_4
y_2^q	ξ_1	ξ_3	ξ_2	ξ_2	ξ_5	ξ_4	ξ_4
w_q	$\widehat{w}_1$	$\widehat{w}_2$	$\widehat{w}_2$	$\widehat{w}_2$	$\widehat{w}_3$	$\widehat{w}_3$	$\widehat{w}_3$

Table III.3.: Symmetric Gauss integration on triangles for $n_q = 1, 3, 4, 6, 7$ points with order 1,2,3,4,5. For details see [Cow73, Dun85, Wal00, Fel04]. The corresponding weights and positions are given below in (III.25) and (III.26).

$$\begin{aligned}
&\widehat{w}_1 = 9/40, \\
&\widehat{w}_2 = \tfrac{1}{1200}(155+\sqrt{15}), \qquad \widehat{w}_3 = \tfrac{1}{1200}(155-\sqrt{15}) \\
&\xi_1 = 1/3, \\
&\xi_2 = \tfrac{1}{21}(6+\sqrt{15}), \qquad \xi_3 = \tfrac{1}{21}(9-2\sqrt{15}), \\
&\xi_4 = \tfrac{1}{21}(6-\sqrt{15}), \qquad \xi_5 = \tfrac{1}{21}(9+2\sqrt{15}).
\end{aligned} \tag{III.25}$$

$$\begin{aligned}
r_1 &= \tfrac{1}{18}(8-\sqrt{10}+\sqrt{38-44\sqrt{\tfrac{2}{5}}}), \quad w_1 = \tfrac{1}{3720}(620+\sqrt{213125-53320\sqrt{10}}), \\
r_2 &= \tfrac{1}{18}(8-\sqrt{10}-\sqrt{38-44\sqrt{\tfrac{2}{5}}}), \quad w_2 = \tfrac{1}{3720}(620-\sqrt{213125-53320\sqrt{10}}),
\end{aligned} \tag{III.26}$$

III.7. FEM – Neumann Boundary conditions

Let $\Omega \subset \mathbb{R}^2$, $\Gamma = \Gamma_D \bigcup \Gamma_N$, $f \in L^2(\Omega)$ and $h \in L^2(\Gamma_N)$. Consider the boundary value problem

$$\begin{cases} -\Delta u = f \quad \text{in} \quad \Omega, \\ u = 0 \quad \text{on} \quad \Gamma_D, \\ \partial_\nu u = h \quad \text{on} \quad \Gamma_N. \end{cases} \tag{III.27}$$

With the test space $V = \{w \in H^1(\Omega) | w|_{\Gamma_D} = 0\}$ the weak form of (III.27) is given by

$$\int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega f v dx + \int_{\Gamma_N} h v ds \quad \forall v \in V. \tag{III.28}$$

Let $V_n = \text{span}\{\varphi_i | i = 1, ..., n\} \subset V$. After the Galerkin projection the weak form reads

$$\int_\Omega \nabla u_n \cdot \nabla v_n dx = \int_\Omega f v_n dx + \int_{\Gamma_N} h v_n ds \quad \forall v_n \in V_n. \tag{III.29}$$

Substituting the Galerkin ansatz $u_n = \sum_{i=1}^n \alpha_i \varphi_i$ implies the linear equation system

$$A\alpha = Mf + \eta \tag{III.30}$$

with $A_{ij} = \int_\Omega \nabla \varphi_i \cdot \nabla \varphi_j dx$, $M = \int_\Omega \varphi_i \varphi_j$, $f_j = f(x_j)$ and $\eta_j = \int_{\Gamma_N} h \varphi_j ds$. In the previous chapters we discussed the method for building the matrices A and M via the elements matrix `e`. In this section it is derived a representation for the vector η.
Let $n_{\text{fN}} \in \mathbb{N}$ be the number of faces on the Neumann boundary. The Neumann boundary can be written as $\Gamma_N = \bigcup_{k=1}^{n_{\text{fN}}} \Gamma_N^k$. Therefore it holds $\eta_j = \sum_{k=1}^{n_{\text{fN}}} \eta_j^k$ with

$$\eta_j^k = \int_{\Gamma_N^k} h \varphi_j ds. \tag{III.31}$$

In 1D FEM we used the reference interval and in 2D FEM we made use of the reference triangle. In 2D FEM the boundary is one-dimensional. Therefore we use the reference interval $\widehat{\Gamma}_N^{\text{ref}} = [0, 1]$. In 1D FEM we introduced a linear map which maps the reference interval to the i-th element of the discretisation. Here we use a map which maps the reference interval $\widehat{\Gamma}_N^{\text{ref}}$ to the k-th element of the Neumann boundary Γ_N^k. Let $z_{\text{fN}}(k, m)$ for $m = 1, 2$, $k = 1, ..., n_{\text{fN}}$ denote the coordinates of the boundary points of element Γ_N^k. The linear map which maps the reference interval to the k-th element of the Neumann boundary now reads

$$S_k \colon \widehat{\Gamma}_N^{\text{ref}} \to \Gamma_N^k, \quad S_k(t) = z_{\text{fN}}(k, 1) + (z_{\text{fN}}(k, 2) - z_{\text{fN}}(k, 1))t. \tag{III.32}$$

On the reference interval the basis functions are given by $\psi_1(t) = 1 - t$ and $\psi_2(t) = t$. As the functions h and φ_j are scalar and the boundary Γ_N^k is one-dimensional, the integral

(III.31) is a scalar line integral for the function $h\varphi_j$ along the curve Γ_N^k parametrized by S_k. Substituting the definition of a scalar line integral implies

$$\eta_j^k = \int_{\Gamma_N^k} h(x)\varphi_j(x)ds \tag{III.33}$$

$$= \int_0^1 h\left(S_k(t)\right)\varphi_j\left(S_k(t)\right)||S_k'(t)||dt \tag{III.34}$$

$$= \int_0^1 h\left(S_k(t)\right)\varphi_j\left(S_k(t)\right)||z_{\mathrm{fN}}(k,2) - z_{\mathrm{fN}}(k,1)||dt \tag{III.35}$$

$$= \int_0^1 h\left(S_k(t)\right)\varphi_j\left(S_k(t)\right)\mathrm{len}(\Gamma_N^k)dt. \tag{III.36}$$

Consider a matrix $\mathrm{fN} \in \mathbb{N}^{n_{\mathrm{fN}},2}$. Let the entries of the k-th row $\mathrm{fN}(k,:)$ be the nodes of the boundary element Γ_N^k. Then it holds $\varphi_j\left(S_k(t)\right) = \psi_m(t)$ for $j = \mathrm{fN}(k,m)$. The integral then reads

$$\eta_j^k = \mathrm{len}(\Gamma_N^k)\int_0^1 h(S_k(t))\psi_m(t)dt, \quad j = \mathrm{fN}(k,m). \tag{III.37}$$

Take for example a constant function $h = h_0 \in \mathbb{R}$ on the Neumann boundary Γ_N. The integral is given by

$$\eta_j^k = \mathrm{len}(\Gamma_N^k)h_0\int_0^1 \psi_m(t)dt \tag{III.38}$$

$$= \frac{1}{2}\mathrm{len}(\Gamma_N^k)h_0\begin{pmatrix}1\\1\end{pmatrix}_m, \quad j = \mathrm{fN}(k,m). \tag{III.39}$$

For each element number of the Neumann boundary k you are interested in the position j of the entry m of the 2×1 vector in (III.39). Consider Figure III.17 at page 117. Take a look at element 4 with the nodes 9 and 22. In the picture there is no enumeration of the boundary elements. Lets declare it as element 3 of the Neumann boundary. Thus it holds $\mathrm{fN}(3,:) = [9,22]$. Therefore the position of entry one of the 2×1 vector is $j = 9$ and entry two is $j = 22$.

III.8. FEM – Analysis

In this section some strategies to analyze the Galerkin FEM are shown. First we fix the mathematical framework. The underlying variational problem is

$$\text{Find } u \in V \text{ such that } a(u,v) = F(v) \text{ for all } v \in V. \tag{V}$$

Here we impose the following assumption.

Assumption III.37: Let $\Omega \subset \mathbb{R}^d$ be a bounded polygonal domain and set $V = H_0^1(\Omega)$ (for Dirichlet boundary conditions) with norm $\|u\|_V = \|u\|_{H^1(\Omega)}$. Further, the bilinear form $a : V \times V \to \mathbb{R}$ and the linear form $F : V \to \mathbb{R}$ satisfy

a) There exists an $\alpha \in (0,\infty)$ such that $|a(v_1,v_2)| \leq \alpha \|v_1\|_V \|v_2\|_V \; \forall\, v_1, v_2 \in H_0^1(\Omega)$.
Then we call a a "continuous/bounded bilinear form".

b) There exists a $\beta \in (0,\infty)$ such that $a(v,v) \geq \beta \|v\|_V^2 \; \forall\, v \in H_0^1(\Omega)$.
Then we say a is "V-elliptic"/"coercive".

c) There exists a $\gamma \in (0,\infty)$ such that $|F(v)| \leq \gamma \|v\|_V \; \forall\, v \in H_0^1(\Omega)$.
Then we call F a "continuous linear form"/"bounded linear form".

Under these conditions we have the following existence theorem:

Theorem III.38 (Theorem of Lax-Milgram)**:** Under Assumption III.37 there exists a unique solution of the variational problem (V) and we have the estimate

$$\|u\|_V \leq \frac{\gamma}{\beta}.$$

Note that a does not have to be a symmetric bilinear form.

Example III.39 (More general elliptic operators)**:** Let $\Omega \subset \mathbb{R}^d$ be as above. Consider

$$Lu(x) = -\nabla \cdot (a(x)\nabla u(x)) + c(x)u(x),$$

where a, $c : \Omega \to \mathbb{R}$ are bounded and smooth functions with

$$a(x) \geq a_0 > 0 \text{ and } c(x) \geq c_0 \geq 0 \text{ for all } x \in \Omega.$$

Recall that by "$\cdot$" we denote the usual inner product in the Euclidean space $\mathbb{R}^d$. The bilinear form is then

$$a(v_1,v_2) = \int_\Omega a(x)\nabla v_1(x) \cdot \nabla v_2(x) + c(x)v_1(x)v_2(x)\,\mathrm{d}x,$$

i. e., here the bilinear form a is symmetric. Now we check that the Assumptions in III.37 are indeed satisfied:

a) Continuity of a: It holds that

$$\begin{aligned} |a(v_1,v_2)| &\leq \int_\Omega |a||\nabla v_1 \cdot \nabla v_2| + |c||v_1||v_2|\,\mathrm{d}x \\ &\leq \int_\Omega |a|\|\nabla v_1\|\|\nabla v_2\|\,\mathrm{d}x + \int_\Omega |c||v_1||v_2|\,\mathrm{d}x \\ &\leq \|a\|_\infty \int_\Omega \|\nabla v_1\|\|\nabla v_2\|\,\mathrm{d}x + \|c\|_\infty \int_\Omega |v_1||v_2|\,\mathrm{d}x \end{aligned}$$

By the Cauchy-Schwarz inequality[1] it follows

$$|a(v_1,v_2)| \leq \|a\|_\infty \|\nabla v_1\|_{L^2(\Omega)} \|\nabla v_2\|_{L^2(\Omega)} + \|c\|_\infty \|v_1\|_{L^2(\Omega)} \|v_2\|_{L^2(\Omega)}.$$

[1] Cauchy-Schwarz inequality for $L^2(\Omega)$:

$$\left| \int_\Omega fg\,\mathrm{d}x \right| \leq \left(\int_\Omega |f|^2\,\mathrm{d}x \right)^{\frac{1}{2}} \left(\int_\Omega |g|^2\,\mathrm{d}x \right)^{\frac{1}{2}} = \|f\|_{L^2(\Omega)} \|g\|_{L^2(\Omega)}$$

Recall that

$$\|v\|_V = \|v\|_{H^1(\Omega)} = \left(\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}},$$

which gives

$$\|\nabla v\|_{L^2(\Omega)} \le \|v\|_V \quad \text{and} \quad \|v\|_{L^2(\Omega)} \le \|v\|_V,$$

and therefore,

$$|a(v_1, v_2)| \le \underbrace{(\|a\|_\infty + \|c\|_\infty)}_{=:\alpha} \|v_1\|_V \|v_2\|_V.$$

Thus, the bilinear form a is continuous/bounded a).

b) It remains to show the ellipticity of a. For every $v \in H^1_0(\Omega)$ the Poincaré inequality reads

$$\|v\|^2_{L^2(\Omega)} = \int_\Omega |v|^2 \,\mathrm{d}x \le C \int_\Omega \|\nabla v\|^2 \,\mathrm{d}x = C\|\nabla v\|^2_{L^2(\Omega)}$$

for some constant C only depending on the domain Ω.

From this it follows that

$$\|v\|^2_V = \|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} \le (C+1) \int_\Omega \|\nabla v\|^2 \,\mathrm{d}x$$

for all $v \in H^1_0(\Omega)$. This gives the estimate

$$\begin{aligned} a(v,v) &= \int_\Omega \underbrace{a(x)}_{\ge a_0} \underbrace{\nabla v \cdot \nabla v}_{=\|\nabla v\|^2} + \underbrace{c(x)}_{\ge 0} \underbrace{v(x)^2}_{\ge 0} \,\mathrm{d}x \ge \int_\Omega a_0 \|\nabla v\|^2 \,\mathrm{d}x + c_0 \int_\Omega v^2 \,dx \\ &\ge \frac{a_0}{C+1} \|v\|^2_V \quad \text{(by the Poincaré inequality).} \end{aligned}$$

For $c_0 > 0$ we have $a(v,v) \ge \min(a_0, c_0)\|v\|^2_V$ and do not need to use the Poincaré inequality. Thus, the bilinear form a is coercive/elliptic b).

Remark III.40 (Proof of the Poincaré inequality in 1D)**:** Assume f is zero at $x = 0$ and differentiable in Ω. The idea is to use $f(x) = f(0) + \int_0^x f'(y) \,\mathrm{d}y$. Thus, $|f(x)| \le \int_0^x |f'(y)| \,\mathrm{d}y$. Then, taking the square and integrating over x yields

$$\begin{aligned} \int_\Omega |f(x)|^2 \,\mathrm{d}x &\le \int_\Omega \left(\int_0^x |f'(y)| \,\mathrm{d}y \right)^2 \mathrm{d}x \\ &\le \int_\Omega \left(\int_0^x 1 \,\mathrm{d}y \right) \left(\int_0^x |f'(y)|^2 \,\mathrm{d}y \right) \mathrm{d}x \le |\Omega|^2 \|f'\|^2_{L^2(\Omega)}, \end{aligned}$$

where we also applied the Cauchy-Schwarz inequality to the interior integral in the second step. As the argument is basically the same as in the derivation of the stability constant for the Laplace operator in the L^2 norm, the constant is basically given by the smallest eigenvalue as $C = 1/\min\lambda$.

Example III.41 (Elliptic problem with convection)**:** Now we indicate the changes if we include $\mathbf{b} : \Omega \to \mathbb{R}^n$ into the elliptic operator

$$Lu(x) = -\nabla \cdot (a(x)\nabla u(x)) + \mathbf{b}(x) \cdot \nabla u(x) + c(x)u(x),$$

where we assume again $V = H_0^1(\Omega)$. The resulting weak formulation of the problem is $a(u, v) = F(x)$ with the bilinear form defined as

$$a(v_1, v_2) = \int_\Omega a(x)\nabla v_1(x) \cdot \nabla v_2(x)\,\mathrm{d}x \quad + \quad b(v_1, v_2), \tag{III.40}$$

$$b(v_1, v_2) = \int_\Omega c(x)v_1(x)v_2(x) + v_2(x)\mathbf{b}(x) \cdot \nabla v_1(x)\,\mathrm{d}x, \tag{III.41}$$

which due to $\mathbf{b}(x)$ is not symmetric anymore. Lets show the assumptions for the Lax-Milgram theorem and focus on the $b(u, v)$ part, as the other parts are analogous to the previous consideration.

a) Boundedness: We focus on b and have

$$\begin{aligned} |b(v_1, v_2)| &\leq \int_\Omega |b||\nabla v_1 \cdot \nabla v_2| + |c||v_1||v_2|\,\mathrm{d}x \\ &\leq \int_\Omega \|b\|_\infty \|\nabla v_1\||v_1| + \|c\|_\infty |v_1||v_2|\,\mathrm{d}x \leq (\|b\|_\infty + \|c\|_\infty)\|v_1\|_V \|v_2\|_V. \end{aligned}$$

b) Coercivity: Using Gauss and $v = 0$ on $\partial\Omega$ we show

$$\begin{aligned} b(v, v) &= \int_\Omega cv^2 + v\mathbf{b} \cdot \nabla v\,\mathrm{d}x = \int_\Omega cv^2 + \tfrac{1}{2}\left(\nabla \cdot (\mathbf{b}v^2) - v^2(\nabla \cdot \mathbf{b})\right)\,\mathrm{d}x \\ &= \int_\Omega \left(c - \tfrac{1}{2}\nabla \cdot \mathbf{b}\right) v^2\,\mathrm{d}x. \end{aligned}$$

Thus, if we additionally require $2c \geq \nabla \cdot b$ in Ω, then we get coercivity $a(v, v) \geq a_0\|\nabla v\|_{L^2(\Omega)} \geq \frac{a_0}{C+1}\|v\|_V$. For more details see the monograph [Eva98, Chap. 6].

Next, we turn to the error analysis. Let $h \in (0, 1)$ be a parameter determining the refinement level of the finite element spaces (e. g., a smaller h means more degrees of freedom and hence a better approximation; h may be the maximum length of the edges of all triangles). We impose the following assumption on the finite element spaces. This assumption enforces that if h is small then there always exists a $v_h \in V_h$ that is "close" to a given $v \in H^r(\Omega)$. We will discuss this assumption for the standard finite element space further below.

Assumption III.42 (Approximation property)**:** Let $(V_h)_{h\in(0,1)} \subset V$ be a sequence of Galerkin finite element spaces that, for some $r \in \mathbb{N}$ and constant C satisfies

$$\inf_{v_h \in V_h} \|v - v_h\|_{L^2(\Omega)} \leq Ch^r \|v\|_{H^r(\Omega)}$$

and

$$\inf_{v_h \in V_h} \|v - v_h\|_{H^1(\Omega)} \leq Ch^{r-1}\|v\|_{H^r(\Omega)}$$

for all $v \in H^r(\Omega)$. In particular, the constant C does not depend on $v \in H^r(\Omega)$.

It can be shown that the above approximation property is satisfied for finite element spaces consisting of piecewise linear functions. For finite element spaces consisting of higher-degree polynomial finite elements, better approximation properties can often be shown (in the sense that the exponent of h gets larger), if v is sufficiently regular.

Theorem III.43 (Lemma of Céa [LT03])**:** Let Assumptions III.37 and III.42 be satisfied with $r = 1$. Assume that $u \in H^2(\Omega)$ is the exact solution for (V) and let $u_h \in V_h$, $h \in (0, 1)$, be a sequence of FEM approximations. Then

$$\|u_h - u\|_{H^1(\Omega)} \leq Ch^1\|u\|_{H^2(\Omega)}$$

(or in general: $\|u_h - u\|_{H^1(\Omega)} \leq Ch^{r-1}\|u\|_{H^r(\Omega)}$).

Proof. By the ellipticity of a we have

$$\begin{aligned}
\|u_h - u\|^2_{H^1(\Omega)} &\leq \frac{1}{\beta} a(u_h - u, u_h - u) \\
&= \frac{1}{\beta} a(u_h - u, u_h - v_h + v_h - u) \\
&= \frac{1}{\beta} a(u_h - u, u_h - v_h) + \frac{1}{\beta} a(u_h - u, v_h - u) \text{ (bilinearity)} \\
&= \frac{1}{\beta} a(u_h - u, v_h - u) \text{ (Galerkin orthogonality)} \\
&\leq \frac{\alpha}{\beta} \|u_h - u\|_{H^1(\Omega)} \|v_h - u\|_{H^1(\Omega)} \text{ (a continuous).}
\end{aligned}$$

By canceling $\|u_h - u\|_{H^1(\Omega)}$ on both sides we get

$$\|u_h - u\|_{H^1(\Omega)} \leq \frac{\alpha}{\beta} \|v_h - u\|_{H^1(\Omega)}.$$

This is true for all $v_h \in V_h$, so we can take the infimum over $v_h \in V_h$ on the right hand side and obtain

$$\begin{aligned}
\|u - u_h\|_{H^1(\Omega)} &\leq \frac{\alpha}{\beta} \inf_{v_h \in V_h} \|v_h - u\|_{H^1(\Omega)} \\
&\leq \underbrace{\widetilde{C}\frac{\alpha}{\beta}}_{=C} h\|u\|_{H^2(\Omega)} \text{ (by Assumption III.42).}
\end{aligned}$$

□

A similar convergence result can also be shown for the L^2-norm.

Theorem III.44: Under the same assumptions as in Theorem III.43, it holds that

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^2 \|u\|_{H^2(\Omega)}.$$

Let us conclude with some final remarks concerning the error analysis.

Remark III.45: a) Compared to the finite difference method from Chapter II.1 we see the following:

- The FDM needs $u \in C^4(\overline{\Omega})$ to obtain convergence of order 2 with respect to $\|\cdot\|_\infty$.
- The FEM needs only $u \in H^2(\Omega)$ to have the same order of convergence but with respect to the norm $\|\cdot\|_{L^2(\Omega)}$.

b) Better orders of convergence can be obtained by using higher degree Lagrange finite elements, but only if the exact solution u is sufficiently regular! For example, if one uses quadratic finite elements on a triangle, then one obtains

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^3 \|u\|_{H^3(\Omega)},$$

if $u \in H^3(\Omega)$. To sum up, the quality of the approximation depends on

- the regularity of u;
- the maximum length h of all edges in the triangulation;
- the minimum interior angle of the elements (which should be bounded away from zero).

III.9. FEM for Parabolic Problems

In this section we consider the approximation of parabolic problems. Therefore let L be a second-order elliptic operator in the spatial domain $\Omega \subset \mathbb{R}^n$ with polygonal boundary. For example you can think of $L = -\Delta$, so that the corresponding problem becomes the heat/diffusion equation. For a given source term $f\colon [0,T] \times \Omega \to \mathbb{R}$ and initial value $u_0\colon \Omega \to \mathbb{R}$ consider the initial boundary value problem

$$\begin{cases} \partial_t u + Lu = f & \text{in } Q_T = (0,T) \times \Omega, \\ u(t,x) = 0 & \text{for all } t \in (0,T), x \in \partial\Omega, \\ u(0,x) = u_0(x) & \text{for all } x \in \Omega. \end{cases} \tag{III.42}$$

In order to solve this problem numerically with the finite element method, we have to reformulate it as a variational problem. Following the same steps as for the Poisson equation we first multiply the differential equation with a test function $v \in H_0^1(\Omega) = V$ and integrate over Ω. An application of the theorem of Gauss (integration by parts) then yields

$$(\partial_t u, v)_{L^2(\Omega)} + a(u,v) = (f,v)_{L^2(\Omega)},$$

where the bilinear form[2] $a(u,v)$ and linear form (f,v) are

$$\begin{aligned} a(u,v) &= \int_\Omega a(x) \nabla u(t,x) \cdot \nabla v(x)\,\mathrm{d}x \quad + \quad b(u,v) \\ (f,v)_{L^2(\Omega)} &= \int_\Omega f(t,x) v(x)\,\mathrm{d}x. \end{aligned}$$

where the extra terms are contained in b[3]

$$b(u,v) = \int_\Omega c(x)u(x)v(x) + v(x)\big[\mathbf{b}(x) \cdot \nabla u(t,x)\big]\,\mathrm{d}x.$$

Note that due to $b(u,v)$ the bilinear form $a(u,v)$ is not symmetric. Observe that the solution u and the source term f depend on the time t, although this is sometimes suppressed in the notation.

The variational problem then reads as follows:

$$\begin{cases} \text{Find } u : [0,T] \times \Omega \to \mathbb{R} \text{ such that} \\ \text{a) For every } t \in (0,T) \text{ we have } u(t,\cdot) \in H_0^1(\Omega). \\ \text{b) } (\partial_t u(t,\cdot), v)_{L^2(\Omega)} + a(u(t,\cdot),v) = (f(t,\cdot),v)_{L^2(\Omega)} \text{ for all } v \in H_0^1(\Omega). \\ \text{c) } u(0,x) = u_0(x) \text{ for all } x \in \Omega. \end{cases} \tag{III.43}$$

Observe that the first condition on u ensures that the boundary conditions are satisfied. The notation $u(t,\cdot) \in H_0^1(\Omega)$ is short hand and means, more formally, that for every

[2] Note the ambiguity $a(u,v)$ being the bilinear form and $a(x)$ the coefficient function in L.
[3] For certain natural boundary conditions one might also integrate by parts in b.

fixed $t \in (0,T)$ the mapping $\Omega \ni x \mapsto u(t,x) \in \mathbb{R}$ is in $H_0^1(\Omega)$. Moreover, since $t = 0$ is excluded from this condition, it is allowed that the initial value u_0 does not satisfy the boundary conditions. This concept (t fixed, x unspecified) motivates a shorter notation. Instead of $u(t,\cdot)$ we will only write $u(t)$, where now $u\colon [0,T] \to H_0^1(\Omega)$, $t \mapsto u(t) \in H_0^1(\Omega)$, instead of $u\colon [0,T] \times \Omega \to \mathbb{R}$. Similarly, we often write $f\colon [0,T] \to L^2(\Omega)$ when we want to say that for every fixed t the mapping $x \mapsto f(t,x) \in \mathbb{R}$ is an element of the space $L^2(\Omega)$.

Next, we should spend a moment in order to check if the variational problem (III.43) is well-posed. However, we skip over this part and only refer to the literature for results on the existence and uniqueness of a weak solution to (III.43).

Concerning the numerical discretization, we first approximate the solution u to (III.43) only with respect to $x \in \Omega$. This will result in a *spatial (semi-)discretization*. To this end let $(V_h)_{h\in(0,1)}$ be a sequence of finite element spaces with refinement parameter $h \in (0,1)$. (e. g., $V_h = \operatorname{span}\{\varphi_j \mid j = 1,\dots,N_h\}$, where φ_j are pyramid basis functions with associated nodes $z_j \in \Omega$. In particular, we assume that V_h respects the boundary conditions, meaning $V_h \subset H_0^1(\Omega)$.)

The semi-discrete problem is to find a mapping $u_h\colon [0,T] \to V_h$ such that

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\,(u_h(t), v_h)_{L^2(\Omega)} + a(u_h(t), v_h) = (f(t), v_h)_{L^2(\Omega)} \quad \text{for all } t \in (0,T),\ v_h \in V_h, \\ u_h(0) = u_{0,h} \in V_h, \end{cases} \tag{III.44}$$

where $u_{0,h} \in V_h$ is some suitable approximation of u_0, for example, obtained by a piecewise linear interpolation.

The mapping u_h is then called the *spatially (semi-)discrete approximation* of u. The semi-discrete problem is now reformulated in terms of the basis functions: Find coefficient functions $\alpha_j\colon [0,T] \to \mathbb{R}$ such that

$$u_h(t) = \sum_{j=1}^{N_h} \alpha_j(t)\varphi_j \in V_h$$

$$u_h(0) = u_{0,h} = \sum_{j=1}^{N_h} \alpha_j(0)\varphi_j.$$

Using the interpolation property of the basis functions ($\varphi_j(z_\ell) = \delta_{j,\ell}$) we directly get

$$u_h(0)(z_\ell) = u_{0,h}(z_\ell) = \sum_{j=1}^{N_h} \alpha_j(0) \underbrace{\varphi_j(z_\ell)}_{=\delta_{j,\ell}} = \alpha_\ell(0)$$

and from (III.44) with $v_h = \varphi_\ell$ we get

$$\begin{aligned} & \frac{\mathrm{d}}{\mathrm{d}t}\,(u_h(t), \varphi_\ell)_{L^2(\Omega)} + a(u_h(t), \varphi_\ell) = (f(t), \varphi_\ell)_{L^2(\Omega)} \\ \Leftrightarrow \quad & \sum_{j=1}^{N_h} \alpha_j'(t)(\varphi_j, \varphi_\ell)_{L^2(\Omega)} + \sum_{j=1}^{N_h} \alpha_j(t) a(\varphi_j, \varphi_\ell) \overset{!}{=} (f(t), \varphi_\ell)_{L^2(\Omega)} \quad \text{for all } \ell = 1,\dots,N_h. \end{aligned}$$

In matrix-vector-formulation, this gives with $\alpha(t) = [\alpha_1(t), \dots, \alpha_{N_h}(t)]^\mathsf{T}$

$$M_h \alpha'(t) + A_h \alpha(t) = f_h(t), \tag{III.45}$$

where $M_h = \left[(\varphi_j, \varphi_\ell)_{L^2(\Omega)}\right]_{j,\ell=1}^{N_h}$ is the mass matrix, $A_h = [a(\varphi_j, \varphi_\ell)]_{j,\ell=1}^{N_h}$ is the stiffness matrix, $f_h(t) = [(f(t), \varphi_\ell)_{L^2(\Omega)}]_{\ell=1}^{N_h}$ denotes the load vector. We further have the initial condition

$$\alpha(0) = \left[\alpha_1(0), \dots, \alpha_{N_h}(0)\right]^\mathsf{T} \overset{!}{=} \left[u_{0,h}(z_1), \dots, u_{0,h}(z_{N_h})\right]^\mathsf{T}.$$

Let us take a close look on the mass matrix M_h. Note that M_h is symmetric (since the $L^2(\Omega)$ inner product is symmetric) and positive definite: Take $\xi = [\xi_i]_{i=1}^{N_h} \in \mathbb{R}^{N_h}$ arbitrarily. Then we get

$$\xi^\mathsf{T} M_h \xi = \sum_{j,\ell=1}^{N_h} \xi_j \xi_\ell (\varphi_j, \varphi_\ell)_{L^2(\Omega)} = \left(\sum_{j=1}^{N_h} \xi_j \varphi_j, \sum_{\ell=1}^{N_h} \xi_\ell \varphi_\ell \right)_{L^2(\Omega)} = \left\| \sum_{j=1}^{N_h} \xi_j \varphi_j \right\|_{L^2(\Omega)}^2 \geq 0.$$

Since $(\varphi_j)_{j=1}^{N_h}$ is a basis of V_h equality only holds if $\xi_j = 0$ for all j. Thus, M_h is positive definite and, consequently, invertible.

We can now multiply (III.45) by M_h^{-1} and obtain a standard ODE (linear, inhomogeneous) for α:

$$\begin{cases} \alpha'(t) + M_h^{-1} A_h \alpha(t) = M_h^{-1} f_h(t), \\ \alpha(0) = [u_{0,h}(z_j)]_{j=1}^{N_h}. \end{cases} \tag{III.46}$$

Standard ODE theory shows that there exists a unique solution to the initial value problem (III.46) (variation of constants formula). This also shows that the semidiscrete approximation u_h exists and is uniquely determined.

However, we stress that the ODE (III.46) is mostly of theoretical interest. One should not use the representation (III.46) for computational purposes, since the computation of the inverse M_h^{-1} is too expensive and results, in general, in a full matrix. Further below, we will present a fully discrete scheme based on the form (III.45).

But before we come to this, let us briefly investigate the error of the semi-discrete approximation.

Theorem III.46: Let u_h and u be the solutions of (III.44) and (III.43) respectively. Then

$$\|u_h(t) - u(t)\|_{L^2(\Omega)} \leq \|u_{0,h} - u_0\|_{L^2(\Omega)} + Ch^2 \left(\|u_0\|_{H^2(\Omega)} + \int_0^t \|\partial_s u(s)\|_{H^2(\Omega)} \, \mathrm{d}s \right)$$

for all $t \in [0, T]$.

Note that the last term on the right hand side constitutes a *regularity requirement* on the exact solution u stating that the mapping $u\colon [0,T] \to H_0^1(\Omega)$ is in fact differentiable with respect to time and the derivative $\frac{\mathrm{d}}{\mathrm{d}s}u$ takes values in $H^2(\Omega)$ and is integrable. This is an additional assumption on u that needs to be verified. If the exact solution is less regular, then the finite element method may fail to converge or may only converge with a reduced order of convergence.

Let us briefly discuss also a fully discrete scheme: The perhaps simplest fully discrete scheme combines the Galerkin FEM with the backward Euler method for the temporal approximation. The idea consists of replacing the time derivative in (III.44) by a difference quotient:

Let $k > 0$, $k = \frac{T}{N_k}$, $N_k \in \mathbb{N}$ be the time step size. Find $(U^n)_{n=0}^{N_k} \subset V_h$ such that

$$\begin{cases} \left(\frac{U^n - U^{n-1}}{k}, v_h\right)_{L^2(\Omega)} + a(U^n, v_h) = (f(t_n), v_h)_{L^2(\Omega)}, \\ U^0 = u_{0,h} \in V_h \quad \forall\, n = 1, \ldots, N_k,\ \forall v_h \in V_h, \end{cases}$$

where $t_n = nk$.

Given U^{n-1} this defines U^n implicitly from the discrete elliptic problem.

$$(U^n, v_h)_{L^2(\Omega)} + ka(U^n, v_h) = (U^{n-1} + kf(t_n), v_h)_{L^2(\Omega)} \quad \forall\, v_h \in V_h.$$

Expressing U^n in terms of the finite element basis $(\varphi_j)_{j=1}^{N_h}$ gives

$$U^n = \sum_{j=1}^{N_h} \alpha_j^n \varphi_j$$

with coefficients $\alpha^n = \left[\alpha_1^n, \ldots, \alpha_{N_h}^n\right]^\mathsf{T}, n = 0, \ldots, N_k$. This leads to the matrix-vector system

$$M_h \alpha^n + kA_h \alpha^n = M_h \alpha^{n-1} + kf_h^n. \tag{III.47}$$

The initial data α^0 contains the values of the FEM approximation $u_{0,h} \in V_h$ of u_0 at the nodes of the triangulation. This stays true for α^n which can therefore directly be used for plotting, data processing, etc.

Moreover, observe that the matrix $M_h + kA_h$ positive definite and, for $b = 0$ symmetric. In addition, A_h does not depend on t (respective n) in this example. However, this could happen if the coefficient functions in $a(u,v)$ depend on time, i.e., $a(t,x), b(t,x), c(t,x)$.

Theorem III.47 (Error estimate)**:** Let U^n and $u(t_n)$ be the solutions to (III.47) and to (III.43) at time $t_n = nk$, $n \in \mathbb{N}$. Then for all $n \in \mathbb{N}$, $h \in (0,1)$, $k = \frac{T}{N_k}$ it holds that

$$\begin{aligned} \|U^n - u(t_n)\|_{L^2(\Omega)} \leq Ch^2 \left(\|u_0\|_{H^2(\Omega)} + \int_0^{t_n} \|\partial_s u(s)\|_{H^2(\Omega)} \,\mathrm{d}s \right) \\ + Ck \left(\int_0^{t_n} \|\partial_s u(s)\|_{L^2(\Omega)} \,\mathrm{d}s \right) \end{aligned}$$

if the initial condition satisfies $\|u_{0,h} - u_0\|_{L^2(\Omega)} \leq Ch^2 \|u_0\|_{H^2(\Omega)}$.

Remark III.48: The error estimate shows that we have order 2 convergence with respect to the spatial approximation and order 1 convergence with respect to the temporal discretization. This suggests to use $h \approx \sqrt{k}$ in order to equilibrate the two errors.

IV. Solving Linear Equation Systems

In this chapter we study some strategies to solve systems of linear equations $A_h u_h = f_h$ where $A_h \in \mathbb{R}^{N_h \times N_h}$ and $N_h \in \mathbb{N}$ is very large, A_h may be symmetric, positive definite, and sparse.

IV.1. A Model Problem

Consider again the Poisson equation on a bounded polygonal domain $\Omega \subset \mathbb{R}^2$

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{IV.1}$$

and its weak formulation

$$\text{Find } u \in V \text{ such that } a(u,v) = F(v) \quad \forall\, v \in V \tag{IV.2}$$

where $V = H_0^1(\Omega)$ and

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x,$$
$$F(v) = \int_\Omega f v \,\mathrm{d}x.$$

In Chapter III we introduced the Galerkin-FEM for (IV.2):

$$\text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h \tag{IV.3}$$

where $V_h = \operatorname{span}\{\varphi_1, \dots, \varphi_{N_h}\}$, $N_h = \dim(V_h)$.

Then (IV.3) is equivalent to solving

$$A_h \alpha = f_h \tag{IV.4}$$

with $\alpha = \big[\alpha_1, \dots, \alpha_{N_h}\big]^\mathsf{T} \in \mathbb{R}^{N_h}$, the stiffness matrix $A_h = [a(\varphi_i, \varphi_j)]_{i,j=1}^{N_h} \in \mathbb{R}^{N_h \times N_h}$, and load vector $f_h = \big[F(\varphi_1), \dots, F(\varphi_{N_h})\big]^\mathsf{T}$.

Recall that the bilinear form $a(\cdot,\cdot)$ is symmetric and elliptic on $H_0^1(\Omega)$ and therefore,

$$\|v\|_a := \sqrt{a(v,v)} = \left(\int_\Omega \nabla v \cdot \nabla v \,\mathrm{d}x \right)^{1/2}$$

induces a norm on $H_0^1(\Omega)$ which is called the *energy norm* on $V = H_0^1(\Omega)$.

Let $\Phi : \mathbb{R}^{N_h} \to V_h$ (embedding) be defined by

$$\Phi(x) = \sum_{j=1}^{N_h} x_j \varphi_j$$

for $x = \begin{bmatrix} x_1, \ldots, x_{N_h} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{N_h}$. Note that $V_h \ni u_h = \Phi(\alpha) = \sum_{j=1}^{N_h} \alpha_j \varphi_j$. In other words, Φ switches between function space $V_h \subset V$ and the Euclidean space $\mathbb{R}^{N_h}$, in particular, Φ is linear and bijective (and therefore, an *isomorphism*).

Then we define the *algebraic energy norm* on $\mathbb{R}^{N_h}$ by

$$\|x\|_{A_h} := \sqrt{a(\Phi(x), \Phi(x))} = \|\Phi(x)\|_a,$$

i. e. the norms under the action of Φ are preserved (we call this an isometry between $(\mathbb{R}^{N_h}, \|\cdot\|_{A_h})$ and $(V_h, \|\cdot\|_a)$.)

We have

$$a(\Phi(x), \Phi(x)) = \sum_{j=1}^{N_h} x_i a(\varphi_i, \varphi_j) x_j = x^\mathsf{T} A_h x.$$

Since A_h is symmetric and positive definite, $\|\cdot\|_{A_h}$ is indeed a norm.

We make the following observation:

a) $\|\cdot\|_a$ is a norm on a function space;

b) $\|\cdot\|_{A_h}$ is a norm on the Euclidean space $\mathbb{R}^{N_h}$;

c) Using the embedding Φ we obtain for $\Psi = \Phi(x)$

$$\|x\|_{A_h}^2 = x^\mathsf{T} A_h x = a(\Phi(x), \Phi(x)) = \|\Psi\|_a^2.$$

Next, let u and u_h be the solutions to (IV.2) and (IV.3), respectively. Then the Galerkin orthogonality gives us

$$a(u - u_h, v_h) = 0 \quad \forall\, v_h \in V_h.$$

Recall that u_h is the best approximation with respect to the energy norm, that is

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a. \tag{IV.5}$$

In practice, we often have $N_h \gg 1$. Solving the linear system (IV.4) exactly is often too expensive.

Let $\alpha^{(j)}$ be an approximation of α (e. g., obtained after the j-th iteration of an iterative solver). Then $u_h^{(j)} = \Phi(\alpha^{(j)})$ is an approximation of $u_h = \Phi(\alpha)$. Then we have the following errors:

- the *FEM error* $u - u_h$ obtained by approximating V by V_h;
- the *algebraic error* $u_h - u_h^{(j)}$ obtained by the inexactness of the solution of the resulting linear system;

- the *total error* $u - u_h^{(j)}$.

If we measure the total error with respect to the energy norm we get

$$\begin{aligned}
\|u - u_h^{(j)}\|_a^2 &= a\big(u - u_h^{(j)}, u - u_h^{(j)}\big) \\
&= a\big(u - u_h + u_h - u_h^{(j)}, u - u_h + u_h - u_h^{(j)}\big) \\
&= a(u - u_h, u - u_h) + 2a\big(u - u_h, \underbrace{u_h - u_h^{(j)}}_{\in V_h}\big) + a\big(u_h - u_h^{(j)}, u_h - u_h^{(j)}\big) \\
&= \|u - u_h\|_a^2 + 0 + \|u_h - u_h^{(j)}\|_a^2 \\
&= \|u - u_h\|_a^2 + \|\alpha - \alpha^{(j)}\|_{A_h}^2.
\end{aligned}$$

Theorem IV.1: The square of the energy norm of the total error is equal to the sum of the square of the FEM error in the energy norm and the square of the algebraic energy norm of the algebraic error:

$$\|u - u_h^{(j)}\|_a^2 = \|u - u_h\|_a^2 + \|\alpha - \alpha^{(j)}\|_{A_h}^2.$$

Conclusion IV.2: Let us sum up the above result:

- The algebraic error and FEM error are orthogonal, since $a\big(u - u_h, u_h - u_h^{(j)}\big) = 0$.
- Since the errors simply sum up it is sufficient to use an approximation of α which has an error, roughly speaking, of the same magnitude as the FEM error.

IV.2. The Conjugate Gradient Method

The CG method is an "iterative" solver that is designed to generate a sequence of iterates $\big(x^{(1)}, x^{(2)}, x^{(3)}, \dots\big)$ that converges to the solution x of the linear system $Ax = b$. Here we assume that the matrix A is symmetric and positive definite, these assumptions will be crucial for the effectiveness of the method. Moreover, we assume that A is large and sparse. Sparse matrices can be stored much more efficiently than full dense matrices, since we only have to store the position and the values of the non-zero entries of A. Moreover, sparse matrices allow for the efficient execution of certain algebraic operations such as the evaluation of matrix-vector products. All iterative solvers presented in the literature are built on a sequence of matrix-vector multiplications and the evaluation of inner products etc. and therefore, it is important that these operations can be executed fast.

Define the functional $J : \mathbb{R}^n \to \mathbb{R}$ by $J[x] := \frac{1}{2}x^\mathsf{T}Ax - x^\mathsf{T}b$. If $A\widehat{x} = b$, then

$$\begin{aligned}
J[x] - J[\widehat{x}] &= \frac{1}{2}x^\mathsf{T}Ax - x^\mathsf{T}b - \frac{1}{2}\widehat{x}^\mathsf{T}A\widehat{x} + \widehat{x}^\mathsf{T}b \\
&= \frac{1}{2}\big(x - \widehat{x}\big)^\mathsf{T}A\big(x - \widehat{x}\big) + x^\mathsf{T}A\widehat{x} - \widehat{x}^\mathsf{T}A\widehat{x} - x^\mathsf{T}b + \widehat{x}^\mathsf{T}b \\
&= \frac{1}{2}(x - \widehat{x})^\mathsf{T}A(x - \widehat{x}) \geq 0.
\end{aligned}$$

In other words, $J[x] \geq J[\widehat{x}]$ for all $x \in \mathbb{R}^n$ and therefore, J attains a global minimum at $\widehat{x}$.

Idea: Construct an iterative method that minimizes J! Recall that $\|x - \widehat{x}\|_A^2 = (x-\widehat{x})^\mathsf{T} A(x-\widehat{x}) = \langle x-\widehat{x}, x-\widehat{x}\rangle_A$ is the algebraic energy norm/algebraic inner product.

Now we will construct the CG method based on the above idea: Let $x^{(j)}$ be the current value of the iteration. Then we have to determine a "suitable descent direction" $d^{(j)} \in \mathbb{R}^n$ and set

$$x^{(j+1)} = x^{(j)} + \alpha_j d^{(j)} \tag{IV.6}$$

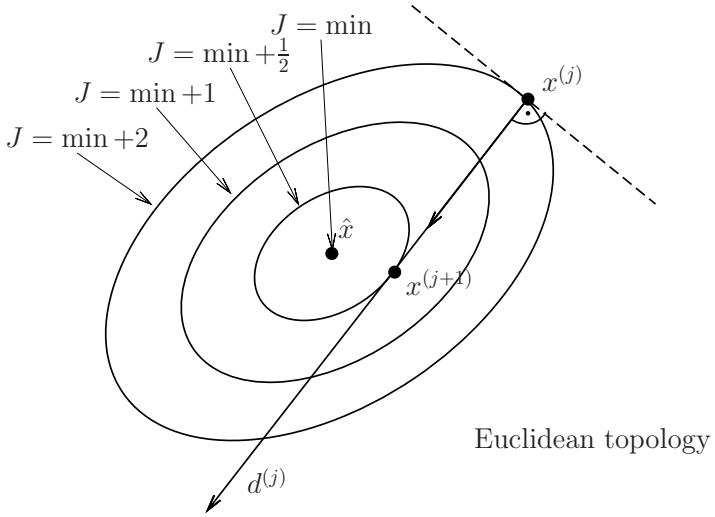
where α_j is chosen in such a way that

$$J\big[x^{(j)} + \alpha_j d^{(j)}\big] = J[x^{(j)}] + \alpha_j \big(d^{(j)}\big)^\mathsf{T} A x^{(j)} + \frac{1}{2}\alpha_j^2 \big(d^{(j)}\big)^\mathsf{T} A d^{(j)} - \alpha_j \big(d^{(j)}\big)^\mathsf{T} b$$

is minimized. By differentiating the above and setting the result to zero, we can compute this optimal value which is

$$\alpha_j = \frac{\big(r^{(j)}\big)^\mathsf{T} d^{(j)}}{\big(d^{(j)}\big)^\mathsf{T} A d^{(j)}} = \frac{\big\langle r^{(j)}, d^{(j)}\big\rangle_2}{\big\langle d^{(j)}, d^{(j)}\big\rangle_A} \tag{IV.7}$$

where $r^{(j)} = b - Ax^{(j)}$ ("residual"). Note that the denominator is always non-zero if $d^{(j)} \neq 0$.

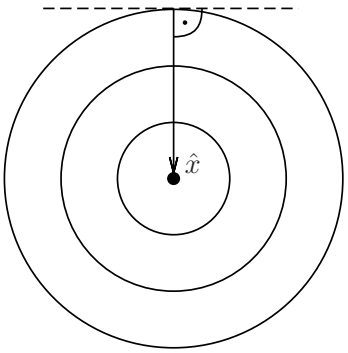


The question how to choose an optimal search direction $d^{(j)}$ remains. Recall that $-\nabla J$ always points to the steepest descent direction of J. For our problem we have

$$\nabla J[x] = Ax - b,$$

so we get

$$-\nabla J\big[x^{(j)}\big] = b - Ax^{(j)} =: r^{(j)} \quad \text{(residual)}.$$



We observe the following:

- If the level sets are circles, the gradient points directly to the center.
- The gradient may not be the best descent direction if the level sets are not circles.
- It holds $J[x] = J[\widehat{x}] + \frac{1}{2}\|x - \widehat{x}\|_A^2$, i. e., the level sets of J are circles with respect to the algebraic energy norm!

We will use this fact to speed up the *method of steepest descent* ($d^{(j)} := r^{(j)}$) and transform this method into the energy norm topology.

Ansatz: Choose the new descent direction as a linear combination of the gradient (i. e., the residual) and the old descent direction, that is

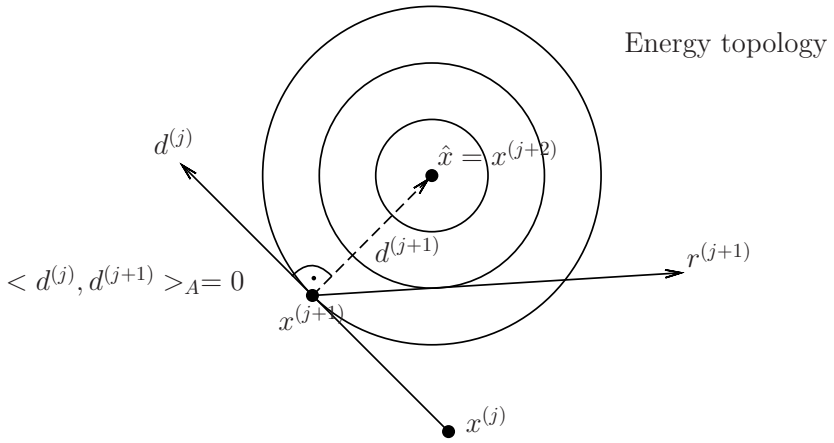
$$d^{(j+1)} := r^{(j+1)} + \beta_j d^{(j)} \tag{IV.8}$$

such that $\langle d^{(j+1)}, d^{(j)} \rangle_A = 0$. Then, $d^{(j+1)}$ is orthogonal to $d^{(j)}$ in the energy inner product and we get

$$0 = \big(d^{(j)}\big)^{\mathsf{T}} A d^{(j+1)} = \big(d^{(j)}\big)^{\mathsf{T}} A r^{(j+1)} + \beta_j \big(d^{(j)}\big)^{\mathsf{T}} A d^{(j)},$$

which gives

$$\beta_j = -\frac{\langle r^{(j+1)}, d^{(j)} \rangle_A}{\langle d^{(j)}, d^{(j)} \rangle_A}. \tag{IV.9}$$



The CG method is fully determined by (IV.7), (IV.8), and (IV.9). Note that $d^{(j+1)} = 0$ only happens if $r^{(j+1)}$ and $d^{(j)}$ are linearly dependent. However, $x^{(j+1)}$ lies on the intersection of the line $\alpha \mapsto x^{(j)} + \alpha d^{(j)}$ and the level set of $J[x^{(j+1)}]$. By the choice of α_j we have

$$-\nabla J\big[x^{(j+1)}\big] = r^{(j+1)} \perp d^{(j)}$$

with respect to the Euclidean inner product. This can only happen at the same time if $r^{(j+1)} = 0$ and hence $Ax^{(j+1)} = b$, in other words, if $x^{(j+1)} \neq \widehat{x}$, then $d^{(j+1)} \neq 0$.

Lemma IV.3: Let $x^{(0)} \in \mathbb{R}^n$ be arbitrary. Set $d^{(0)} := r^{(0)} = b - Ax^{(0)}$. Let $(x^{(j)})_{j\in\mathbb{N}}$ and $(d^{(j)})_{j\in\mathbb{N}}$ be defined by (IV.6) and (IV.8). If $x^{(l)} \neq \widehat{x}$ for $l = 0, \ldots, j$ then for all $0 \leq l < j$ it holds

a) $\big\langle r^{(j)}, d^{(l)} \big\rangle_2 = 0$;

b) $\big\langle r^{(j)}, r^{(l)} \big\rangle_2 = 0$;

c) $\big\langle d^{(j)}, d^{(l)} \big\rangle_A = 0$.

Proof. For each $k \geq 0$ we have $Ax^{(k+1)} = Ax^{(k)} + \alpha_k Ad^{(k)}$ and therefore

$$\begin{aligned} r^{(k+1)} &= b - Ax^{(k+1)} \\ &= b - Ax^{(k)} - \alpha_k Ad^{(k)} \\ &= r^{(k)} - \alpha_k Ad^{(k)}. \end{aligned} \tag{IV.10}$$

By inserting the value of α_k we get

$$\begin{aligned} \big\langle r^{(k+1)}, d^{(k)} \big\rangle_2 &= \big\langle r^{(k)} - \alpha_k Ad^{(k)}, d^{(k)} \big\rangle_2 \\ &= \big\langle r^{(k)}, d^{(k)} \big\rangle - \alpha_k \big\langle d^{(k)}, d^{(k)} \big\rangle_A = 0. \end{aligned} \tag{IV.11}$$

The remainder of the proof is now done by induction.

Base case $j = 1$: If $k = 0$ then (IV.11) is (a). Since $d^{(0)} = r^{(0)}$ this also gives (b). Finally (c) follows directly from (IV.8) (by how $d^{(j+1)}$ is chosen).

Induction hypothesis: Assume that a), b), c) are correct for each $\widetilde{j} \leq j$.

Induction step $j \to j+1$: First we get from (IV.11) that $\langle r^{(j+1)}, d^{(j)} \rangle = 0$. In addition, by a), c) and (IV.10) we obtain

$$\begin{aligned} \langle r^{(j+1)}, d^{(l)} \rangle_2 &= \langle r^{(j)} - \alpha_j A d^{(j)}, d^{(l)} \rangle_2 \\ &= \underbrace{\langle r^{(j)}, d^{(l)} \rangle_2}_{=0 \text{ by a)}} - \underbrace{\alpha_j \langle d^{(j)}, d^{(l)} \rangle_A}_{=0 \text{ by c)}} \end{aligned}$$

for all $l < j$. This proves a) for $j + 1$.

Next, due to (IV.8) we have

$$r^{(l)} = d^{(l)} - \beta_{l-1} d^{(l-1)}, \quad l = 1, \ldots, j.$$

This gives

$$\langle r^{(j+1)}, r^{(l)} \rangle_2 = \langle r^{(j+1)}, d^{(l)} \rangle_2 - \beta_{l-1} \langle r^{(j+1)}, d^{(l-1)} \rangle_2 \overset{a)}{=} 0.$$

This gives us b).

Regarding c) we get the following: The case $\langle d^{(j+1)}, d^{(j)} \rangle_A = 0$ is true by how $d^{(j+1)}$ is constructed. Consider now $l < j$. This yields

$$\begin{aligned} \langle d^{(j+1)}, d^{(l)} \rangle_A &\overset{(IV.8)}{=} \langle r^{(j+1)}, d^{(l)} \rangle_A + \beta_j \langle d^{(j)}, d^{(l)} \rangle_A \\ &= \langle r^{(j+1)}, A d^{(l)} \rangle_2 \\ &= \frac{1}{\alpha_l} \langle r^{(j+1)}, r^{(l)} - r^{(l+1)} \rangle_2 \overset{b)}{=} 0 \end{aligned}$$

Therefore, we either have $\langle d^{(j+1)}, d^{(l)} \rangle_A = 0$ as claimed or $\alpha_l = 0$. In the following we thus show that $\alpha_l \neq 0$: For this assume $\alpha_l = 0$. Due to (IV.7) this is equivalent to

$$\begin{aligned} 0 = \langle r^{(l)}, d^{(l)} \rangle_2 &\overset{(IV.8)}{=} \langle r^{(l)}, r^{(l)} + \beta_{l-1} d^{(l-1)} \rangle_2 \\ &= \langle r^{(l)}, r^{(l)} \rangle_2 + \beta_{l-1} \underbrace{\langle r^{(l)}, d^{(l-1)} \rangle_2}_{=0 \text{ by a)}} \\ &= \| r^{(l)} \|_2^2 \end{aligned}$$

for each $0 < l < j$ and

$$0 = \langle r^{(0)}, d^{(0)} \rangle_2 = \langle r^{(0)}, r^{(0)} \rangle_2 = \| r^{(0)} \|_2^2,$$

if $l = 0$. In both cases, this gives $r^{(l)} = 0$ which contradicts with $x^{(l)} \neq \widehat{x}$ for all $x^{(l)}$, $l \leq j + 1$ and therefore $\alpha_l \neq 0$. This implies that $\langle d^{(j+1)}, d^{(l)} \rangle_A = 0$ for all $0 \leq l < j + 1$. This completes the proof of c). □

Corollary IV.4: Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. Then, the CG method converges to the exact solution $\widehat{x}$ in at most n steps, that is, there exists a j with $0 \leq j \leq n$ such that $x^{(j)} = \widehat{x}$.

Proof. Due to lemma IV.3 all descent directions $(d^{(j)})_{j\in\mathbb{N}}$ are pairwise orthogonal with respect to $\langle\cdot,\cdot\rangle_A$. The maximum number of pairwise orthogonal vectors is equal to the dimension of the linear system. Thus there exists a j with $0 \leq j \leq n$ such that $d^{(j)} = 0$, therefore, $r^{(j)} = 0$ which gives $x^{(j)} = \widehat{x}$. □

Implementation of the CG-method: For the implementation one should not use (IV.7) and (IV.9) directly. Instead we use that by Lemma IV.3 a) we have

$$\left\langle r^{(j)}, d^{(j)}\right\rangle_2 = \left\langle r^{(j)}, r^{(j)} + \beta_{j-1}d^{(j-1)}\right\rangle_2 = \left\|r^{(j)}\right\|_2^2,$$

and therefore,

$$\alpha_j = \frac{\left\|r^{(j)}\right\|_2^2}{\left\|d^{(j)}\right\|_A^2}. \tag{IV.12}$$

Similarly we get

$$\begin{aligned}\left\langle r^{(j+1)}, d^{(j)}\right\rangle_A &= \left\langle r^{(j+1)}, Ad^{(j)}\right\rangle_2 \\ &= \frac{1}{\alpha_j}\left\langle r^{(j+1)}, r^{(j)} - r^{(j+1)}\right\rangle_2 \\ &= -\frac{1}{\alpha_j}\left\|r^{(j+1)}\right\|_2^2 = -\frac{\left\|r^{(j+1)}\right\|_2^2}{\left\|r^{(j)}\right\|_2^2}\left\|d^{(j)}\right\|_A^2.\end{aligned}$$

Hence,

$$\beta_j = \frac{\left\|r^{(j+1)}\right\|_2^2}{\left\|r^{(j)}\right\|_2^2}. \tag{IV.13}$$

These observations lead to the Algorithm 1.

Let us summarize the computational cost of Algorithm 1:

a) In every step of the algorithm we have to evaluate the matrix-vector product $Ad^{(j)}$ (which is the most expensive step of the iteration), the two inner products $\left\langle d^{(j)}, Ad^{(j)}\right\rangle_2$ and $\left\langle r^{(j+1)}, r^{(j+1)}\right\rangle_2$, and three sums of vectors.

b) The memory requirements are very low, since it is only necessary to have the data available from the previous iteration (short recurrences!). Therefore, in the $j+1$-st iteration, the data of the $j-1$-st iteration can be overwritten by the updated results.

Remark IV.5:

- Although corollary IV.4 assures convergence to the exact solution after at most n iterations, the CG method is usually used as an iterative (incomplete) solver. We stop much earlier than before the iteration, which would give us the exact solution.

Algorithm 1 Conjugate gradient method for linear systems

Input: A symmetric and positive definite matrix $A \in \mathbb{R}^{n\times n}$, right-hand side $b \in \mathbb{R}^n$.
Output: Approximation $\widetilde{x} \in \mathbb{R}^n$ to the solution $\widehat{x} \in \mathbb{R}^n$ of $Ax = b$.
Initialization: Choose $x^{(0)} \in \mathbb{R}^n$ arbitrarily and set $r^{(0)} := b - Ax^{(0)}$, $d^{(0)} := r^{(0)}$.
for $j = 0, 1, 2, \ldots$ (until convergence) **do**
Set $\alpha_j := \|r^{(j)}\|_2^2 / \langle d^{(j)}, Ad^{(j)}\rangle_2$.
Set $x^{(j+1)} := x^{(j)} + \alpha_j d^{(j)}$.
Set $r^{(j+1)} := r^{(j)} - \alpha_j A d^{(j)}$.
Set $\beta_j := \|r^{(j+1)}\|_2^2 / \|r^{(j)}\|_2^2$.
Set $d^{(j+1)} := r^{(j+1)} + \beta_j d^{(j)}$.
end for
Return $\widetilde{x} := x^{(j+1)}$.

- In practice, the orthogonality of $(d^{(j)})_{j\in\mathbb{N}}$ is often lost because of round-off errors.
- The speed of convergence depends on the *condition number* $\kappa(A)$ of A. One can show that
$$\|\widehat{x} - x^{(j)}\|_A \leq 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^j \|\widehat{x} - x^{(0)}\|_A$$
where $\kappa(A) := \|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ where $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ denote the largest and the smallest eigenvalue of A, respectively.

Next we present the proof of the convergence rate of the CG method. Observe that for j we have that

$$\begin{aligned} x^{(j)} &= x^{(j-1)} + \alpha_{j-1} d^{(j-1)} \\ &= x^{(j-2)} + \alpha_{j-1} d^{(j-1)} + \alpha_{j-2} d^{(j-2)} \\ &= \ldots \\ &= x^{(0)} + \sum_{i=0}^{j-1} \alpha_i d^{(i)}. \end{aligned} \tag{IV.14}$$

Therefore, $x^{(j)} \in x^{(0)} + \mathcal{W}_j$, where $\mathcal{W}_j = \operatorname{span}\{d^{(0)}, \ldots, d^{(j-1)}\}$. For the main result we now make use of the fact that $\mathcal{W}_j$ is a so-called *Krylov subspace*. This means that the following lemma is satisfied:

Lemma IV.6: It holds that

$$\mathcal{W}_j = \operatorname{span}\left\{r^{(0)}, Ar^{(0)}, A^2 r^{(0)}, \ldots, A^{j-1} r^{(0)}\right\}. \tag{IV.15}$$

Proof. First recall that

$$\begin{aligned} r^{(0)} &= d^{(0)}, \\ r^{(i)} &= d^{(i)} - \beta_{i-1} d^{(i-1)}, \quad i = 1, \ldots, j-1. \end{aligned}$$

Thus we see that

$$\operatorname{span}\left\{r^{(0)}, \ldots, r^{(j-1)}\right\} \subseteq \operatorname{span}\left\{d^{(0)}, \ldots, d^{(j-1)}\right\}.$$

On the other hand,

$$\begin{aligned} d^{(i)} &= r^{(i)} + \beta_{i-1} d^{(i-1)} \\ &= r^{(i)} + \beta_{i-1}\left(r^{(i-1)} + \beta_{i-2} d^{(i-2)}\right) \\ &= \ldots \\ &\in \operatorname{span}\left\{r^{(0)}, \ldots, r^{(i)}\right\}, \end{aligned}$$

and thus,

$$\operatorname{span}\left\{d^{(0)}, \ldots, d^{(j-1)}\right\} \subseteq \operatorname{span}\left\{r^{(0)}, \ldots, r^{(j-1)}\right\}.$$

Let us prove the statement of the lemma by induction. The base step $j = 1$ is clear, since $r^{(0)} = d^{(0)}$. Assume that the statement is true for all $j \leq k$. We show that then the statement is also true for $j = k + 1$: From (IV.14) we obtain

$$\begin{aligned} r^{(k)} &= b - Ax^{(k)} \\ &= b - Ax^{(0)} - A\sum_{i=0}^{k-1} \alpha_i d^{(i)} \\ &= r^{(0)} - A\sum_{i=0}^{k-1} \alpha_i d^{(i)} \\ &\in r^{(0)} + A\mathcal{W}_k \\ &= r^{(0)} + A \cdot \operatorname{span}\left\{r^{(0)}, Ar^{(0)}, A^2 r^{(0)}, \ldots, A^{k-1} r^{(0)}\right\} \quad \text{(induction hypothesis)} \\ &\subseteq \operatorname{span}\left\{r^{(0)}, Ar^{(0)}, A^2 r^{(0)}, \ldots, A^k r^{(0)}\right\}. \end{aligned}$$

Overall, this gives

$$\begin{aligned} \operatorname{span}\left\{d^{(0)}, \ldots, d^{(k)}\right\} &= \operatorname{span}\left\{r^{(0)}, \ldots, r^{(k)}\right\} \\ &\subseteq \operatorname{span}\left\{r^{(0)}, Ar^{(0)}, A^2 r^{(0)}, \ldots, A^k r^{(0)}\right\}. \end{aligned}$$

Since the $(d^{(i)})_{i=0}^{k}$ are pairwise A-orthogonal vectors, we have that

$$\dim \operatorname{span}\left\{d^{(0)}, \ldots, d^{(k)}\right\} = k + 1,$$

provided that none of the $d^{(j)}$ is zero (the case that one of the $d^{(j)}$ is zero is left out for brevity). But since

$$\dim \operatorname{span}\left\{r^{(0)}, Ar^{(0)}, A^2 r^{(0)}, \ldots, A^k r^{(0)}\right\} \leq k + 1,$$

we must have

$$\mathcal{W}_{k+1} = \operatorname{span}\left\{d^{(0)}, \ldots, d^{(k)}\right\} = \operatorname{span}\left\{r^{(0)}, Ar^{(0)}, A^2r^{(0)}, \ldots, A^kr^{(0)}\right\}.$$

□

The benefit of the fact that $\mathcal{W}_j$ is a Krylov subspace is that we can write

$$r^{(j)} = P(A)r^{(0)},$$

where P is a polynomial of degree at most $j-1$ (we write $P \in \mathcal{P}_{j-1}$). This allows us to formulate the following error estimate:

Lemma IV.7: With the notation above we have

$$\left\|\widehat{x} - x^{(j)}\right\|_A = \min_{P\in\mathcal{P}_j,\, P(0)=1} \left\|P(A)\left(\widehat{x} - x^{(0)}\right)\right\|_A.$$

Proof. Since $r^{(j)}$ is orthogonal to $\mathcal{W}_j$, we have

$$\left(\widehat{x} - x^{(j)}\right)^{\mathsf{T}} Ay = \left(r^{(j)}\right)^{\mathsf{T}} y = 0 \quad \forall\, y \in \mathcal{W}_j.$$

Denote $w^{(j)} := x^{(j)} - x^{(0)} \in \mathcal{W}_j$ and $e^{(0)} := \widehat{x} - x^{(0)}$, the above implies

$$0 = \left(\widehat{x} - x^{(j)}\right)^{\mathsf{T}} Ay = \left(e^{(0)} - w^{(j)}\right)^{\mathsf{T}} Ay = 0 \quad \forall\, y \in \mathcal{W}_j.$$

Therefore, $w^{(j)}$ is an A-orthogonal projection of $e^{(0)}$ onto the space $\mathcal{W}_j$ which gives us the best approximation property

$$\left\|e^{(0)} - w^{(j)}\right\|_A = \min_{w\in\mathcal{W}_j} \left\|e^{(0)} - w\right\|_A.$$

But since $w \in \mathcal{W}_j$, we know that $w = Q(A)r_0$ for some $Q \in \mathcal{P}_{j-1}$. Since $Ae^{(0)} = r^{(0)}$ and $e^{(0)} - w = (I_n - Q(A)A)e^{(0)} =: P(A)e^{(0)}$ we obtain

$$\left\|\widehat{x} - x^{(j)}\right\|_A = \left\|e^{(0)} - w^{(j)}\right\|_A = \min_{P\in\mathcal{P}_j,\, P(0)=1} \left\|P(A)e^{(0)}\right\|_A.$$

□

Note that for any symmetric matrix $A \in \mathbb{R}^{n\times n}$ there exists an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ ($V^{\mathsf{T}}V = I_n$) such that $A = V^{\mathsf{T}}\Lambda V$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix containing the eigenvalues of A on its diagonal (with the assumption $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$). This gives

$$P(A) = V^{\mathsf{T}}P(\Lambda)V,$$

therefore

$$\begin{aligned}\left\|P(A)e^{(0)}\right\|_A &= \sqrt{\left(e^{(0)}\right)^{\mathsf{T}} P(A)AP(A)e^{(0)}} \\ &= \left\|P(A)A^{\frac{1}{2}}e^{(0)}\right\|_2 \\ &\leq \left\|P(A)\right\|_2 \cdot \left\|A^{\frac{1}{2}}e^{(0)}\right\|_2 \\ &= \left\|V^{\mathsf{T}}P(\Lambda)V\right\|_2 \cdot \left\|e^{(0)}\right\|_A \\ &\leq \underbrace{\left\|V^{\mathsf{T}}\right\|_2 \cdot \left\|V\right\|_2}_{=1} \cdot \underbrace{\left\|P(\Lambda)\right\|_2}_{=\max_{\lambda\in\Lambda(P(A))}|\lambda|} \cdot \left\|e^{(0)}\right\|_A.\end{aligned}$$

Then with the *spectral radius of* A given by $\rho(P(A)) := \max_{\lambda\in\Lambda(P(A))}|\lambda| = \rho(P(\Lambda))$, for any polynomial $P \in \mathcal{P}_j$ we get

$$\begin{aligned}\left\|\widehat{x} - x^{(j)}\right\|_A &= \min_{P\in\mathcal{P}_j,\, P(0)=1} \left\|P(A)e^{(0)}\right\|_A \leq \min_{P\in\mathcal{P}_j,\, P(0)=1} \rho(P(A))\left\|e^{(0)}\right\|_A \\ &= \underbrace{\min_{P\in\mathcal{P}_j,\, P(0)=1} \max_{1\leq k\leq n} |P(\lambda_k)|}_{=:c(\lambda_1,\ldots,\lambda_n)} \left\|e^{(0)}\right\|_A.\end{aligned}$$

Moreover, it is clear that

$$\min_{P\in\mathcal{P}_j,\, P(0)=1} \max_{1\leq k\leq n} |P(\lambda_k)| \leq \min_{P\in\mathcal{P}_j,\, P(0)=1} \max_{\lambda\in[\lambda_1,\,\lambda_n]} |P(\lambda)| =: \min_{P\in\mathcal{P}_j,\, P(0)=1} \|P\|_{\infty,[\lambda_1,\lambda_n]}.$$

Now the question remains whether we can determine a polynomial $S_j \in \mathcal{P}_j$ such that

$$\|S_j\|_{\infty,[\lambda_1,\lambda_n]} = \min_{P\in\mathcal{P}_j,\, P(0)=1} \|P\|_{\infty,[\lambda_1,\lambda_n]}.$$

Indeed, we can express S_j by so-called *Chebyshev polynomials* which are defined by

$$T_j(\xi) := \frac{1}{2}\left(\left(\xi + \sqrt{\xi^2-1}\right)^j + \left(\xi - \sqrt{\xi^2-1}\right)^j\right).$$

Then one can show that

$$S_j(\lambda) = T_j\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)^{-1} T_j\left(\frac{\lambda_n+\lambda_1-2\lambda}{\lambda_n-\lambda_1}\right)$$

and

$$\|S_j\|_{\infty,[\lambda_1,\lambda_n]} = \left|T_j\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)\right|^{-1}. \tag{IV.16}$$

With these observations we can now conclude the following convergence result:

Theorem IV.8: The error after j steps of the CG algorithm can be bounded by

$$\left\|\widehat{x} - x^{(j)}\right\|_A \leq 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^j \left\|\widehat{x} - x^{(0)}\right\|_A,$$

where $\kappa(A) := \|A\|_2 \cdot \left\|A^{-1}\right\|_2 = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ where $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ denote the largest and the smallest eigenvalue of A, respectively.

Proof. With the considerations from above we must compute the value of (IV.16). First note that

$$\xi := \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1} = \frac{\frac{\lambda_n}{\lambda_1} + \frac{\lambda_1}{\lambda_1}}{\frac{\lambda_n}{\lambda_1} - \frac{\lambda_1}{\lambda_1}} = \frac{\kappa(A)+1}{\kappa(A)-1}.$$

This gives

$$\begin{aligned}\xi \pm \sqrt{\xi^2 - 1} &= \frac{\kappa(A)+1}{\kappa(A)-1} \pm \frac{2\sqrt{\kappa(A)}}{\kappa(A)-1} \\ &= \frac{\kappa(A)+1 \pm 2\sqrt{\kappa(A)}}{\kappa(A)-1} \\ &= \frac{\left(\sqrt{\kappa(A)} \pm 1\right)^2}{\left(\sqrt{\kappa(A)}-1\right)\left(\sqrt{\kappa(A)}+1\right)} = \frac{\sqrt{\kappa(A)} \pm 1}{\sqrt{\kappa(A)} \mp 1}.\end{aligned}$$

This gives us

$$T_j\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right) = \frac{1}{2}\left(\left(\frac{\sqrt{\kappa(A)}+1}{\sqrt{\kappa(A)}-1}\right)^j + \left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^j\right)$$

and

$$\|S_j\|_{\infty,[\lambda_1,\lambda_n]} = \left|T_j\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)\right|^{-1} = \frac{2}{\left(\frac{\sqrt{\kappa(A)}+1}{\sqrt{\kappa(A)}-1}\right)^j + \left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^j} \leq 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^j.$$

□

If $\kappa(A) \gg 1$, then the convergence of the CG method can be slow. If this case, one often uses a *preconditioner*. The idea consists of considering a linear system

Instead of $Ax = b$, we are interested in the *preconditioned* linear system

$$M^{-1}Ax = M^{-1}b, \tag{IV.17}$$

where M is the *preconditioner*. If the matrix M is positive definite and symmetric, then there exists a *Cholesky factorization* of the form $M = LL^\mathsf{T}$, where L is a lower triangular matrix with positive diagonal entries. Then (IV.17) is equivalent to solving

$$L^{-1}AL^{-\mathsf{T}}z = L^{-1}b \tag{IV.18}$$

with $x = L^{-\mathsf{T}}z$. Then we apply the CG method to (IV.18), where hopefully, we have that $\kappa\left(L^{-1}AL^{-\mathsf{T}}\right) \ll \kappa(A)$. Then the CG method should converge faster to the solution of (IV.18), which can be transformed back to the solution of (IV.17). There are several ways of determining preconditioners for the CG method. In general, a good preconditioner should fulfill the following properties:

- It should be cheap to apply the preconditioner. In the above considerations, this is for example the case, if the matrix L is sparse.
- It should hold that $L^{-1}AL^{-\mathsf{T}} \approx I_n$ (in a vague sense) or the eigenvalues of $L^{-1}AL^{-\mathsf{T}}$ should at least be concentrated in a few clusters. Recall that

$$\left\|\widehat{z} - z^{(j)}\right\|_{L^{-1}AL^{-\mathsf{T}}} \leq \min_{P\in\mathcal{P}_j,\, P(0)=1} \max_{1\leq k\leq n} |P(\lambda_k)| \left\|e^{(0)}\right\|_{L^{-1}AL^{-\mathsf{T}}},$$

  where $\widehat{z}$ is the exact solution of (IV.18), $z^{(j)}$ is the j-th iterate of the preconditioned CG iteration, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the eigenvalues of $L^{-1}AL^{-\mathsf{T}}$. If there are only a few eigenvalue clusters, then one can easily find a polynomial P of *low degree* with $P(0) = 1$ such that $P(\lambda_k)$ is small for all $k \in \{1, \ldots, n\}$, so only a few iterations are necessary to make the error small.

One popular preconditioner for the CG method is the incomplete Cholesky decomposition. Roughly speaking it consists of a Cholesky-like decomposition $A \approx LL^{\mathsf{T}}$ in which the lower triangular factor $L = [l_{ij}]_{i,j=1}^n$ is constructed in such a way that $l_{ij} = 0$ if $a_{ij} = 0$. Then L has the same sparsity pattern as A and since L is further triangular, it is very efficient to solve with L and L^{T}.

Remark IV.9: For all iterative solvers, we need a suitable stopping criterion. Mostly, for the (P)CG method one checks if

$$\begin{aligned} \left\|r^{(j)}\right\|_2 &< \text{tol} \quad \text{(small residual) or} \\ \frac{\left\|r^{(j)}\right\|_2}{\|b\|_2} &< \text{tol} \quad \text{(small relative residual).} \end{aligned}$$

Alternatively, one may use

$$\begin{aligned} \left\|x^{(j+1)} - x^{(j)}\right\|_2 &< \text{tol} \quad \text{or} \\ \frac{\left\|x^{(j+1)} - x^{(j)}\right\|_2}{\left\|x^{(j+1)}\right\|_2} &< \text{tol}\,. \end{aligned}$$

However, it is important to note that in general, a small residual does *not* imply that the error is small. Namely, the residual and the error are related via

$$\left\|\widehat{x} - x^{(j)}\right\|_2 \leq \left\|A^{-1}\right\| \cdot \left\|r^{(j)}\right\|_2.$$

A simple example where the residual is very small, but the error is very big is given by

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-16} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Then the exact solution is $\widehat{x} = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ and $\left\|A^{-1}\right\|_2 = 10^{16}$. Now with $x^{(j)} = \left[\begin{smallmatrix} 0 \\ 10^{8} \end{smallmatrix}\right]$, we see that

$$\left\|b - Ax^{(j)}\right\|_2 = \left\| - \begin{bmatrix} 0 \\ 10^{-8} \end{bmatrix} \right\|_2 = 10^{-8},$$

but

$$\left\|\widehat{x} - x^{(j)}\right\|_2 = \left\| - \begin{bmatrix} 0 \\ 10^8 \end{bmatrix} \right\|_2 = 10^8.$$

This observation is another motivation for precondition, since then $\left\|A^{-1}\right\|_2$ will in general be smaller, therefore, the computed solutions will also be more accurate.

IV.3. Multigrid Methods

First of all, we consider "classical" iterative methods that are based on a so-called *splitting* of the matrix A. A splitting of A is an additive decomposition $A = M - N$ with $\det(M) \neq 0$, where M should be easily invertible. Then it holds that $Ax = b$ is equivalent to

$$(M - N)x = b,$$

therefore

$$Mx = Nx + b \Leftrightarrow x = M^{-1}Nx + M^{-1}b.$$

The right-hand side of the latter equation can be interpreted as a function f that depends on x and x is a solution of the equation

$$f(x) = x,$$

one says that x is a fixed point of f. Then we can construct the fixed point iteration

$$x^{(j+1)} = M^{-1}Nx^{(j)} + M^{-1}b, \quad j = 0, 1, 2, \ldots,$$

where $x^{(0)}$ is a given initial vector. The question is whether this iteration converges and if yes, whether it converges to the exact solution $\widehat{x}$. Consider the error

$$\begin{aligned} \widehat{x} - x^{(j+1)} &= \widehat{x} - M^{-1}Nx^{(j)} - M^{-1}b \\ &= \widehat{x} - M^{-1}Nx^{(j)} - M^{-1}A\widehat{x} \\ &= \widehat{x} - M^{-1}Nx^{(j)} - M^{-1}(M - N)\widehat{x} \\ &= \widehat{x} - M^{-1}Nx^{(j)} - \big(I - M^{-1}N\big)\widehat{x} \\ &= M^{-1}N\big(\widehat{x} - x^{(j)}\big). \end{aligned}$$

We obtain

$$\widehat{x} - x^{(j+1)} = \big(M^{-1}N\big)^{j+1}\big(\widehat{x} - x^{(0)}\big).$$

Then for every norm $\|\cdot\|$ on $\mathbb{R}^n$ (and associated matrix norm $|||\cdot|||$) we get

$$\begin{aligned} \left\|\widehat{x} - x^{(j+1)}\right\| = \left\|(M^{-1}N)^{j+1}(\widehat{x} - x^{(0)})\right\| &\leq \left|\left|\left|(M^{-1}N)^{j+1}\right|\right|\right| \left\|\widehat{x} - x^{(0)}\right\| \\ &\leq \left|\left|\left|M^{-1}N\right|\right|\right|^{j+1} \left\|\widehat{x} - x^{(0)}\right\|. \end{aligned}$$

The method converges, if $\|\widehat{x}-x^{(j+1)}\| \to 0$ for $j \to \infty$. This is satisfied if $(M^{-1}N)^j \to 0$ for $j \to \infty$. For this, a sufficient condition is $\|\|M^{-1}N\|\| < 1$.

If $M^{-1}N$ is diagonalizable, i. e. there exists an invertible matrix $X \in \mathbb{R}^{n\times n}$ such that $M^{-1}N = XDX^{-1}$ with a diagonal matrix D, then it holds that $(M^{-1}N)^j = XD^jX^{-1}$ and $(M^{-1}N)^j \to 0$, if and only if $|d_{ii}| < 1$ for all i, where d_{ii} are the diagonal entries of D.

So, one would like to find a splitting of A, in which all eigenvalues of $M^{-1}N$ have a magnitude smaller than one (the smaller $\max_{1\le i\le n}|d_{ii}|$, the better, since this will lead to faster convergence).

More flexibility can be gained by introducing a so-called relaxation parameter ω: For all $\omega > 0$ it holds that

$$\begin{aligned} M\widehat{x} = N\widehat{x} + b \quad &\Leftrightarrow \quad \omega M\widehat{x} = \omega N\widehat{x} + \omega b \\ &\Leftrightarrow \quad (\omega-1)M\widehat{x} + M\widehat{x} = \omega N\widehat{x} + \omega b \\ &\Leftrightarrow \quad M\widehat{x} = \omega N\widehat{x} + (1-\omega)M\widehat{x} + \omega b \\ &\Leftrightarrow \quad \widehat{x} = \omega M^{-1}N\widehat{x} + (1-\omega)\widehat{x} + \omega M^{-1}b \\ &\Leftrightarrow \quad \widehat{x} = (\omega M^{-1}N + (1-\omega)I_n)\widehat{x} + \omega M^{-1}b. \end{aligned}$$

We obtain the relaxed fixed point iteration

$$x^{(j+1)} = R(\omega)x^{(j)} + \omega M^{-1}b, \quad j = 0,\, 1,\, 2,\, \ldots$$

with $R(\omega) = \omega M^{-1}N + (1-\omega)I_n$. When using $R(1)$ we obtain the *unrelaxed method*. We can now try to make the eigenvalues of $R(\omega)$ as small as possible to speed up convergence by a clever choice of ω.

An example for a relaxed splitting method is the *Jacobi splitting*: We split A as

$$\begin{aligned} A &= \text{proper lower triangle} + \text{diagonal} + \text{proper upper triangle} \\ &= L + D + U \end{aligned}$$

By choosing $M = D = \operatorname{diag}(d_{11}, \ldots, d_{nn})$ (we need $d_{ii} \neq 0$ for $i = 1, \ldots, n$ for invertibility) and $N = -(L+U)$ we have

$$R(\omega) = -\omega D^{-1}(L+U) + (1-\omega)I_n.$$

Example IV.10: Consider the matrix

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

Then

$$M = D = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad N = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

therefore,

$$R(\omega) = \omega \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + (1-\omega) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1-\omega & \frac{\omega}{2} \\ \frac{\omega}{2} & 1-\omega \end{bmatrix}.$$

We now consider the eigenvalues of $R(\omega)$. The characteristic polynomial of $R(\omega)$ is

$$(\lambda - (1-\omega))^2 - \frac{\omega^2}{4} = \lambda^2 - 2\lambda(1-\omega) + (1-\omega)^2 - \frac{\omega^2}{4}$$

and the eigenvalues are

$$\lambda_{1,2}(\omega) = 1 - \omega \pm \sqrt{\frac{\omega^2}{4}} = \begin{cases} 1 - \frac{\omega}{2}, \\ 1 - \frac{3}{2}\omega. \end{cases}$$

The optimal value is $\omega = 1$.

Now let A be the system matrix obtained by using the five-point stencil of the finite difference method for the differential operator $-\Delta$, that is,

$$A = \frac{1}{h^2} \begin{bmatrix} T & -I_n & & \\ -I_n & \ddots & \ddots & \\ & \ddots & \ddots & -I_n \\ & & -I_n & T \end{bmatrix} \in \mathbb{R}^{n^2 \times n^2} \text{ with}$$

$$T = \begin{bmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad h = \frac{1}{n+1}$$

Then by using Jacobi splitting and with

$$\widetilde{T} = \begin{bmatrix} 0 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 0 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

we get

$$\begin{aligned}
R(\omega) &= -\omega D^{-1}(L+U) + (1-\omega)I_{n^2} \\
&= -\frac{1}{4}\omega \begin{bmatrix} \widetilde{T} & -I_n & & \\ -I_n & \ddots & \ddots & \\ & \ddots & \ddots & -I_n \\ & & -I_n & \widetilde{T} \end{bmatrix} + (1-\omega)I_{n^2} \\
&= I_{n^2} - \left(\frac{1}{4}\omega \begin{bmatrix} \widetilde{T} & -I_n & & \\ -I_n & \ddots & \ddots & \\ & \ddots & \ddots & -I_n \\ & & -I_n & \widetilde{T} \end{bmatrix} + \omega I_{n^2} \right) \\
&= I_{n^2} - \frac{1}{4}\omega h^2 A.
\end{aligned}$$

Then the eigenvalues of $R(\omega)$ with Jacobi splitting are given by

$$\lambda_{k,l} = 1 - \omega + \frac{\omega}{2}\left(\cos\left(\frac{k\pi}{n+1}\right) + \cos\left(\frac{l\pi}{n+1}\right)\right) \qquad k,\, l = 1,\, \ldots,\, n.$$

Moreover, the eigenvectors of $R(\omega)$ are the same as the ones of A, they are given by

$$u_{k,l} = \begin{bmatrix} u_{k,l}^{1,1} \\ \vdots \\ u_{k,l}^{1,n} \\ \vdots \\ u_{k,l}^{n,1} \\ \vdots \\ u_{k,l}^{n,n} \end{bmatrix}$$

with

$$u_{k,l}^{p,q} = \gamma_{k,l} \sin\left(\frac{kp\pi}{n+1}\right) \sin\left(\frac{lq\pi}{n+1}\right), \quad p,\, q = 1,\, \ldots, n,$$

where $\gamma_{k,l}$ is chosen such that $\left\|u_{k,l}\right\|_2 = 1$.

With

$$\begin{aligned}
U &:= \begin{bmatrix} u_{1,1} & \ldots & u_{1,n} & \ldots & u_{n,1} & \ldots & u_{n,n} \end{bmatrix}, \\
\Lambda &:= \operatorname{diag}(\lambda_{1,1},\, \ldots,\, \lambda_{1,n},\, \ldots,\, \lambda_{n,1},\, \ldots,\, \lambda_{n,n})
\end{aligned}$$

we can write

$$R(\omega) = U\Lambda U^{\mathsf{T}}.$$

Moreover, assume we have

$$\begin{aligned}\widehat{x} - x^{(0)} &= \alpha_{1,1}u_{1,1} + \ldots + \alpha_{1,n}u_{1,n} + \ldots + \alpha_{n,1}u_{n,1} + \ldots + \alpha_{n,n}u_{n,n} \\ &= U\alpha\end{aligned}$$

for $\alpha = \begin{bmatrix}\alpha_{1,1} & \ldots & \alpha_{1,n} & \ldots & \alpha_{n,1} & \ldots & \alpha_{n,n}\end{bmatrix}^{\mathsf{T}}$. Therefore, we can write

$$\begin{aligned}\widehat{x} - x^{(j)} &= R(\omega)^j\left(\widehat{x} - x^{(0)}\right) \\ &= U\Lambda^j U^{\mathsf{T}} U\alpha \\ &= U\Lambda^j\alpha \\ &= \lambda_{1,1}^j\alpha_{1,1}u_{1,1} + \ldots + \lambda_{1,n}^j\alpha_{1,n}u_{1,n} + \ldots + \lambda_{n,1}^j\alpha_{n,1}u_{n,1} + \ldots + \lambda_{n,n}^j\alpha_{n,n}u_{n,n}.\end{aligned}$$

For $\omega = 1$ we have $|\lambda_{k,l}| \approx 1$ for k and l near 1 ("lower eigenfrequencies") as well as k and l near n ("high eigenfrequencies"). Therefore, in the Jacobi iteration $x_{j+1} = R(1)x_j + M^{-1}b$ low as well as high frequency parts in the error

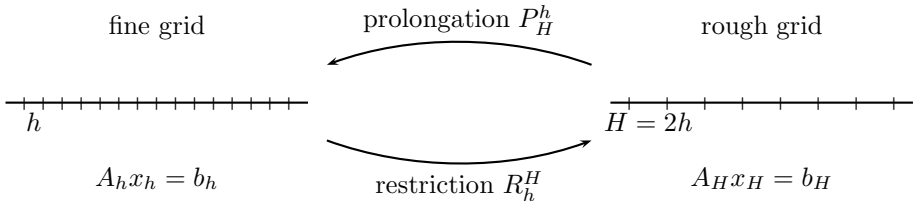
$$\widehat{x} - x^{(j+1)} = R(\omega)^{j+1}\left(\widehat{x} - x^{(0)}\right)$$

are "badly damped", i. e., these parts of the error tend to zero slowly.

On the other hand, for the relaxed iteration with $\omega = \frac{1}{2}$ it holds $|\lambda_{k,l}| \approx 0$ for k and l near n and $|\lambda_{k,l}| \approx 1$ for k and l near 1. In other words, high eigenfrequencies are damped well.

Observation IV.11: Relaxation methods smoothen the error by reducing its high-frequency parts. After a few steps, e. g., by applying $R(\frac{1}{2})$, the error hopefully only contains low-frequency parts.

The idea of geometric multigrid consists of reducing low-frequent parts in the error on a rough grid. The basic scheme is as follows:



a) Smoothening step I: Set $x_h^{(j)} = \text{Relax}\left(A_h, b_h, x_h^{(j-1)}, m_1\right)$ (e. g., m_1 steps of a Jacobi iteration).

b) Correction:

   i) Compute the defect $d_h^{(j)} := b_h - A_h x_h^{(j)}$. If the defect is zero, we have the exact solution.

   ii) Restriction: Set $d_H^{(j)} = R_h^H d_h^{(j)}$.

iii) Solve the *defect equation* $A_H x_H^{(j)} = d_H^{(j)}$.

iv) Prolongation: Set $\widetilde{x}_h^{(j+1)} := P_H^h x_H^{(j)}$.

v) Update the solution: Set $x_h^{(j+1)} := x_h^{(j)} + \widetilde{x}_h^{(j+1)}$.

c) Smoothening step II: Set $x_h^{(j+2)} = \text{Relax}\left(A_h, b_h, x_h^{(j+1)}, m_2\right)$.

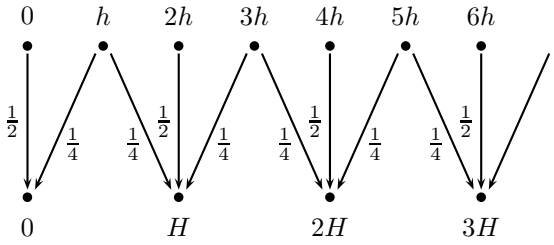
What is the "defect equation"? Let $\widetilde{x}$ be an approximation of the solution of $Ax = b$. Then $r = b - A\widetilde{x}$ is the residual and $e = \widehat{x} - \widetilde{x}$ is the error. Then $Ay = r$ is called the "defect equation". Its solution fulfills

$$y = A^{-1}r = A^{-1}\left(b - A\widetilde{x}\right) = \widehat{x} - \widetilde{x} = e,$$

i. e., the solution of the defect equation is the error. Moreover, $y + \widetilde{x} = \widehat{x}$ is the exact solution of $Ax = b$.

In multigrid methods, the defect equation is not solved with the "large" matrix A_h (which would give x_h), but with the small matrix A_H. Prolongation and restriction are typically given by *linear interpolation*.

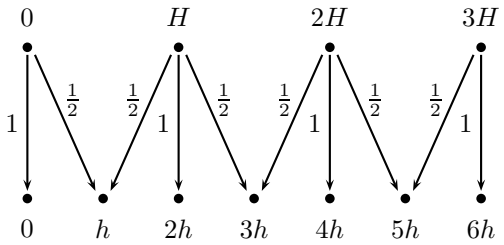
Example IV.12: The restriction



is given by

$$R_h^H = \frac{1}{4}\begin{bmatrix} 2 & 1 & & & & & & \\ & 1 & 2 & 1 & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & 2 & 1 \\ & & & & & & 1 & 2 \end{bmatrix}.$$

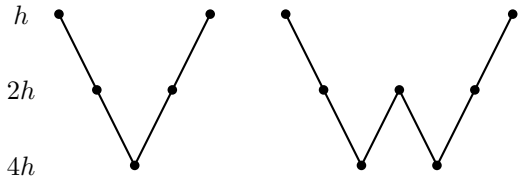
The prolongation



is given by

$$P_H^h = \begin{bmatrix} 1 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & \frac{1}{2} & \frac{1}{2} \\ & & & & 1 \end{bmatrix} = 2(R_h^H)^\mathsf{T}.$$

Often, multiple steps of the basic scheme are carried out on a hierarchical grid.



V cycle W cycle

The applicability is strongly depending on the PDE. Convergence proofs are typically hard.

Remark IV.13: So called algebraic multigrid methods are similar to geometric ones. In contrast to geometric multigrid methods there does not exist an actual grid (on which a solution is defined), so an analogous choice of the restriction and prolongation matrices as in geometric multigrid is not possible. There are many possible choices for such matrices that lead to different methods (such as there are many different Krylov subspace methods for solving linear systems).

Bibliography

[Cow73] GR Cowper. Gaussian quadrature formulas for triangles. *International Journal for Numerical Methods in Engineering*, 7(3):405–408, 1973.

[Dun85] David A Dunavant. High degree efficient symmetrical gaussian quadrature rules for the triangle. *International journal for numerical methods in engineering*, 21(6):1129–1148, 1985.

[Eva98] L. C. Evans. *Partial Differential Equations*, volume 19 of *Grad. Stud. Math.* AMS, Providence, RI, USA, 1998.

[Fel04] Carlos A Felippa. A compendium of fem integration formulas for symbolic work. *Engineering Computations*, 21(8):867–890, 2004.

[Hac92] W. Hackbusch. *Elliptic Differential Equations: Theory and Numerical Treatment*, volume 18 of *Springer Ser. Comput. Math.* Springer-Verlag, Berlin, Heidelberg, 1st edition, 1992.

[HB09] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens.* Vieweg+Teubner Verlag, 3rd edition, 2009.

[LT03] S. Larsson and V. Thomée. *Partial Differential Equations with Numerical Methods*, volume 45 of *Texts Appl. Math.* Springer-Verlag, Berlin, Heidelberg, 1st edition, 2003.

[Ste14] Erwin Stein. History of the finite element method–mathematics meets mechanics–part i: Engineering developments. In *The History of Theoretical, Material and Computational Mechanics-Mathematics Meets Mechanics and Engineering*, pages 399–442. Springer, 2014.

[Wac13] E. Wachspress. *The ADI Model Problem.* Springer-Verlag, New York, NY, USA, 1st edition, 2013.

[Wal00] Noel Walkington. *Quadrature on simplices of arbitrary dimension.* Carnegie Mellon University, Department of Mathematical Sciences, Center for . . . , 2000.

[Zwi98] D. Zwillinger. *Handbook of differential equations*, volume 1. Gulf Professional Publishing, 1998.