Lecture Notes of the Course

Numerical Mathematics II for Engineers

Technische Universität Berlin
Winter Term 2019/20
lectured by Dr. Dirk Peschka

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

Latest changes: November 8, 2019

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

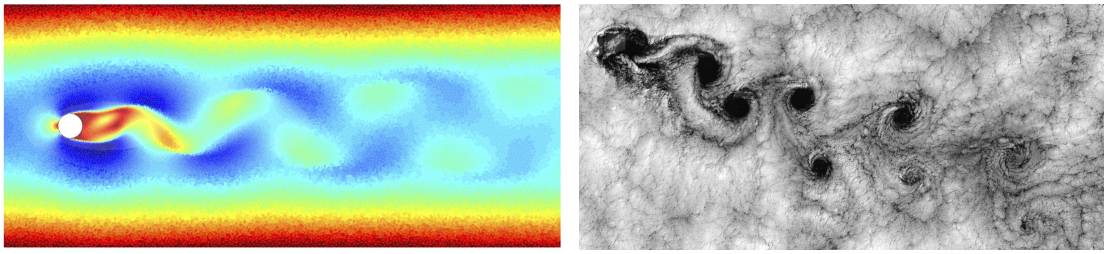


Figure I.1.: **(left)** Kármán vortex street as an instationary solution of the incompressible Navier-Stokes equation (I.1b) around an obstacle and **(right)** Landsat 7 image by NASA showing clouds near the Juan Fernandez Islands.

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

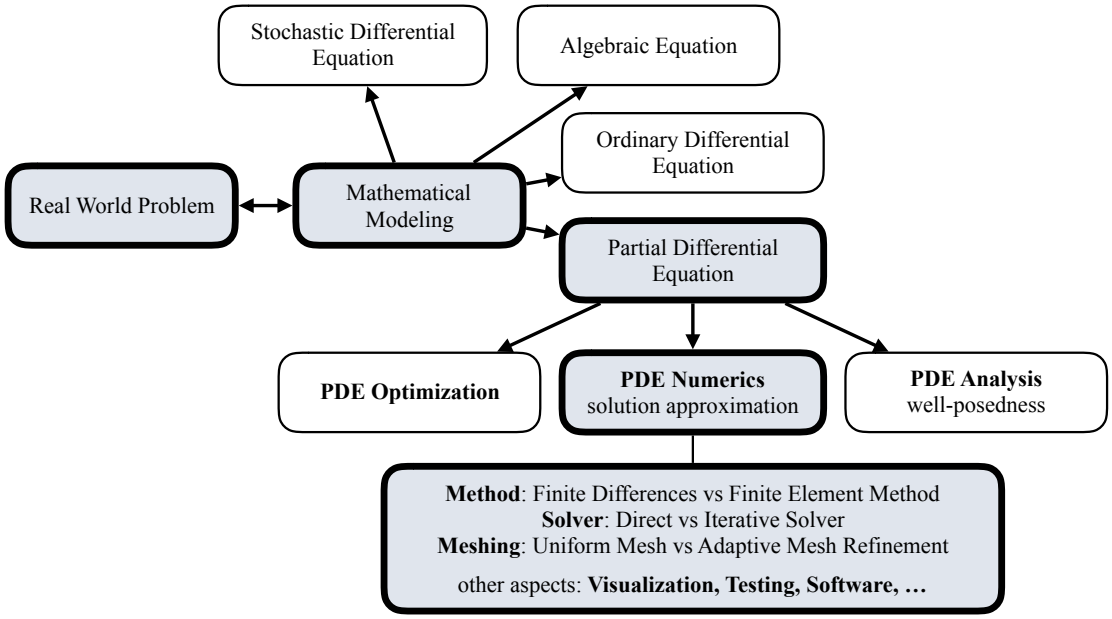


Figure I.2.: From Real World Problems to PDE Numerics

It is important to know that in this general setting there are many sources for errors: The modeling step itself present 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. Getting the hand on the 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 a 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

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

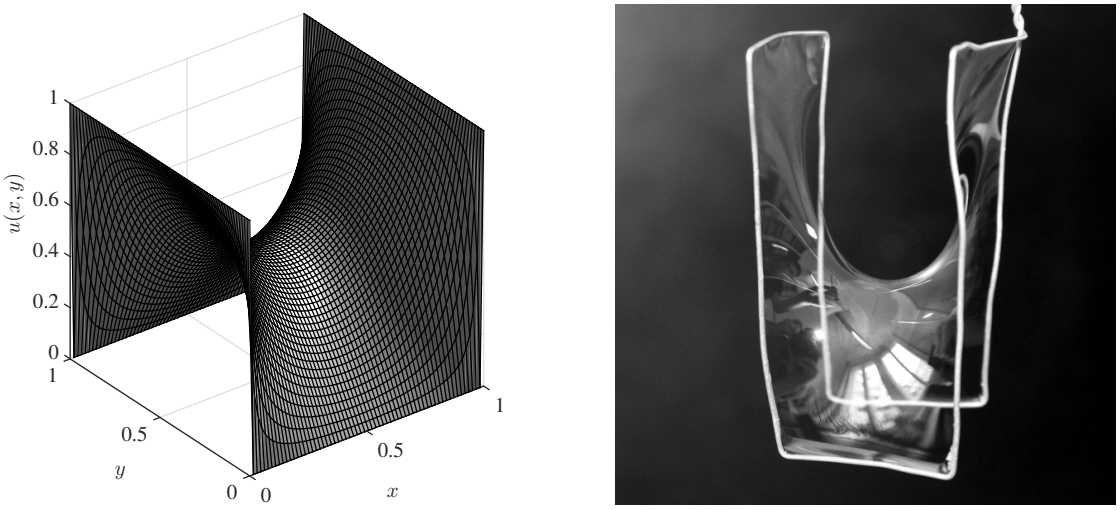


Figure I.3.: **(left)** Solution $u(x, y)$ of 2-dimensional homogeneous Poisson equation (I.5) on $\Omega = [0, 1]^2$ with boundary conditions $g(0, y) = g(1, y) = 1$ for $0 < y < 1$ and $g(x, 0) = g(x, 1) = 0$ for $0 < x < 1$ and **(right)** corresponding minimal (soap) surface on a wire. Note, the 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 give 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 district 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

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

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 denotes 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 $u \in C^k(\Omega)$ is u is k times differentiable and if all k-th partial derivative 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 the 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.

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

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

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 $\nabla \times \mathbf{f} = \operatorname{curl}(\mathbf{f})$ denotes the curl of a vector field defined as

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

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 $Df(x)(\mathbf{u})$ or 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 to 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 to many conditions or have a too restrictive solution concept. If infinitely many solutions exist, then we might have set to 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 in c)**:** 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.

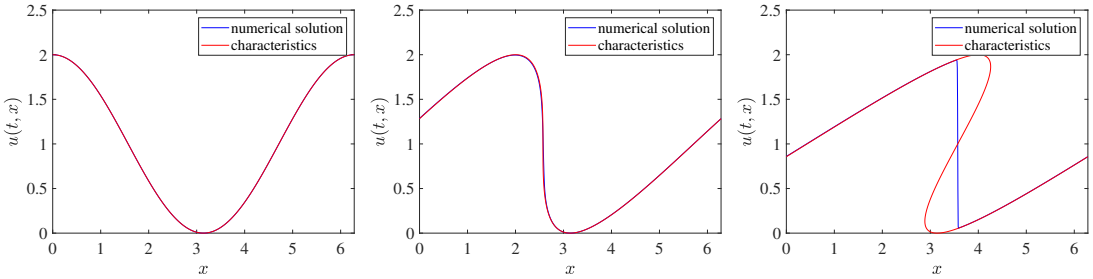


Figure I.4.: Solution of the Burgers equation at **(left)** $t = 0$ **(middle)** $t = 0.5$ and **(right)** $t = 1$.

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) ode 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 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 $T = L/V$. The 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 application 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 expression 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 for 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

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

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 an 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 $u_0(x_0) = 1 + \cos(x_0)$.

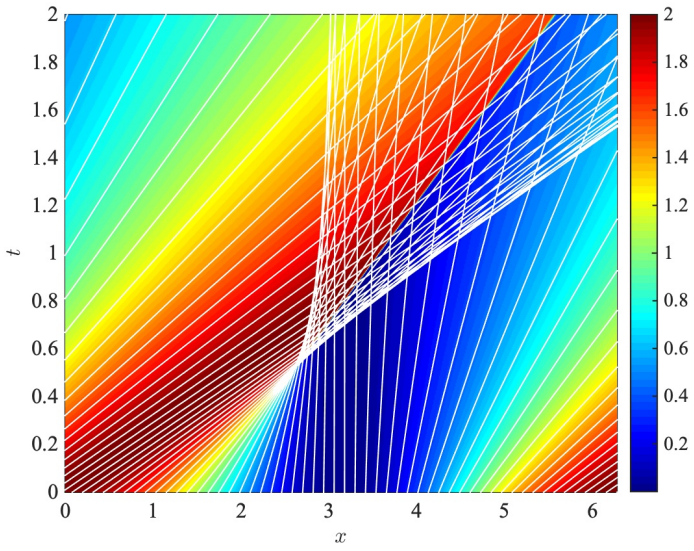


Figure I.5.: Numerical solution of Burgers equation with shading showing $u(t,x)$ at (t,x) and white lines are the characteristic curves $(t,\xi(t))$ in a periodic domain.

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 a 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 employes 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

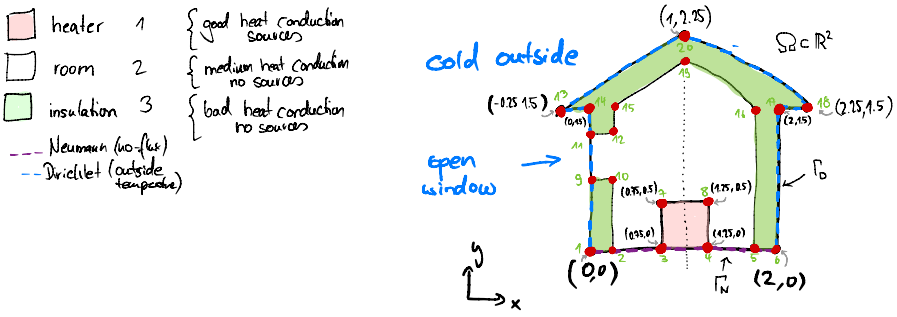


Figure I.6.: Geometric model of (diffusive) heat transport in house.

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.

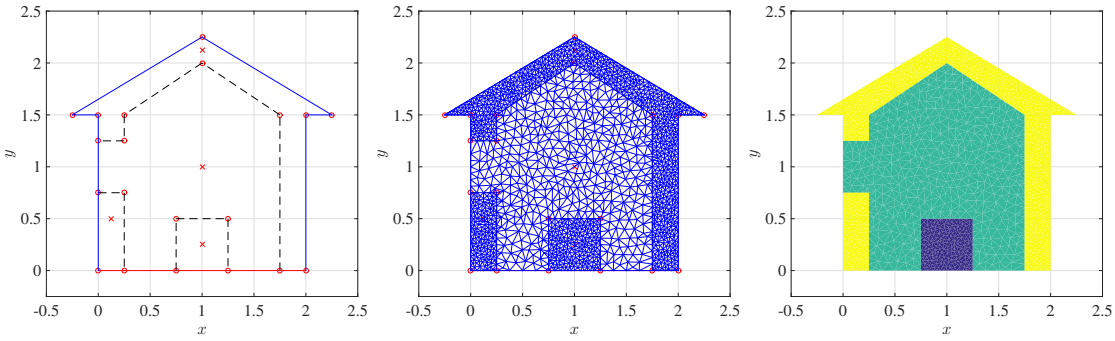


Figure I.7.: **(left)** CAD geometry description from 20 vertices (red dots) with line segments for interior interfaces (black dashed), Dirichlet boundary (blue full) and insulating boundary (red full) **(middle)** computational mesh (triangulation **(right)** distinct subdomains Ω_i encoded with different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

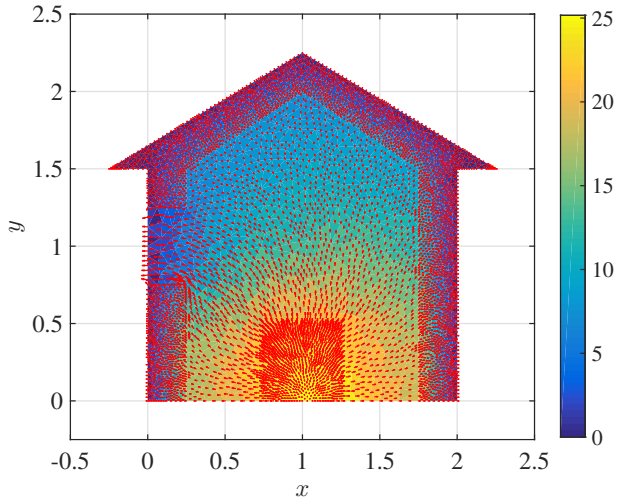


Figure I.8.: Numerical solution of the stationary heat flow problem (linear, elliptic PDE) showing temperature $T : \Omega \to \mathbb{R}$ using shading/color and the heat flux $\mathbf{q} : \Omega \to \mathbb{R}^2$ using red arrows.

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 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 $\overline{\Omega} = [0, L_1] \times [0, L_n] \subset \mathbb{R}^n$ which depending on n would be called interval $n = 1$, square/reactangle $n = 2$, 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

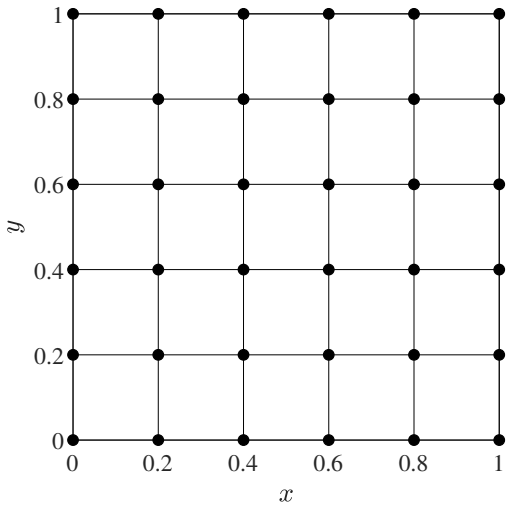


Figure II.1.: Example discretization mesh

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., hyperreactangles 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 one 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 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)`.

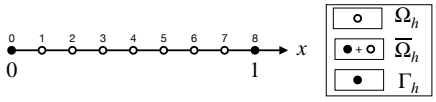


Figure II.2.: Example discretization 1D mesh

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

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

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

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

In fact, (II.DP) is a linear system of equations. Then $x_0 = 0$ and $x_{N+1} = 1$. For the unknowns $u_h(x_0), \ldots, u_h(x_{N+1})$ 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 $n \in \{1, \ldots, N\}$. 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 $u_h(x_1), \ldots, u_h(x_N)$. 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.

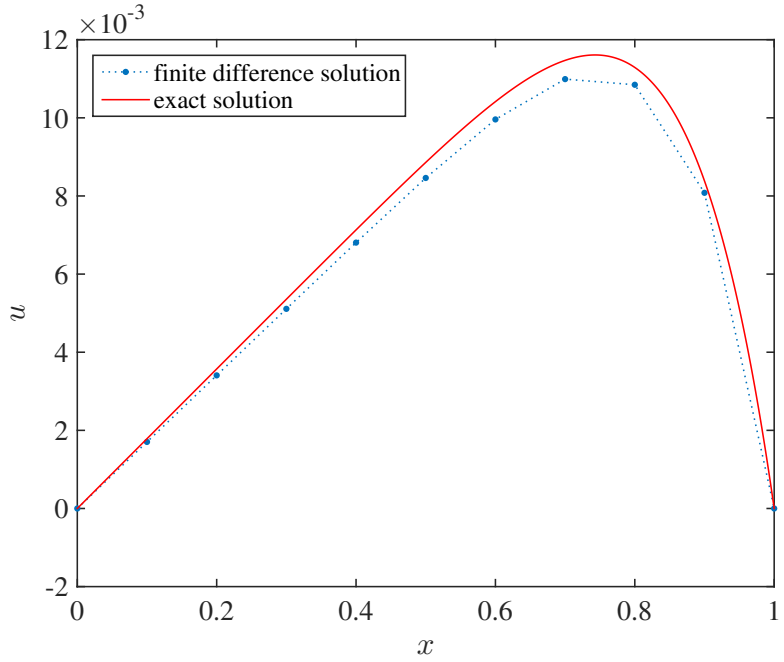


Figure II.3.: Comparison of numerical solution of $L_h u_h = f$ on $\overline{\Omega}_h$ generated from $\overline{\Omega} = [0,1]$ with on a coarse grid with $h = 1/(N+1) = 1/10$ and parameters $a = 1, \alpha = \beta = b = c = 0$ with $f(x) = x^6$ and comparison with exact solution $u(x) = 1/56(x -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

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

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 $\Delta_{h,k}$ 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 are represent compact difference stencils 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.

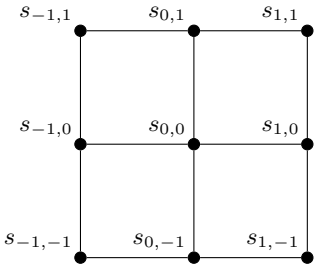


Figure II.4.: General compact 9-point stencil in 2D

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 if with $\delta x_n = (x_{n+1/2} - x_{n-1/2})$, 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 $s_{\pm1,0,0} = s_{0,\pm1,0} = s_{0,0,\pm1} = 1$ 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 $-1 \leq i, j, k \leq 1$. 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 $\sum_i |\alpha_i| = 1$.

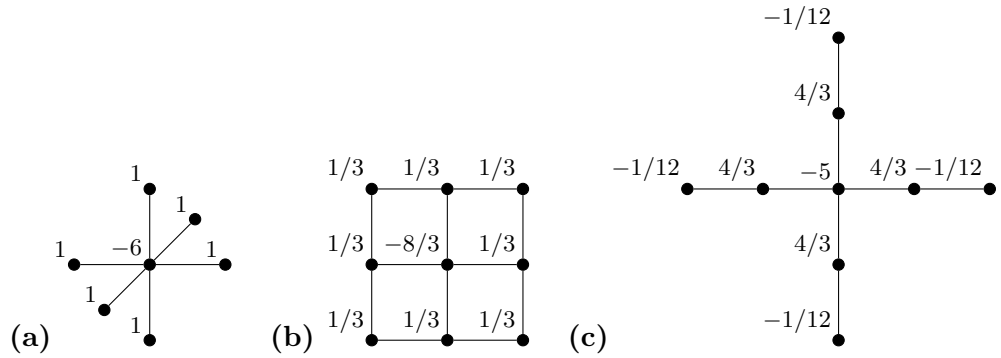


Figure II.5.: **(a)** standard 7-point difference stencil in 3D **(b)** alternative compact 9-point stencil in 2D **(c)** alternative noncompact 9-point stencil in 2D with consistency order 4, assuming $u \in C^6(\overline{\Omega})$.

One can show that in two dimensions there is no compact 9-point stencil of consistency order 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 mathmatical 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

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

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, in particular, 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:

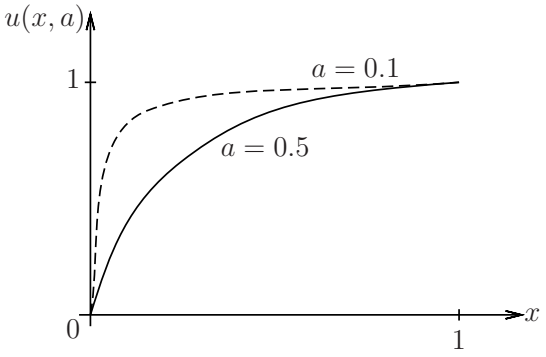
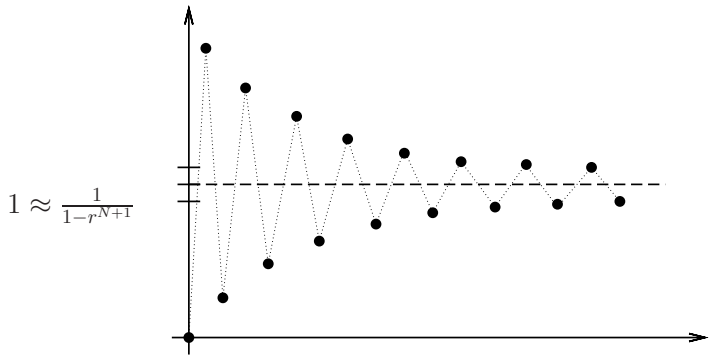


Figure II.6.: Exact solution to (II.16) for two different values of the parameter a.

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

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

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 ever 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.
- 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, the 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.

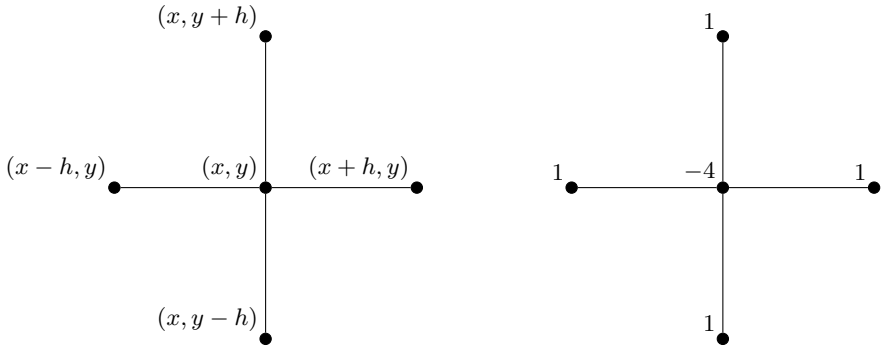


Figure II.7.: 5-point difference stencil for $-\Delta u$ in two spatial dimensions showing (left) to vertices/grid points and (right) the corresponding weights.

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:

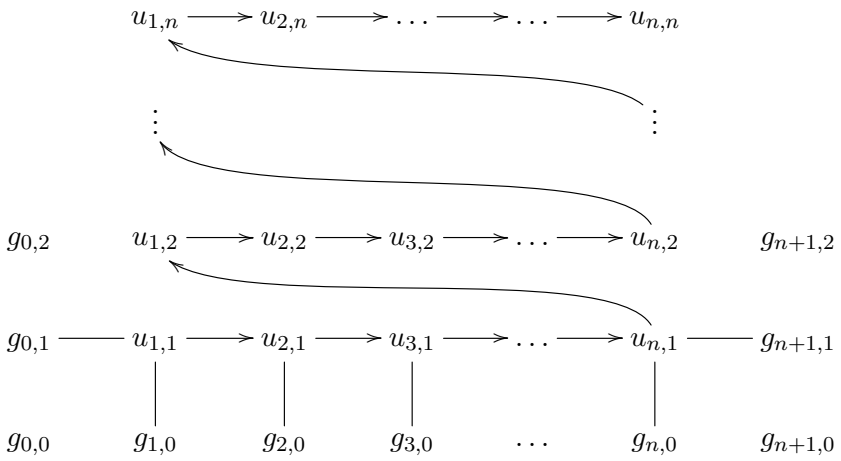
$$\begin{aligned} u_{k,l} &:= u_h(kh, lh), \quad k, l = 0, \ldots, N+1, \\ f_{k,l} &:= f(kh, lh), \quad k, l = 1, \ldots, N, \\ g_{k,l} &:= g(kh, lh), \quad (kh, lh) \in \Gamma_h. \end{aligned}$$

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

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

This leads to a system of linear equations on the N^2 unknowns $u_{k,l}$, $k, l = 1, \ldots, N$. We want to write this as $L_h u_h = f_h$. Here, $L_h \in \mathbb{R}^{N^2 \times N^2}$, $u_h \in \mathbb{R}^{N^2}$, $f_h \in \mathbb{R}^{N^2}$.

Book keeping is important: Keep track of which entry in u_h corresponds to which grid point. The standard way is to use the lexicographical ordering:



Then we get

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

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.8.

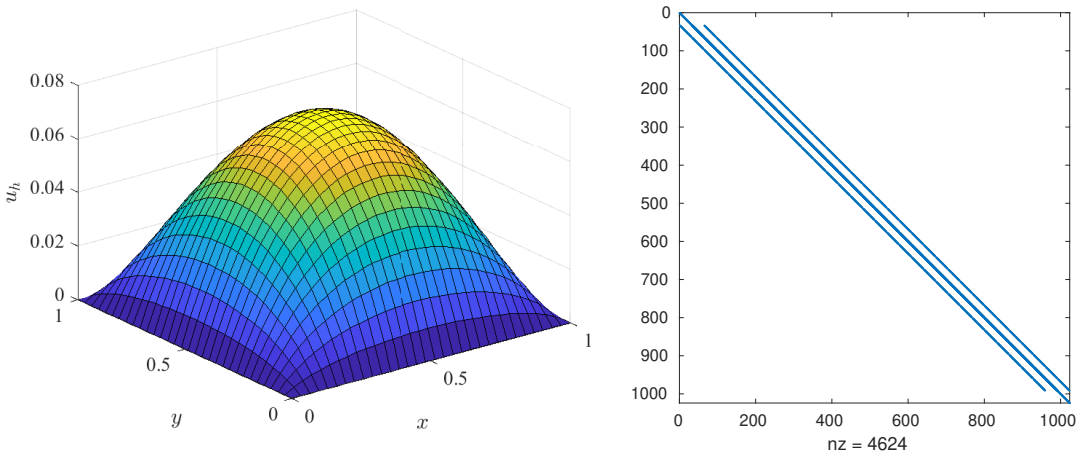


Figure II.8.: **(left)** Numerical solution u_h of the 2D Poisson problem $L_h u_h = f$ with homogeneous Dirichlet boundary conditions on $\Omega = (0,1)^2$ with $f(x) = 1$ using the standard 5-point stencil for $Lu = -\Delta u$. **(right)** Sparsity pattern of the matrix $L_h \in \mathbb{R}^{(N+2)^2 \times (N+2)^2}$ (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

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

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

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

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

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

or, equivalently,

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

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

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

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) $\left\|\left|L_h^{-1}\right|\right\|_\infty \le \frac{1}{8}$ 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 $w_h = R_h w$ 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 $L_h w_h$ is

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

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

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

□

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 and all points can be reached by 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 function with $\Delta_h u_h \geq 0$ on Ω_h. Then $\max_{\Omega_h} u_h \leq \max_{\Gamma_h} u_h$ and equality holds if and only if u_h is constant.

Proof by contradiction. Suppose $\max_{\Omega_h} u_h > \max_{\Gamma_h} u_h$ and let $x_0 \in \Omega_h$ where the maximum is attained and let x_i for $i = 1 \dots 2n$ its nearest neighbors. Using the standard $(2n+1)$-point stencil we have

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

where in the last step we used $u_h(x_i) \leq u(x_0)$. This requires $u_h(x_i) = u_h(x_0)$ 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 (or bigger on the boundary). □

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

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

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.

Proof. Consider a domain $\overline{\Omega}_h$ contained in ball of radius R around x_0. Defining $\phi(x) = R^2 - \|x - x_0\|^2$, then $-\Delta_h \phi = 2n > 0$ and $0 \le \phi \le R^2$. Now define the function $v(x) = \max_{\Gamma_h} |g| + \frac{1}{2n}\phi(x) \max_{\Omega_h} |f|$. By construction we have $-\Delta_h v = \max_{\Omega_h} |f| \ge |-\Delta_h u_h|$ in Ω_h and $v \ge |u|$ on Γ_h. Apply maximum principle to $\pm(u_h - v)$ and find $-v \le u^h \le v$ and hence $|u^h| \le C \max_{\Omega_h} |f| + \max_{\Gamma_h} g$ with $C = R^2/(2n)$. □

Corollary II.25: The constant C in Thm. II.24 also provides constant for the stability estimate $\left\|\left\|L_h^{-1}\right\|\right\|_h \le C$. 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

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

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

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

and therefore, we get

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

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 $\widetilde{h} \ll h$.

Remark II.26: The matrix L_h 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 $\text{const.} \cdot (nmpq)$ 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.

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 exist, integrate (PP) over the domain, i.e.,

$$\int_\Omega f \mathrm{d}x = \int_\Omega (-\Delta u) \mathrm{d}x \overset{\text{Gauss}}{=} \int_\Gamma -\frac{\partial u}{\partial \nu} \,\mathrm{d}A = -\int_\Gamma g \,\mathrm{d}A, \tag{II.19}$$

which turns out to be a nontrivial condition on the data. This leads us to conclude that solutions of (PP) can only exist (but are nonunique), if the data f, g satisfy the *solvability condition* (II.19). However, even with such a solvability condition, these solution are 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

$$C[u] = \int_\Omega u \,\mathrm{d}x = \sigma, \tag{II.20}$$

which removes at 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 (PP) can be found by minimizing the functional

$$A[u] = \int_\Omega \left(\tfrac{1}{2} |\nabla u(x)|^2 - f(x)u(x) \right) \,\mathrm{d}x - \int_\Gamma g(x)u(x) \,\mathrm{d}A, \tag{II.21}$$

over all suitable functions u, without imposing conditions on Γ. In order to satisfy (II.20) we minimize (II.21) subject to the constraint (II.20), which introduces a scalar Lagrange multiplier $\lambda \in \mathbb{R}$ such that we need to solve the modified problem

$$-\Delta u + \lambda = f \quad \text{in } \Omega \subset \mathbb{R}^n, \qquad \frac{\partial u}{\partial \nu} = g \quad \text{on } \Gamma = \partial\Omega, \qquad \int_\Omega u \,\mathrm{d}x = \sigma. \tag{II.22}$$

Then there are two alternatives:

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

2. The data does not satisfy the solvability condition: Then $\lambda \neq 0$ such that u is the unique solution of the modified problem, 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 $\Omega_h = \{jh \mid 1 \le j \le N\}$. The boundary is then $\Gamma_h = \{0, 1\}$. The second order derivative is replaced by the usual 3-point difference stencil

$$-\frac{1}{h^2}\begin{pmatrix} 1 & -2 & 1 \end{pmatrix} u_h = f \qquad \text{in } \Omega_h, \tag{II.23}$$

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

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

or equivalently,

$$\frac{1}{h} g_0 = \frac{1}{h^2}(u_h(h) - u_h(0)), \qquad \frac{1}{h} g_1 = \frac{1}{h^2}(u_h(1) - u_h(1-h)). \tag{II.24}$$

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.24) we obtain

$$u_h(0) = u_h(h) - h g_0, \quad u_h(1) = u_h(1-h) + h g_1.$$

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

$$\begin{aligned} f(h) &= -\frac{1}{h^2}(u_h(0) - 2u_h(h) + u_h(2h)) \\ &= -\frac{1}{h^2}(u_h(h) - h g_0 - 2u_h(h) + u_h(2h)). \end{aligned}$$

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

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

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

$$h\sum_{n=1}^{N} f(x_n) = -(g_1 - g_0), \tag{II.25}$$

corresponding to (II.19).

However, the discrete solvability condition (II.25) may not be satisfied after discretizing the continuous solvability condition.

How do we then 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 arbitrary (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.25) is satisfied for $f_h - \lambda \mathbb{1}$.

Remark II.27: 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,

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

and

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

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

$$\frac{h}{2}\big(f(0) + f(1)\big) + h\sum_{n=1}^{N} f(x_n) = -(g_1 - g_0), \tag{II.26}$$

which by midpoint rule is a second-order approximation of the continuous solvability condition (II.19). 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.28 (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 two 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

$$\alpha u_h(0) + \frac{\beta}{2h}(u_h(h) - u_h(-h)) = g(0), \tag{II.27}$$

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.29 (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

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

e.g. see [Hac92].

II.6.2. Periodic Boundary conditions

For simplicity we counder 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

$$\int_0^1 f(x)\,\mathrm{d}x = 0. \tag{II.28}$$

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

$$\frac{-1}{h^2}\begin{pmatrix} 1 & -2 & 1 \end{pmatrix} u_h(x_n) = f(x_n) \tag{II.29}$$

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

$$L_h^p u_h - \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{bmatrix}\begin{bmatrix} u_h(0) \\ u_h(h) \\ \vdots \\ u_h(1-2h) \\ u_h(1-h) \end{bmatrix} = \begin{bmatrix} f(0) \\ f(h) \\ \vdots \\ f(1-2h) \\ f(1-h) \end{bmatrix} \qquad \text{in } \mathbb{R}^{N+1},$$

with compatibility condition

$$\sum_{i=0}^{N} f(x_i)h = 0. \tag{II.30}$$

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

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

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

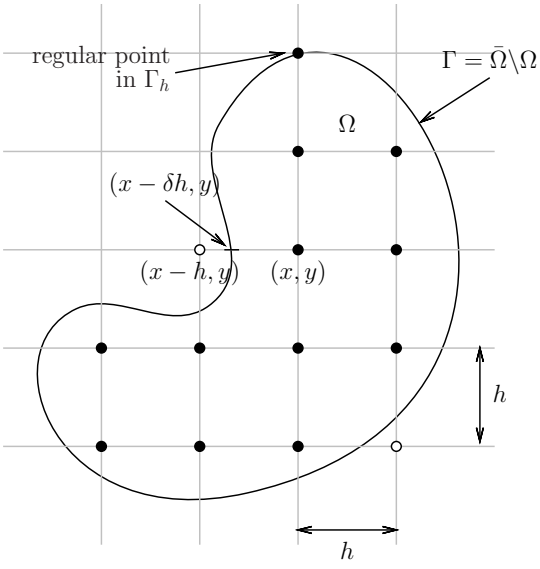
Consider again

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

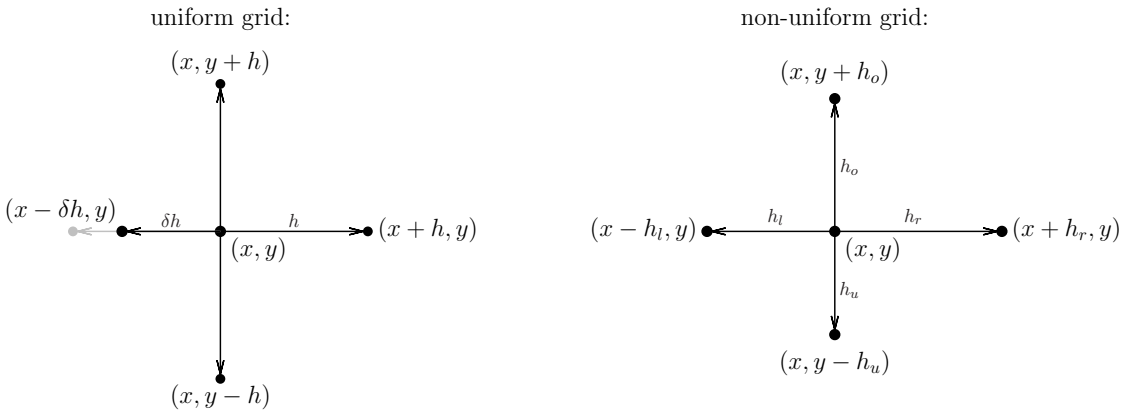
Here, Ω is not necessarily a "nice" domain like the unit square. One way to define an equidistant mesh is

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

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

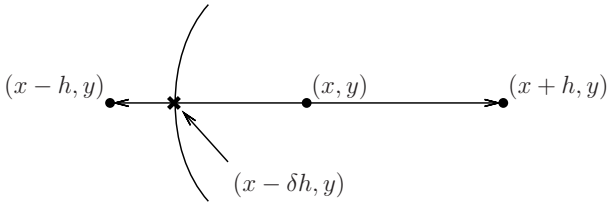
$$\begin{array}{ccccc} & & \frac{2}{h_o(h_u+h_o)} & & \\ & & | & & \\ \frac{2}{h_l(h_l+h_r)} & — & -\left(\frac{2}{h_l h_r}+\frac{2}{h_o h_u}\right) & — & \frac{2}{h_r(h_l+h_r)}. \\ & & | & & \\ & & \frac{2}{h_u(h_u+h_o)} & & \end{array}$$

Remark II.30: 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.31: The consistency is the same as above (order 1).

Theorem II.32: 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

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

where

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

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

$$Lv_k = \lambda_k v_k \qquad \text{in } \Omega \tag{II.31}$$

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.33 (Laplace operator in 1D)**:** For example, consider $\Omega = (0, 1)$ and the problem

$$-v_k''(x) = \lambda_k v_k(x), \qquad \text{in } (0, 1). \tag{II.32}$$

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 $v_k^1 = \sin(2n\pi x)$ 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 $\lambda = 4kn^2\pi^2$, 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

$$\partial_t u + Lu = 0, \qquad \text{in } Q_T = (0, \infty) \times \Omega, u(t = 0, \cdot) = u_0, \qquad \text{in } \Omega, \tag{II.33}$$

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.,

$$u_0(x) = \sum_k a_k v_k(x), \tag{II.34}$$

then we find the solution

$$u(t, x) = \sum_k a_k v_k(x) \exp(-\lambda_k t). \tag{II.35}$$

For inhomogeneous boundary conditions we simply find

$$u(t,x) = u_{\text{hom}}(x) + \sum_k \bar{a}_k v_k(x) \exp(-\lambda_k t), \qquad u_0 - u_{\text{hom}} = \sum_k \bar{a}_k v_k(x), \tag{II.36}$$

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 $L_h \in \mathbb{R}^{N\times N}$. Then we can compute

$$\left\| L_h^{-1} \right\|^2 = \sup_{v\in\mathbb{R}^N, \|v\|=1} \|L_h^{-1} v\|^2 = \sup_{\xi} \sum_k \lambda_k^{-1} \xi_k = \lambda_{\min}^{-1}. \tag{II.37}$$

where $0 \le \xi_k \le 1$ and $\sum_k \xi_k = 1$. In this sense the eigenvalues provided in Example II.33 provide stability/continuity constants for the elliptic boundary value problem. In case b,c) we have $\left\| L_h^{-1} \right\| = \infty$. 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

$$\begin{bmatrix} L_h & \mathbb{1} \\ \mathbb{1}^\top & 0 \end{bmatrix} \bar{v}_h^k = \lambda_k \begin{bmatrix} \mathbb{I}_N & 0 \\ 0 & 0 \end{bmatrix} \bar{v}_h^k \tag{II.38}$$

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.

Example II.34 (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

$$-\frac{1}{h^2}(\ 1, -2,\ 1) v_k = \lambda_k v_k \qquad \text{in } \Omega_h \tag{II.39}$$

and $u_h(0) = u_h(1) = 0$. We make the ansatz $w(x_n) = e^{i\mu n h}$ and plug this into the the finite difference quotient

$$\begin{aligned} L_h w(x_n) &= -\frac{1}{h^2}(e^{i\mu h(n-1)} + e^{i\mu h(n+1)} - 2e^{i\mu h n}) = \frac{-1}{h^2}(e^{-i\mu} + e^{+i\mu} - 2) e^{i\mu h n} \\ &= \frac{1}{h^2}(2 - (e^{-ih\mu} + e^{+hi\mu})) w(x_n) = \lambda w(x_n) \end{aligned}$$

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

$$v_k(x_n) = \sin(\pi k n h), \qquad \lambda_k = 4h^{-2}\sin^2(\tfrac{h\pi k}{2}). \tag{II.40}$$

We can expand the eigenvalues for $h \to 0$ and fixed k as

$$\lambda_k = 4h^{-2}\sin^2(\tfrac{h\pi k}{2}) = \pi^2 k^2 + \mathcal{O}(h^2), \tag{II.41}$$

which coincides with the exact eigenvalues to leading order as $h \to 0$.

Example II.35 (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

$$-\Delta v = \lambda v, \qquad \text{in } \Omega, \tag{II.42}$$

$$v = 0, \qquad \text{on } \partial\Omega. \tag{II.43}$$

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

$$-(R''F + r^{-1}R'F + r^{-2}RF'') = \lambda RF, \tag{II.44}$$

or equivalently

$$-\frac{r^2(R'' + r^{-1}R' - \lambda R)}{R} = -\frac{F''}{F} = \sigma \in \mathbb{R}, \tag{II.45}$$

where $\sigma \geq 0$ such that $F(\phi) = a_n \cos(n\varphi) + b_n \cos(n\varphi)$ for $n \in \mathbb{N}_0$ and $\sigma = n^2$.

$$r^2R'' + rR' + (n^2 - r^2\lambda)R = 0 \tag{II.46}$$

Transforming $R(r) = a(x)$ with $x = \sqrt{\lambda}r$ we get

$$x^2a'' + xa' + (n^2 - x^2)a = 0, \tag{II.47}$$

which is Bessel's differential equation. Since we want the solution to be finite at the origin $x = 0$, we have $a(x) = J_n(x)$ where J_n are Bessel functions of the first kind as shown in Fig. II.10.

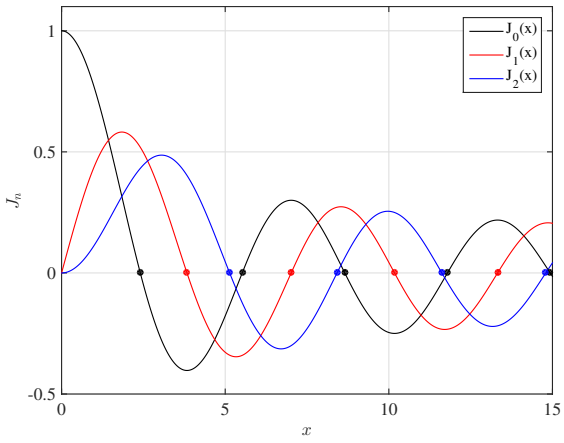


Figure II.9.: Bessel functions of first kind $J_n(x)$ for $n = 0, 1, 2$ with corresponding zeros.

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.

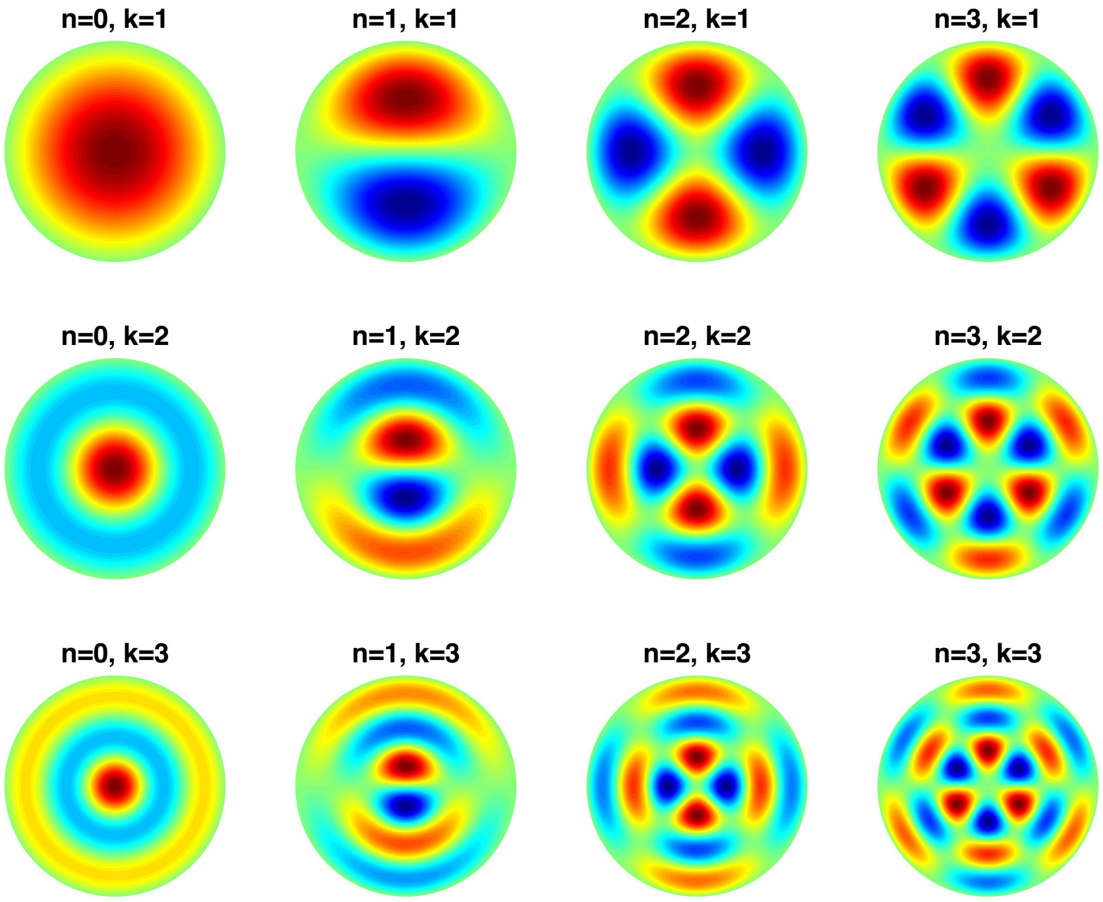


Figure II.10.: Eigenfunctions $v = \cos(n\varphi)J_n(rK_{n,k})$ of Laplace operator where $K_{n,k} = \sqrt{\lambda}$ is the k-th zero of the Bessel function J_n (Note: multiplicity).

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

$$\partial_t u + Lu = f, \qquad \text{in } Q_T \tag{II.48}$$

$$u(t=0,\cdot) = u_0, \qquad \text{in } \Omega \tag{II.49}$$

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.11.

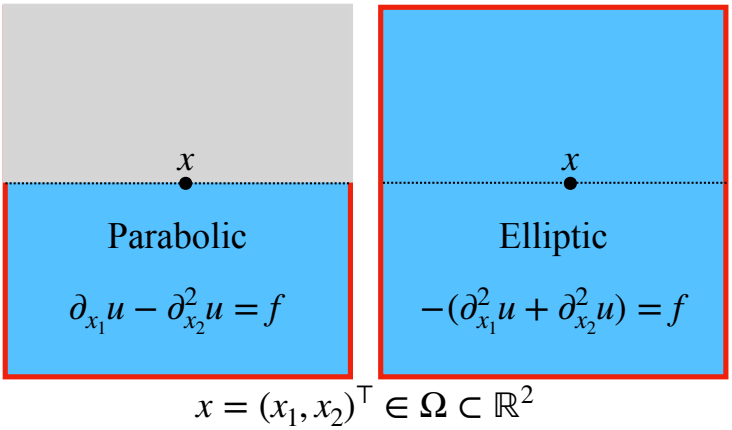


Figure II.11.: **Domain of dependence:** The solution $u(x)$ at a point $x = (x_1, x_2)^\top$ depends on the data provided on the boundary marked in red. While for elliptic problems we need to specify conditions on the entire boundary $\partial\Omega$, for a parabolic problem we only need to know the 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 iterative 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 $\overline{\Omega}_h = \{0, h, \ldots, (N+1)h = L\}$ with grid spacing $h = L/(N+1)$. We 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 with omit an extra index h or τ). As before we will discretize the second spatial derivatives using

$$D_x^+ D_x^- u_n^k = \frac{1}{h^2}(u_{n-1}^k - 2u_n^k + u_{n+1}^k) \tag{II.50}$$

and first time-derivatives using

$$D_t^+ u_n^k = \frac{1}{\tau}(u_n^{k+1} - u_n^k). \tag{II.51}$$

Definition II.36 (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 $k = 1 \ldots M$ and $n = 1 \ldots N$ 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 with 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** $\theta = 0$:

$$u_n^{k+1} = (1-2\gamma)u_n^k + \gamma(u_{n+1}^k + u_{n-1}^k) + \tau f(t^k, x_n) \tag{II.52}$$

ii) **Implicit/Backward Euler scheme** $\theta = 1$:

$$(1+2\gamma)u_n^{k+1} - \gamma(u_{n+1}^{k+1} + u_{n-1}^{k+1}) = u_n^k + +\tau f(t^{k+1}, x_n) \tag{II.53}$$

iii) **Crank-Nicolson scheme** $\theta = 1/2$:

$$(1+\gamma)u_n^{k+1} - \tfrac{\gamma}{2}(u_{n+1}^{k+1} + u_{n-1}^{k+1}) = (1-\gamma)u_n^k + \tfrac{\gamma}{2}(u_{n+1}^k + u_{n-1}^k) + \tau \bar{f}_n^k \tag{II.54}$$

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

Theorem II.37 (Consistency of θ-scheme)**:** The θ-scheme defined in II.36 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. Proven as usual using Taylor's theorem. □

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

$$\begin{aligned} &-\gamma\theta(u_{n-1}^{k+1} + u_{n+1}^{k+1}) + (2\theta\gamma + 1)u_n^{k+1} = F_n^k, \\ &\text{where} \quad F_n^k = (1-\theta)\gamma(u_{n-1}^k + u_{n+1}^k) + (1 - 2(1-\theta)\gamma)u_n^k + \tau\bar{f}_n^k. \end{aligned}$$

The diagonal dominance of the first equation implies

$$\max_n |u_n^{k+1}| \le \max_n |F_n^k|.$$

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

$$\max_{k,n} |u_n^{k+1}| = \max_n |u_0(x_n)| + \tau \sum_{k=1}^{M} \max_n |\bar{f}_n^k|. \tag{II.55}$$

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 restruction $(1-2r) \ge 0$ is translated into a restriction of the time-step size

$$(1-\theta)\tau \le \tfrac{1}{2}h^2, \tag{II.56}$$

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 $(1-\theta)\tau \leq \frac{1}{2}h^2$ and $u \in C^{2,4}(\bar{Q}_T)$ and $\bar{f}_n^k = f(t^k, x_n)$. Then we have

$$\max_{k,n} |u_n^k - u(t^k, x_n)| = \mathcal{O}(h^2 + \tau). \tag{II.57}$$

For the Crank-Nicolson scheme with $\tau \leq h^2$ we have

$$\max_{k,n} |u_n^k - u(t^k, x_n)| = \mathcal{O}(h^2 + \tau^2). \tag{II.58}$$

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

$$\max_{k,n} |w_n^k| = \underbrace{\max_n |w_n^0|}_{0} + \tau \sum_{k=1}^{M} \max_n |\varepsilon| \leq T|\varepsilon|.$$

□

II.8.4. Generalization of θ-scheme

Lets consider the θ-scheme with a general discrete elliptic operator L_h. We have

$$[\mathbb{I} + \theta\tau L_h]u_n^{k+1} = F_n^k = \bar{f}_n^k + [\mathbb{I} - \tau(1-\theta)L_h]u_n^k,$$

Theorem II.39: Let $M \in \mathbb{R}^{n\times n}$ weakly diagonal dominant. Then

$$\|v\|_\infty \leq \|(\mathbb{I} + M)v\|_\infty. \tag{II.59}$$

Furthermore assume $0 \leq M_{ii} \leq 1$. Then

$$\|(1-M)v\|_\infty \leq \|v\|_\infty. \tag{II.60}$$

Corollary II.40: If L_h is weakly diagonal domiant, then the discrete parabolic IVP is stable in the max-norm $\|\|_\infty$.

Proof. Use first estimate with $M = \theta\tau L_h$ and then the second estimate using the CFL-type condition $0 \leq \tau(1-\theta)(L_h)_{ii} \leq 1$ analogously gives stability in max-norm. □

Remark II.41: 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.42 (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.43 (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.44 (Stability of Implicit Euler scheme)**:** The implicit Euler scheme reads

$$(\mathbb{I} + \tau L_h)u^{k+1} = \tau f^k + u^k \tag{II.61}$$

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

$$\|u^{k+1}\| \leq \left\|\left|(\mathbb{I} + \tau L_h)^{-1}\right|\right\| (\tau\|\bar{f}^k\| + \|u^k\|) \leq \tau\|\bar{f}^k\| + \|u^k\| \tag{II.62}$$

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

$$u_h(t^k, x_n) = u_n^k = \sum_m \omega_m^k v_m(x_n), \tag{II.63}$$

which plugging into the discretized system and using the fact that v_m is an eigenfunction gives

$$\frac{\omega_m^{k+1} - \omega_m^k}{\tau} = -\lambda_m(\theta\omega_m^{k+1} + (1-\theta)\omega_n^k), \tag{II.64}$$

for all eigenfunctions m. The initial coefficients can be obtained via $\omega_m^0 = (u_0, v_m)_h$. We can write (II.64) in the form

$$\omega_m^{k+1} = q(\tau\lambda_m)\omega_m^k, \qquad \text{where} \qquad q(s) = \frac{1-(1-\theta)s}{1+\theta s} \tag{II.65}$$

with the special cases

$$q(s) = 1 - s, \qquad \text{explicit Euler } \theta = 0, \tag{II.66}$$

$$q(s) = \frac{1}{1+s}, \qquad \text{implicit Euler } \theta = 1, \tag{II.67}$$

$$q(s) = \frac{1 - \frac{1}{2}s}{1 + \frac{1}{2}s}, \qquad \text{Crank-Nicolson } \theta = \tfrac{1}{2}. \tag{II.68}$$

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.45: 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 an unrealistic regularity of the exact solution (e. g. $\|u\|_{C^4(\overline{\Omega})} < \infty$) which is often not given even in simple cases. Consider the problem

$$\begin{cases} -\Delta u = 0 & \text{in } \Omega = (0,1) \times (0,1), \\ u(x,y) = x^2 & \text{on } \partial\Omega. \end{cases}$$

This PDE admits a unique solution, but since $\Delta u(x,y) = 2$ on the boundary, it follows that it is not a classical solution $u \notin C^2(\overline{\Omega})$.

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 is often 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.

Bibliography

[Eva98] L. C. Evans. *Partial Differential Equations*, volume 19 of *Grad. Stud. Math.* AMS, Providence, RI, USA, 1998.

[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.

[Wac13] E. Wachspress. *The ADI Model Problem*. Springer-Verlag, New York, NY, USA, 1st edition, 2013.

[Zwi98] D. Zwillinger. *Handbook of differential equations*, volume 1. Gulf Professional Publishing, 1998.