# Numerical Analysis and Simulation of Partial Differential Equations

Roland Pulch

Lecture in Summer Term 2011

University of Wuppertal
Applied Mathematics/Numerical Analysis

Contents:

Literature:

Ch. Großmann, H.-G. Roos, M. Stynes: Numerical Treatment of Partial Differential Equations. Springer, Berlin 2007. (parts of Chapters 1-4)

H.R. Schwarz, J. Waldvogel: Numerical Analysis: A Comprehensive Introduction. John Wiley & Sons, 1989. (Chapter 10)

D. Braess: Finite Elements. (3rd ed.) Cambridge University Press, 2007.

R.J. LeVeque: Numerical Methods for Conservation Laws. Birkhäuser, Basel, 1992.

# Contents

Chapter 1

# Examples and Classification

This lecture deals with the numerical solution of *partial differential equations* (PDEs). The exact solution depends on several independent variables, which are often the time and the space coordinates. Different types of PDE models exist even in the linear case. Each class exhibits certain properties and thus requires corresponding numerical methods. Initial and/or boundary conditions appear.

In contrast, systems of *ordinary differential equations* (ODEs) can be written in the general form

$$y'(x) = f(x, y(x)) \qquad (y : \mathbb{R} \to \mathbb{R}^n,\ f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n).$$

The independent variable $x$ often represents the time. Thus initial value problems $y(x_0) = y_0$ are the most important task. An analytical solution is not feasible in general. Hence we need numerical methods to achieve an approximate solution. Nevertheless, a convergent numerical method can resolve an arbitrary system of ODEs.

## 1.1 Examples

We present three important examples, which illustrate the three classes of PDE models.

### Poisson equation

We consider an open bounded domain $\Omega \subset \mathbb{R}^2$. For example, we can choose $\Omega = (0,1) \times (0,1)$. For $u \in C^2(\Omega)$, the Laplace operator is defined via

$$\Delta u := \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{1.1}$$

The Poisson equation (in two space dimensions $x, y$) reads

$$-\Delta u = f$$

with a predetermined function $f : \Omega \to \mathbb{R}$. The special case $f \equiv 0$ reproduces the Laplace equation. Hence the solution of the Poisson equation (1.1) is stationary, i.e., it does not change in time.

Now we specify boundary value problems. Let $\partial\Omega$ be the boundary of $\Omega$. Boundary conditions of Dirichlet type read

$$u(x,y) = g(x,y) \quad \text{for } (x,y) \in \partial\Omega$$

with a given function $g : \partial\Omega \to \mathbb{R}$. Boundary conditions of Neumann type specify the derivative of the solution perpendicular to the boundary, i.e.,

$$\frac{\partial u}{\partial \nu}(x,y) := \langle \nu(x,y), \nabla u(x,y) \rangle = h(x,y) \quad \text{for } (x,y) \in \partial\Omega$$

with the normal vector $\nu$ ($\|\nu\|_2 = 1$) and a given function $h : \partial\Omega \to \mathbb{R}$. Often mixed boundary conditions

$$u(x,y) = g(x,y) \text{ for } (x,y) \in \Gamma_D, \quad \frac{\partial u}{\partial \nu}(x,y) = h(x,y) \text{ for } (x,y) \in \Gamma_N$$

appear with $\Gamma_D \cup \Gamma_N = \partial\Omega$ and $\Gamma_D \cap \Gamma_N = \emptyset$.

We derive the Poisson equation for electric fields in case of three space dimensions ($x = (x_1, x_2, x_3)$). Let $E : \mathbb{R}^3 \to \mathbb{R}^3$ be the electric field and $\Phi : \mathbb{R}^3 \to \mathbb{R}$ the corresponding potential. It holds

$$E(x) = -\nabla\Phi(x)$$

with the gradient $\nabla\Phi = (\frac{\partial\Phi}{\partial x_1}, \frac{\partial\Phi}{\partial x_2}, \frac{\partial\Phi}{\partial x_3})$. It follows

$$\mathrm{div}E(x) = -\Delta\Phi(x).$$

Let $\rho : \mathbb{R}^3 \to \mathbb{R}$ describe the charge distribution and $\varepsilon > 0$ be the permittivity. The first Maxwell's equation (Gauss' law) reads

$$\mathrm{div}E(x) = \frac{\rho(x)}{\varepsilon}.$$

Comparing the two relations, we obtain the Poisson equation

$$-\Delta\Phi(x) = \frac{\rho(x)}{\varepsilon},$$

where $\rho$ is given and $\Phi$ is unknown.

A connection to complex analysis is given for holomorphic functions. Let $g : \mathbb{C} \to \mathbb{C}$ be holomorphic and $\Omega \subset \mathbb{C}$ be bounded, connected and open. On the one hand, Cauchy's integral formula yields

$$g(z) = \frac{1}{2\pi\mathrm{i}} \oint_{\partial\Omega} \frac{g(\zeta)}{\zeta - z}\, \mathrm{d}\zeta \qquad \text{for } z \in \Omega.$$

It follows that $g$ is already determined uniquely inside $\Omega$ by its values on the boundary $\partial\Omega$. On the other hand, the formulas of complex differentiation (Cauchy-Riemann-PDEs) imply

$$\Delta(\mathrm{Re}\ g) = 0 \qquad \text{and} \qquad \Delta(\mathrm{Im}\ g) = 0,$$

i.e., real and imaginary part are solutions of the Laplace equation in two dimensions. The values of $g$ on $\partial\Omega$ specify Dirichlet boundary conditions. It follows a unique solution for the real and the imaginary part in $\Omega$, respectively. Thus the two theoretical concepts agree.

## Wave equation

In a single space dimension, the wave equation reads

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \tag{1.2}$$

where the real constant $c > 0$ is the wave speed. The solution $u$ depends on space as well as time. We solve the wave equation using d'Alembert's method. New variables are introduced via

$$\xi = x - ct, \qquad \eta = x + ct.$$

It follows

$$\begin{aligned}
\frac{\partial}{\partial x} &= \frac{\partial}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial}{\partial \eta}\frac{\partial \eta}{\partial x} = \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}, \\
\frac{\partial^2}{\partial x^2} &= \frac{\partial^2}{\partial \xi^2} + 2\frac{\partial^2}{\partial \xi \partial \eta} + \frac{\partial^2}{\partial \eta^2}, \\
\frac{\partial}{\partial t} &= \frac{\partial}{\partial \xi}\frac{\partial \xi}{\partial t} + \frac{\partial}{\partial \eta}\frac{\partial \eta}{\partial t} = c\left(-\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}\right), \\
\frac{\partial^2}{\partial t^2} &= c^2\left(\frac{\partial^2}{\partial \xi^2} - 2\frac{\partial^2}{\partial \xi \partial \eta} + \frac{\partial^2}{\partial \eta^2}\right).
\end{aligned}$$

We obtain the transformed PDE

$$c^2\left(\frac{\partial^2}{\partial \xi^2} - 2\frac{\partial^2}{\partial \xi \partial \eta} + \frac{\partial^2}{\partial \eta^2}\right) u(\xi, \eta) = c^2\left(\frac{\partial^2}{\partial \xi^2} + 2\frac{\partial^2}{\partial \xi \partial \eta} + \frac{\partial^2}{\partial \eta^2}\right) u(\xi, \eta)$$

and thus

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0.$$

It is straightforward to verify that the general solution is given by

$$u(\xi, \eta) = \Phi(\xi) + \Psi(\eta)$$

with arbitrary functions $\Phi, \Psi \in C^2(\mathbb{R})$. A special case is $\Phi \equiv \Psi$ (only for $\frac{\partial u}{\partial t}(x, 0) \equiv 0$). It follows

$$u(x, t) = \Phi(x - ct) + \Psi(x + ct).$$

The functions $\Phi, \Psi$ follow from initial conditions. As an interpretation, we rewrite

$$\Phi(x - ct) = \Phi(x + c\Delta t - c(t + \Delta t)) = \Phi(x^* - ct^*)$$

with $x^* := x + c\Delta t$ and $t^* := t + \Delta t$. In the period $\Delta t$, the information travels from the point $x$ to the point $x^*$ with $\Delta x = c\Delta t$. Hence the term

$\Phi(x-ct)$ represents a wave moving at speed $c$. Likewise, the term $\Psi(x+ct)$ yields a wave moving at speed $-c$. The solution $u$ is the superposition of two waves travelling at opposite speeds.

For initial values $u(x,0) = u_0(x)$, $\frac{\partial u}{\partial t}(x,0) = cu_1(x)$ at $t = 0$, a formula for the exact solution is available, i.e.,

$$u(x,t) = \frac{1}{2}\left(u_0(x+ct) + u_0(x-ct) + \int_{x-ct}^{x+ct} u_1(s)\,\mathrm{d}s\right). \tag{1.3}$$

It follows that the solution $u$ at a point $(x^*, t^*)$ for $t^* > 0$ depends on initial values at $t = 0$ in the interval $x \in [x^* - ct^*, x^* + ct^*]$ only. Hence the wave equation includes a transport of information at a finite speed.

The linear PDE (1.2) of second order can be transformed into a corresponding system of PDEs of first order. We define $v_1 := \frac{\partial u}{\partial t}$ and $v_2 := \frac{\partial u}{\partial x}$. Assuming $u \in C^2$, the theorem of Schwarz yields

$$\frac{\partial v_1}{\partial x} = \frac{\partial^2 u}{\partial x \partial t} = \frac{\partial v_2}{\partial t}.$$

The PDE (1.2) implies

$$\frac{\partial v_1}{\partial t} = c^2 \frac{\partial v_2}{\partial x}.$$

It follows the system

$$\frac{\partial}{\partial t}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} 0 & -c^2 \\ -1 & 0 \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{1.4}$$

The resulting matrix exhibits the eigenvalues $+c$ and $-c$, i.e., the wave speeds. To obtain the solution $u$ of (1.2), an integration using $v_1, v_2$ still has to be done.

### Heat equation

In a single space dimension, the heat equation reads

$$\frac{\partial u}{\partial t} = \lambda \frac{\partial^2 u}{\partial x^2} \tag{1.5}$$

with a constant $\lambda > 0$. To demonstrate some solutions of this PDE, we choose $\lambda = 1$ and consider the finite domain $x \in [0, \pi]$ without loss of generality. We arrange homogeneous boundary conditions

$$u(0,t) = 0, \quad u(\pi,t) = 0 \qquad \text{for all } \ t \geq 0. \tag{1.6}$$

The functions

$$v_k(x,t) := \mathrm{e}^{-k^2 t} \sin(kx) \qquad \text{for } \ k \in \mathbb{N} \tag{1.7}$$

satisfy the heat equation (1.5) and the boundary conditions (1.6). Given initial values $u(x,0) = u_0(x)$ for $x \in [0,\pi]$ with $u_0(0) = u_0(\pi) = 0$, we can apply a Fourier expansion

$$u_0(x) = \sum_{k=1}^{\infty} a_k \sin(kx)$$

with coefficients $a_k \in \mathbb{R}$. Since the heat equation (1.5) is linear, we obtain the solution as a superposition of the functions (1.7)

$$u(x,t) = \sum_{k=1}^{\infty} a_k \mathrm{e}^{-k^2 t} \sin(kx)$$

for $t \geq 0$.

Alternatively, boundary conditions of Neumann type can be specified. Homogeneous conditions

$$\frac{\partial u}{\partial x}(0,t) = 0, \quad \frac{\partial u}{\partial x}(\pi,t) = 0 \qquad \text{for all } \ t \geq 0$$

imply that there is no heat flux through the boundaries.

Given initial conditions $u(x,0) = u_0(x)$ in the whole space domain, it follows a formula for the exact solution of the heat equation ($\lambda = 1$)

$$u(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\xi^2/4t} u_0(x-\xi) \, \mathrm{d}\xi \tag{1.8}$$

provided that the integral exists. We recognise that the solution in some point $(x,t)$ depends on the initial values $u_0(\xi)$ for all $\xi \in \mathbb{R}$. Hence the

transport of information proceeds at infinite speed. However, the magnitude of the information is damped exponentially for increasing distances.

We derive the heat equation in three space dimensions. Let $T : \mathbb{R}^3 \to \mathbb{R}$ be the temperature, $F : \mathbb{R}^3 \to \mathbb{R}^3$ the heat flux and $\kappa > 0$ the diffusion constant. It follows

$$F = -\kappa \nabla T.$$

For the energy $E : \mathbb{R}^3 \to \mathbb{R}$, we obtain

$$\frac{\partial E}{\partial t} = -\mathrm{div}\, F = \kappa \mathrm{div} \nabla T = \kappa \Delta T.$$

Let $\alpha := \frac{\partial E}{\partial T}$ be a constant material parameter. It holds $\frac{\partial E}{\partial t} = \frac{\partial E}{\partial T}\frac{\partial T}{\partial t}$. Consequently, the heat equation

$$\frac{\partial T}{\partial t} = \frac{\kappa}{\alpha}\Delta T$$

is achieved with $\lambda = \frac{\kappa}{\alpha}$.

**Black-Scholes equation**

The above examples are motivated by physics and technical applications. We discuss shortly an example from financial mathematics. Let $S$ be the price of a stock and $V$ be the fair price of a European call option based on this stock. It follows a PDE with solution $V$, where the independent variables are the time $t \geq 0$ and the value $S \geq 0$. The famous Black-Scholes equation reads

$$\frac{\partial V}{\partial t} + \tfrac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0, \tag{1.9}$$

with constants $r, \sigma > 0$. Although the Black-Scholes equation (1.9) looks complicated, it can be transformed into the heat equation (1.5) by transformations in the domain of dependence. Thus the properties of the Black-Scholes equation are the same as for the heat equation.

**Remark:** The wave equation (1.2), the heat equation (1.5) and the Black-Scholes equation (1.9) are relatively simple such that formulas for corresponding solutions exist, see (1.3) and (1.8). Hence numerical methods are

not required necessarily. However, an analytical solution is often not feasible any more if a source term appears, i.e.,

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} + f(x,t,u) \qquad \text{or} \qquad \frac{\partial u}{\partial t} = \lambda \frac{\partial^2 u}{\partial x^2} + f(x,t,u).$$

Now we need numerical schemes to obtain an approximative solution. Nevertheless, the fundamental properties of the PDEs do not change by adding a source term.

## 1.2 Classification

We consider a linear PDE of second order

$$\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} = f\left(x_1, \ldots, x_n, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}\right) \tag{1.10}$$

with $n \geq 2$ independent variables. Let the solution $u : \Omega \to \mathbb{R}$ satisfy $u \in C^2(\Omega)$ for some open domain $\Omega \subseteq \mathbb{R}^n$. We apply the abbreviations $x = (x_1, \ldots, x_n)$ and $\nabla u = (\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n})$. The types of PDEs with respect to the degree of linearity read:

- *linear PDE*: The coefficients $a_{ij}$ are constants or depend on $x$ only and the right-hand side is linear ($f = b(x) + c(x)u + d_1(x)\frac{\partial u}{\partial x_1} + \cdots + d_n(x)\frac{\partial u}{\partial x_n}$).
- *semi-linear PDE*: The coefficients $a_{ij}$ are constants or depend on $x$ only and the right-hand side $f$ is nonlinear.
- *quasi-linear PDE*: The coefficients $a_{ij}$ depend on $u$ and/or $\nabla u$. (The right-hand side $f$ can be linear or nonlinear.)

The definition of well-posed problems is as follows.

**Definition 1** *A PDE or system of PDEs with corresponding initial and/or boundary conditions is well-posed if and only if a unique solution exists and the solution depends continuously on the input data. Otherwise, the problem is called ill-posed.*

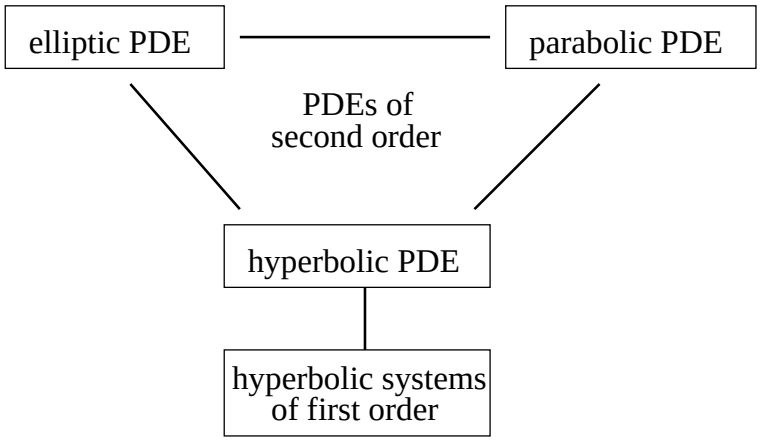



Figure 1: Classification of PDEs.

The coefficients $a_{ij}$ form a matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$. Without loss of generality, we assume that the matrix $A$ is symmetric due to $u \in C^2$. Thus $A$ is diagonalisable and all eigenvalues (EVs) $\lambda_1, \ldots, \lambda_n$ are real. The classification of PDEs (1.10) is based on the definiteness of $A$. (The classification is independent of the right-hand side $f$.) In case of two dimensions ($n = 2$), it holds $\det(A) = \lambda_1\lambda_2$, i.e., the definiteness follows from the sign of the determinant. The quadratic form

$$q(z) := z^\top Az \qquad (A \in \mathbb{R}^{2\times 2}, z \in \mathbb{R}^2)$$

can be investigated for a geometrical interpretation of the definiteness. This yields the nomenclature of the types of PDEs. Fig. 1 illustrates the corresponding classes.

**Case 1:** $A$ (pos. or neg.) definite $\rightarrow$ *elliptic PDE*
(all EVs of $A$ are positive or all EVs of $A$ are negative)
Elliptic PDEs describe stationary solutions. Boundary conditions yield well-posed problems.

For $n = 2$, the set $\{z \in \mathbb{R}^2 : z^\top A z = \pm 1\}$ represents an ellipse.

Example: *Poisson equation*
In $n$ dimensions, the Poisson equation reads

$$-\Delta u = -\frac{\partial^2 u}{\partial {x_1}^2} - \cdots - \frac{\partial^2 u}{\partial {x_n}^2} = f(x_1, \ldots, x_n). \tag{1.11}$$

It follows that the matrix $A \in \mathbb{R}^{n \times n}$ is diagonal an all diagonal elements (eigenvalues) are equal to $-1$. Consequently, the PDE (1.11) is elliptic.

**Case 2:** $A$ indefinite, $\det A \neq 0 \rightarrow$ *hyperbolic PDE*
(at least two EVs have opposite sign, all EVs are non-zero)
Hyperbolic PDEs model transport processes. Initial conditions (possibly additional boundary conditions) result in well-posed problems.

For $n = 2$, the set $\{z \in \mathbb{R}^2 : z^\top A z = 1\}$ represents a hyperbola.

Example: *Wave equation*
In $n$ space dimensions, the wave equation is given by

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = \frac{\partial^2 u}{\partial t^2} - c^2 \left( \frac{\partial^2 u}{\partial {x_1}^2} + \cdots + \frac{\partial^2 u}{\partial {x_n}^2} \right) = 0 \tag{1.12}$$

with wave speed $c > 0$. Again the coefficient matrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ is diagonal. It follows a simple EV $+1$ and a multiple EV $-c^2$. Hence the wave equation (1.12) is a hyperbolic PDE.

Often one EV exhibits a different sign than all other EVs (the time differs qualitatively from the space coordinates). If two pairs of eigenvalues have an opposite sign, then the PDE is called *ultrahyperbolic* ($n \geq 4$ necessary). However, we will not consider ultrahyperbolic PDEs in this lecture.

In the case $n = 2$, a hyperbolic PDE of second order can be transformed into an equivalent hyperbolic PDE of first order. An example has been given

in (1.4). Vice versa, not each hyperbolic PDE of first order is equivalent to a PDE of second order.

**Case 3:** $\det A = 0 \rightarrow$ *parabolic PDE*
(at least one EV is equal to zero)
Parabolic PDEs describe diffusion, for example. Initial conditions in addition to boundary conditions yield well-posed problems.

For $n = 2$, the set $\{z \in \mathbb{R}^2 : z^\top A z = 1\}$ corresponds to straight lines. If linear terms are added to the quadratic form ($q(x) := z^\top A z + b^\top z$), then a parabola can appear.

Example: *Heat equation*
In $n$ space dimensions, the heat equation reads

$$\frac{\partial u}{\partial t} - \lambda \Delta u = \frac{\partial u}{\partial t} - \lambda \left( \frac{\partial^2 u}{\partial {x_1}^2} + \cdots + \frac{\partial^2 u}{\partial {x_n}^2} \right) = 0 \tag{1.13}$$

including a constant $\lambda > 0$. The coefficient matrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ is diagonal. A simple EV zero and a multiple eigenvalue $-\lambda$ appears. It follows that the PDE (1.13) is parabolic.

The classification is unique in case of constant coefficients $a_{ij}$. For $a_{ij}(x)$, the same PDE (1.10) may exhibit different types in different domains $\Omega$. For $a_{ij}(u)$, the type of the PDE may even depend on the corresponding solution $u$. However, this happens rather seldom in practice.

### Scaling

Multiplying the PDE (1.10) by a coefficient $\alpha \neq 0$ changes the matrix $A$ into $\alpha A$. The differences in the signs of the eigenvalues remain the same. Thus the type of the PDE is invariant with respect to this scaling.

### Basis transformations

Now we investigate the invariance of the type of a PDE (1.10) with respect to basis transformations in the domain of dependence $\Omega$. We consider constant coefficients $a_{ij}$, since a generalisation to non-constant coefficients is straightforward. Let $y = Bx$ using a regular matrix $B = (b_{ij}) \in \mathbb{R}^{n\times n}$. In the new coordinates $y$, the solution is $\tilde{u}(y) = \tilde{u}(Bx) = u(x)$. The chain rule of multivariate differentiation implies

$$\begin{aligned}\frac{\partial \tilde{u}(Bx)}{\partial x_i} &= \sum_{k=1}^{n} \frac{\partial \tilde{u}(Bx)}{\partial y_k} b_{ki},\\ \frac{\partial^2 \tilde{u}(Bx)}{\partial x_j \partial x_i} &= \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{\partial^2 \tilde{u}(Bx)}{\partial y_l \partial y_k} b_{ki} b_{lj}.\end{aligned}$$

It follows

$$\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u(x)}{\partial x_j \partial x_i} = \sum_{i,j=1}^{n} a_{ij} \sum_{k,l=1}^{n} \frac{\partial^2 \tilde{u}(Bx)}{\partial y_l \partial y_k} b_{ki} b_{lj} = \sum_{k,l=1}^{n} \left[ \sum_{i,j=1}^{n} a_{ij} b_{ki} b_{lj} \right] \frac{\partial^2 \tilde{u}(y)}{\partial y_l \partial y_k}.$$

Let $\tilde{A} = (\tilde{a}_{kl})$ be the coefficients in the new basis. It holds

$$\tilde{A} = BAB^\top. \tag{1.14}$$

The matrix $\tilde{A}$ is always symmetric. However, the eigenvalues of $A$ are invariant just for orthogonal matrices $B$, i.e., $B^{-1} = B^\top$. Hence orthogonal transformations do not change the type of the PDE. Each symmetric matrix $A$ can be diagonalised via $D = SAS^\top$ using an orthogonal matrix $S$. Thus each PDE (1.10) can be transformed into an equivalent equation with diagonal coefficient matrix of the same type.

Non-orthogonal basis transformations may change the type of the PDE in case of $n \geq 3$. Nevertheless, the type is invariant for an arbitrary basis transformation in case of $n = 2$. Thereby, the type depends just on the sign of $\det(A)$, i.e., elliptic for $\det(A) > 0$, hyperbolic for $\det(A) < 0$ and parabolic for $\det(A) = 0$. The transformation (1.14) yields

$$\det(\tilde{A}) = \det(B) \cdot \det(A) \cdot \det(B^\top) = (\det(B))^2 \det(A).$$

Hence the sign of $\det(\tilde{A})$ is identical to the sign of $\det(A)$.

# Chapter 2

# Elliptic PDEs

In this chapter, we discuss the numerical solution of boundary value problems of second-order elliptic PDEs. Thereby, the Poisson equation represents a benchmark problem. Two classes of numerical methods exist: finite difference methods and finite element methods.

## 2.1 Maximum principle

We consider the Poisson equation

$$-\Delta u(x) := -\sum_{i=1}^{n} \frac{\partial^2 u(x)}{\partial x_i^2} = f(x_1, \ldots, x_n) \tag{2.1}$$

in $n \geq 2$ space dimensions. Let $\Omega \subset \mathbb{R}^n$ be an open and bounded domain. Boundary conditions of Dirichlet type read

$$u(x) = g(x) \qquad \text{for } \; x \in \partial\Omega \tag{2.2}$$

with a predetermined function $g : \partial\Omega \to \mathbb{R}$. We assume the existence of a solution $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$. (Corresponding theorems on existence require some assumptions and are hard to prove.) The uniqueness as well as the continuous dependence on the input data follows from the maximum principle.

**Theorem 1 (maximum principle)** *Let $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$. It holds:*

*(i)* *maximum principle: If $-\Delta u = f \leq 0$ in $\Omega$, then $u$ exhibits its maximum on the boundary $\partial\Omega$.*

*(ii)* *minimum principle: If $-\Delta u = f \geq 0$ in $\Omega$, then $u$ exhibits its minimum on the boundary $\partial\Omega$.*

*(iii)* *comparison: If $v \in C^2(\Omega) \cap C^0(\bar{\Omega})$ and $-\Delta u \leq -\Delta v$ in $\Omega$ and $u \leq v$ on $\partial\Omega$, then it follows $u \leq v$ in $\Omega$.*

Proof:

We show the property (i) first. We assume $f < 0$ in $\Omega$. If $\xi \in \Omega$ exists with

$$u(\xi) = \sup_{x \in \Omega} u(x) > \sup_{x \in \partial\Omega} u(x),$$

then $\xi$ is also a local maximum. It follows $\nabla u(\xi) = 0$ and the Hesse matrix $\nabla^2 u(\xi) = (u_{x_i,x_j}(\xi))$ is negative semi-definite. In particular, the entries on the diagonal are not positive. Thus it holds

$$-(u_{x_1,x_1}(\xi) + \cdots + u_{x_n,x_n}(\xi)) \geq 0.$$

This is a contradiction to $-\Delta u = f < 0$. Hence the maximum must be on the boundary $\partial\Omega$.

Now let $f \leq 0$ and $\eta \in \Omega$ with

$$u(\eta) = \sup_{x \in \Omega} u(x) > \sup_{x \in \partial\Omega} u(x).$$

We define $h(x) := (\eta_1 - x_1)^2 + \cdots + (\eta_n - x_n)^2$ and $w(x) := u(x) + \delta \cdot h(x)$ using a real number $\delta > 0$. Since $h \in C^2(\Omega) \cap C^0(\bar{\Omega})$ holds, the function $w$ exhibits its maximum inside $\Omega$ for sufficiently small $\delta$. It follows

$$-\Delta w(x) = -\Delta u(x) - \delta \Delta h(x) = f(x) - 2\delta n < 0.$$

Again a contradiction appears. Hence $\eta$ must be situated on the boundary $\partial\Omega$.

The property (ii) follows from (i) applying the maximum principle for the function $v := -u$.

To verify the property (iii), we define $w := v - u$. It follows

$$-\Delta w = -\Delta v + \Delta u \geq 0$$

due to the assumptions. It holds $w \geq 0$ on the boundary $\partial\Omega$. The minimum principle implies $w(x) \geq 0$ for all $x \in \Omega$. □

Using this maximum principle, we achieve the following estimate.

**Theorem 2** *For $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$, it holds*

$$|u(x)| \leq \sup_{z\in\partial\Omega} |u(z)| + c \sup_{z\in\Omega} |\Delta u(z)|. \tag{2.3}$$

*for each $x \in \Omega$ with some constant $c \geq 0$.*

Proof:

The bounded domain $\Omega$ is situated inside a circle of radius $R$ with its center at $x = 0$. We define

$$w(x) := R^2 - \sum_{i=1}^{n} x_i^2.$$

It follows $w_{x_i,x_j} = -2\delta_{ij}$. It holds $-\Delta w = 2n$ and $0 \leq w \leq R^2$ in $\Omega$. Now we arrange

$$v(x) := \sup_{z\in\partial\Omega} |u(z)| + w(x) \cdot \tfrac{1}{2n} \sup_{z\in\Omega} |\Delta u(z)| \geq 0.$$

Due to this construction, we have $-\Delta v \geq |\Delta u|$ in $\Omega$ and $v \geq |u|$ on $\partial\Omega$. Theorem 1 (iii) implies $-v(x) \leq u(x) \leq +v(x)$ in $\Omega$. Since $w \leq R^2$ holds, it follows (2.3) with $c := \frac{R^2}{2n}$. □

Let $u_1$ and $u_2$ be two solutions of the boundary value problem (2.1),(2.2), i.e., it holds $-\Delta u_1 = f_1$, $-\Delta u_2 = f_2$ in $\Omega$ and $u_1 = g_1$, $u_2 = g_2$ on $\partial\Omega$. Inserting the difference $u_1 - u_2$ into (2.3) yields

$$|u_1(x) - u_2(x)| \leq \sup_{z\in\partial\Omega} |g_1(z) - g_2(z)| + c \sup_{z\in\Omega} |f_1(z) - f_2(z)| \tag{2.4}$$

for all $x \in \Omega$. Hence the solution is Lipschitz-continuous with respect to the input data. Moreover, the solution of a boudary value problem of Dirichlet type (2.1),(2.2) is unique (choose $f_1 \equiv f_2$, $g_1 \equiv g_2$).

Boundary value problems of Dirichlet type are well-posed, see Definition 1, due to (2.4) (just existence is not shown but assumed). We give an example that initial value problems are ill-posed. We consider the Laplace equation $\Delta u = 0$ in the domain $\Omega = \{(x, y) \in \mathbb{R}^2 : y \geq 0\}$. Let initial values be specified at $y = 0$

$$u(x, 0) = \tfrac{1}{n} \sin(nx), \quad \tfrac{\partial u}{\partial y}(x, 0) = 0.$$

It follows the unique solution

$$u(x, y) = \tfrac{1}{n} \cosh(ny) \sin(nx),$$

which grows like $\mathrm{e}^{ny}$. It holds $|u(x, 0)| \leq \frac{1}{n}$, whereas $u$ becomes larger and larger at $y = 1$ for $n \to \infty$. Considering the limit case $u(x, 0) = 0$, the solution does not depend continuously on the initial data.

Now we consider a general differential operator of elliptic type.

**Definition 2** *The linear differential operator* $L : C^2(\Omega) \to C^0(\Omega)$

$$L := -\sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2}{\partial x_i \partial x_j} \tag{2.5}$$

*is called elliptic, if the matrix* $A = (a_{ij})$ *is positive definite for each* $x \in \Omega$, *i.e., it holds* $\xi^\top A(x)\xi > 0$ *for all* $\xi \in \mathbb{R}^n$. *The operator (2.5) is called uniformly elliptic in* $\Omega \subset \mathbb{R}^n$, *if a constant* $\alpha > 0$ *exists such that*

$$\xi^\top A(x)\xi \geq \alpha \|\xi\|_2^2 \quad \textit{for all } \xi \in \mathbb{R}^n \textit{ and all } x \in \Omega. \tag{2.6}$$

The maximum principle given in Theorem 1 also holds with $L$ instead of $-\Delta$ in case of a general elliptic operator (2.5). The estimate from Theorem 2 is valid with $L$ instead of $-\Delta$ for uniformly elliptic operators (2.5).

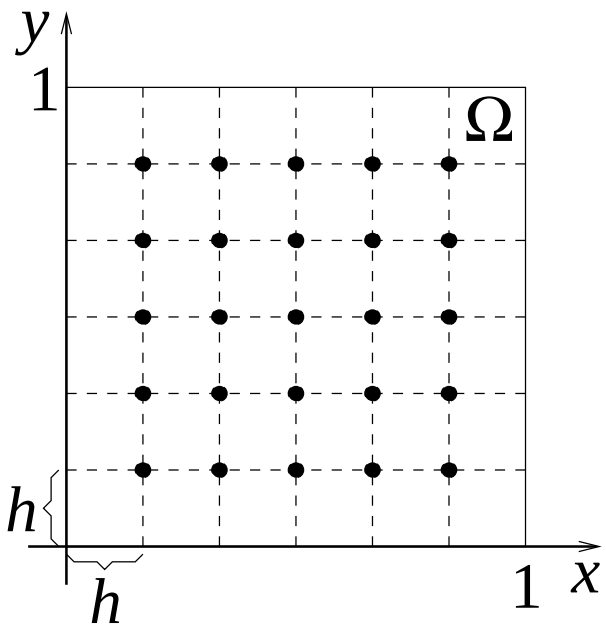



Figure 2: Grid in finite difference method.

## 2.2 Finite Difference Methods

We introduce the class of finite difference methods for boundary value problems of elliptic PDEs in two space dimensions.

### Laplace operator on unit square

As benchmark, we consider a boundary value problem of Dirichlet type on the unit square $\Omega := \{(x, y) : 0 < x, y < 1\}$ for the Poisson equation

$$\begin{aligned} -\Delta u = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} &= f(x,y) \qquad (x,y) \in \Omega \\ u(x,y) &= 0, \qquad (x,y) \in \partial\Omega. \end{aligned} \tag{2.7}$$

We introduce a uniform grid in the domain of dependence $\Omega$ using a step size $h := \frac{1}{M+1}$ for some $M \in \mathbb{N}$

$$\Omega_h := \{(x_i, y_j) = (ih, jh) : i, j = 1, \ldots, M\}, \tag{2.8}$$

see Figure 2. We construct a difference formula via Taylor expansion. It

holds for $u \in C^4(\Omega)$

$$\begin{aligned} u(x+h,y) &= u(x,y) + hu_x(x,y) + h^2\tfrac{1}{2}u_{xx}(x,y) + h^3\tfrac{1}{6}u_{xxx}(x,y) \\ &\quad + h^4\tfrac{1}{24}u_{xxxx}(x+\vartheta_1 h,y) \\ u(x-h,y) &= u(x,y) - hu_x(x,y) + h^2\tfrac{1}{2}u_{xx}(x,y) - h^3\tfrac{1}{6}u_{xxx}(x,y) \\ &\quad + h^4\tfrac{1}{24}u_{xxxx}(x-\vartheta_2 h,y) \end{aligned}$$

with $0 < \vartheta_1, \vartheta_2 < 1$. It follows

$$u(x+h,y) + u(x-h,y) = 2u(x,y) + h^2 u_{xx}(x,y) + h^4\tfrac{1}{12}u_{xxxx}(x+\vartheta h,y)$$

with $-1 < \vartheta < 1$ and thus

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2}(x,y) &= \frac{u(x+h,y) - 2u(x,y) + u(x-h,y)}{h^2} + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\vartheta h,y) \\ \frac{\partial^2 u}{\partial y^2}(x,y) &= \frac{u(x,y+h) - 2u(x,y) + u(x,y-h)}{h^2} + \frac{h^2}{12}\frac{\partial^4 u}{\partial y^4}(x,y+\eta h) \end{aligned}$$

with $-1 < \vartheta, \eta < 1$. These difference formulas are of order two. Now we replace the derivatives in (2.7) by these difference formulas using the grid points (2.8). Let $u_{i,j} := u(x_i, y_j)$, $f_{i,j} := f(x_i, y_j)$ and

$$(\Delta_h u)_{i,j} := \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2}$$

Omitting the remainder terms, it follows the linear system

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} \doteq h^2 f_{i,j} \tag{2.9}$$

for $i, j = 1, \ldots, M$. This discretisation can be illustrated in form of a five-point star, see Figure 3. The homogeneous boundary conditions imply

$$u_{0,j} = u_{M+1,j} = u_{i,0} = u_{i,M+1} = 0 \quad \text{for all } \; i, j.$$

We arrange the unknowns and the evaluations of the right-hand side $f$ in the form

$$\begin{aligned} U_h &= (u_{1,1}, u_{2,1}, \ldots, u_{M,1}, u_{1,2}, \ldots, u_{M,2}, \ldots, u_{1,M}, \ldots, u_{M,M})^\top \\ F_h &= (f_{1,1}, f_{2,1}, \ldots, f_{M,1}, f_{1,2}, \ldots, f_{M,2}, \ldots, f_{1,M}, \ldots, f_{M,M})^\top. \end{aligned} \tag{2.10}$$

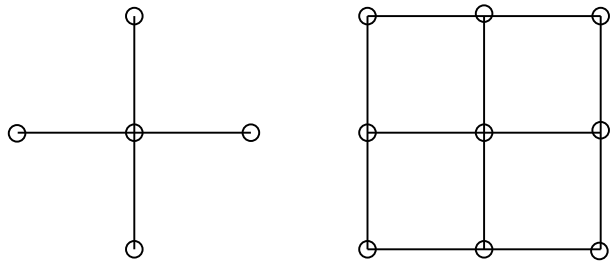


Figure 3: Five-point star (left) and nine-point star (right).

We obtain a linear system $A_h U_h = F_h$ of dimension $n = M^2$. The matrix $A_h$ is a band matrix

$$A_h = \frac{1}{h^2} \begin{pmatrix} C & -I & & \\ -I & C & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & C \end{pmatrix} \quad \text{with} \quad C = \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}. \quad (2.11)$$

The matrix $A_h$ is sparse, since each row includes at most five non-zero elements. Obviously, the matrix $A_h$ is symmetric. It can be shown that $A_h$ is always positive definite. Hence the matrix is regular. The corresponding solution $U_h = A_h^{-1} F_h$ represents an approximation of the solution $u$ of the PDE in the grid points.

### Laplace operator on general domain

Now we consider an arbitrary open and bounded domain $\Omega \subset \mathbb{R}^2$, see Figure 4. The application of a finite difference method requires the construction of a grid. We define an auxiliary (infinite) grid

$$G_h := \{(x, y) = (ih, jh) : i, j \in \mathbb{Z}\}.$$

Now the used (finite) grid reads

$$\Omega_h := G_h \cap \Omega.$$

Let $\Omega_h = \{z_1, \ldots, z_R\}$ be the grid points. Boundary conditions appear in the points of

$$\partial\Omega_h := (\{(ih, y) : i \in \mathbb{Z}, y \in \mathbb{R}\} \cup \{(x, jh) : j \in \mathbb{Z}, x \in \mathbb{R}\}) \cap \partial\Omega.$$

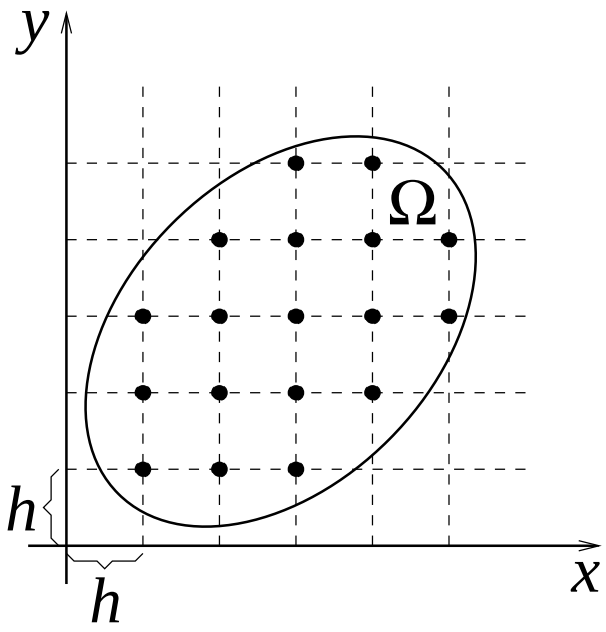



Figure 4: Grid in a general domain Ω.

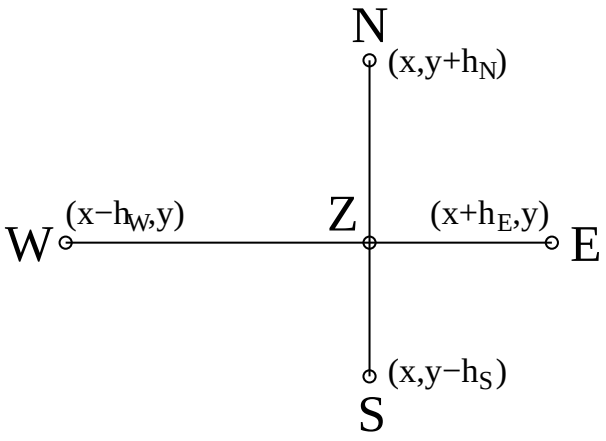



Figure 5: Five-point star for variable step sizes.

To arrange the difference formulas near the boundary, variable step sizes have to be considered.

We apply a five-point star again, see Figure 5. Taylor expansion yields

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_Z = \frac{2}{h_E(h_E+h_W)}u_E - \frac{2}{h_E h_W}u_Z + \frac{2}{h_W(h_E+h_W)}u_W + \mathcal{O}(h)$$

$$\left.\frac{\partial^2 u}{\partial y^2}\right|_Z = \frac{2}{h_N(h_N+h_S)}u_N - \frac{2}{h_N h_S}u_Z + \frac{2}{h_S(h_N+h_S)}u_S + \mathcal{O}(h)$$

provided that $u \in C^3(\Omega)$. Hence four (possibly) different step sizes are involved. Let $h := \max\{h_E, h_W, h_N, h_S\}$. This scheme to include the boundary data is also called the Shortley-Weller star.

In general, a five-point star and a nine-point star are specified via their coefficients

$$\begin{bmatrix} & \alpha_N & \\ \alpha_W & \alpha_Z & \alpha_E \\ & \alpha_S & \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NE} \\ \alpha_W & \alpha_Z & \alpha_E \\ \alpha_{SW} & \alpha_S & \alpha_{SE} \end{bmatrix},$$

respectively, cf. Figure 3. A discretisation of the Poisson equation (2.7) using arbitrary five point stars reads

$$\sum_{l=\mathrm{Z,E,S,W,N}} \alpha_l U_l = f_Z$$

for each $Z \in \Omega_h$. A general difference formula can be written in the form

$$L_h U := \sum_l \alpha_l U_l,$$

where the sum is over all non-zero coefficients $\alpha_l$. The discrete operator $L_h$ depends on the step sizes. We apply the notation

$$L_h u := \sum_l \alpha_l u(Z_l),$$

where a function $u : \bar{\Omega} \to \mathbb{R}$ (usually a solution of the PDE problem) is evaluated at the nodes $Z_l \in \Omega_h \cup \partial\Omega_h$.

We outline the algorithm of the finite difference method for the Dirichlet problem of the Poisson equation on a general domain $\Omega$.

**Algorithm:** *FDM for Dirichlet problem*

1. Choose a step size $h > 0$ and determine $\Omega_h$ as well as $\partial\Omega_h$.
2. Choose a numbering of the unknown $U_Z$ for $Z \in \Omega_h$.
3. Arrange the difference formulas

$$\alpha_{\mathrm{Z}} U_{\mathrm{Z}} + \alpha_{\mathrm{E}} U_{\mathrm{E}} + \alpha_{\mathrm{W}} U_{\mathrm{W}} + \alpha_{\mathrm{N}} U_{\mathrm{N}} + \alpha_{\mathrm{S}} U_{\mathrm{S}} = f_{\mathrm{Z}}$$

   for each $Z \in \Omega_h$.
4. If boundary values $U_B$ with $B \in \partial\Omega_h$ appear in the left-hand side of a difference formula, then replace $U_B$ by $g_B$ and shift this term to the right-hand side.
5. Arrange and solve the linear system

$$A_h U_h = F_h$$

   with the chosen numbering of the unknowns $U_h = (U_{Z_i})$ for $Z_i \in \Omega_h$.

### General differential operator

Difference schemes for mixed derivatives also exist. For example, it holds

$$\frac{\partial^2 u}{\partial x \partial y}(x_i, y_j) = \frac{u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}}{4h^2} + \mathcal{O}(h^2). \quad (2.12)$$

We verify this formula via (multivariate) Taylor expansions of the neighbouring points around the central point $u \equiv u_{i,j}$.

$$\begin{aligned}
u_{i+1,j+1} &= u + hu_x + hu_y + \tfrac{1}{2}h^2(u_{xx} + 2u_{xy} + u_{yy}) \\
&\quad + \tfrac{1}{6}h^3(u_{xxx} + 3u_{xxy} + 3u_{xyy} + u_{yyy}) + \mathcal{O}(h^4) \\
u_{i-1,j+1} &= u - hu_x + hu_y + \tfrac{1}{2}h^2(u_{xx} - 2u_{xy} + u_{yy}) \\
&\quad + \tfrac{1}{6}h^3(-u_{xxx} + 3u_{xxy} - 3u_{xyy} + u_{yyy}) + \mathcal{O}(h^4) \\
u_{i+1,j-1} &= u + hu_x - hu_y + \tfrac{1}{2}h^2(u_{xx} - 2u_{xy} + u_{yy}) \\
&\quad + \tfrac{1}{6}h^3(u_{xxx} - 3u_{xxy} + 3u_{xyy} - u_{yyy}) + \mathcal{O}(h^4) \\
u_{i-1,j-1} &= u - hu_x - hu_y + \tfrac{1}{2}h^2(u_{xx} + 2u_{xy} + u_{yy}) \\
&\quad + \tfrac{1}{6}h^3(-u_{xxx} - 3u_{xxy} - 3u_{xyy} - u_{yyy}) + \mathcal{O}(h^4)
\end{aligned}$$

$$\Rightarrow \quad u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1} = 4h^2 u_{xy} + \mathcal{O}(h^4),$$

Now we can discretise an arbitrary differential operator of second order

$$Lu := a\,\frac{\partial^2 u}{\partial x^2} + 2b\,\frac{\partial^2 u}{\partial x \partial y} + c\,\frac{\partial^2 u}{\partial y^2} \quad (2.13)$$

with $a, b, c \in \mathbb{R}$ for $n = 2$. The operator (2.13) is elliptic, if and only if $ac > b^2$ holds. The derivatives $\frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}$ are replaced by the difference formulas in $\Delta_h$ and the mixed derivative $\frac{\partial^2 u}{\partial x \partial y}$ is substituted by (2.13). It follows the (discrete) difference operator

$$\begin{aligned}
L_h u &= \tfrac{a}{h^2}[u_{i-1,j} - 2u_{i,j} + u_{i+1,j}] \\
&+ \tfrac{b}{2h^2}[u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}] \\
&+ \tfrac{c}{h^2}[u_{i,j-1} - 2u_{i,j} + u_{i,j+1}].
\end{aligned} \quad (2.14)$$

The difference formula represents a nine-point star, see Figure 3.

### Consistency, Stability and Convergence

We are interested in the convergence of a finite difference method, i.e., if the global error converges to zero for small step size. The consistency of the difference formula with respect to the underlying differential operator alone is not sufficient for the convergence. We require an additional stability property to guarantee the convergence. The conclusions are similar as in solving initial value problems of ODEs by multistep methods.

As for initial value problems of ordinary differential equations, we define a local error and a global error.

**Definition 3 (local and global error)** *Given $\Omega \subset \mathbb{R}^n$ and a (finite) grid $\Omega_h \subset \Omega$. Let $L$ be a differential operator and $L_h$ be a difference operator. For a sufficiently smooth function $u : \Omega \to \mathbb{R}$, the local error is given by $\tau(h) := Lu - L_h u$ on $\Omega_h$. If $u$ is a solution of the PDE $Lu = f$ and $U \in \mathbb{R}^{|\Omega_h|}$ a numerical solution on $\Omega_h$, then the global error is $\eta(z_i) := u(z_i) - U_i$ for each $z_i \in \Omega_h$.*

Now the definition of the convergence is based on the global error.

**Definition 4 (convergence)** *A numerical method using a difference operator $L_h$ is convergent, if the global error satisfies*

$$\lim_{h \to 0} \max_{z_i \in \Omega_h} |\eta(z_i)| = 0.$$

*The method is convergent of order $p$ (at least), if*

$$\max_{z_i \in \Omega_h} |\eta(z_i)| = \mathcal{O}(h^p).$$

In case of $h \to 0$, the number of grid points usually tends to infinity, i.e., $|\Omega_h| \to \infty$. To achieve a convergent method, the consistency of the difference scheme represents a crucial property.

**Definition 5 (consistency)** *A difference operator $L_h$ is called consistent with respect to a differential operator $L : V \to W$, if the local error satisfies*

$$\lim_{h \to 0} \tau(h) = 0 \qquad \text{uniformly on } \Omega_h$$

*for all functions $u \in V$. The method is consistent of order $p$ (at least), if*

$$\tau(h) = \mathcal{O}(h^p) \qquad \text{uniformly on } \Omega_h$$

*for all functions $u \in V$.*

For example, the difference operator (2.14) is consistent of order $p = 2$ with respect to the differential operator (2.13) for $V = C^4(\bar{\Omega})$.

We analyse the convergence of the Dirichlet problem of the Poisson equation, i.e., the discretisation $\Delta_h$ of the Laplace operator $\Delta$. It holds

$$A_h U_h = F_h, \qquad A_h \hat{U}_h = F_h + R_h,$$

where $\hat{U}_h = (u(z_i))$ represents the data of the exact solution in the grid points. Thus $R_h$ is a vector, which contains the local errors $\tau(h)$. Since the difference formula is consistent of order $p = 2$, it holds $\|R_h\|_\infty = \mathcal{O}(h^2)$. We apply the maximum norm, since the size of the vectors depends on $h$. Assuming that $A_h$ is regular for all $h > 0$, we obtain

$$U_h - \hat{U}_h = A_h^{-1} F_h - A_h^{-1}(F_h + R_h) = -A_h^{-1} R_h$$

and thus

$$\|U_h - \hat{U}_h\|_\infty \leq \left\|A_h^{-1}\right\|_\infty \cdot \|R_h\|_\infty \leq C \left\|A_h^{-1}\right\|_\infty h^2$$

with some constant $C > 0$ for sufficiently small $h$. However, to guarantee the convergence, we require a condition like

$$\left\|A_h^{-1}\right\|_\infty \leq K \qquad \text{or} \qquad \left\|A_h^{-1}\right\|_\infty \leq \frac{K}{h}$$

for all $h < h_0$ uniformly with some constant $K > 0$. Such a condition corresponds to the stability of the finite difference method.

We obtain a more general stability criterion by the following theoretical result. Thereby, we have to assume that the grid $\Omega_h$ is connected.

**Definition 6 (connected grid)** *A grid $\Omega_h \subset \Omega \subset \mathbb{R}^2$ is connected if each two points of the grid can be connected by piecewise straight lines, which remain inside the lines corresponding to the grid as well as inside $\Omega$.*

Typically, a connected grid is guaranteed for sufficiently small step size. Now we can formulate a discrete analogon of Theorem 1.

**Theorem 3 (discrete maximum principle)**
*Let the elliptic PDE $Lu = f$ with $f \leq 0$ in $\Omega$ and Dirichlet boundary values be given. Let $L_h$ be a five-point star difference operator on a connected grid $\Omega_h$ with negative coefficients outside the center and the sum of all coefficients is zero. Let the data $\{U_Z : Z \in \Omega_h \cup \partial\Omega_h\}$ satisfy the finite difference scheme. Then either all values $U_Z$ are constant or the maximum of the values $U_Z$ is not situated on $\Omega_h$ but on the boundary $\partial\Omega_h$.*

Proof:

We assume that the condition

$$\max_{Z \in \Omega_h} U_Z \geq \max_{B \in \partial\Omega_h} U_B$$

holds and show that then $U_Z$ is constant on $\Omega_h \cup \partial\Omega_h$, which also means that the discrete maximum appears on the boundary. Let $U_Z$ be a maximum within $\Omega_h$. Hence the neighbours satisfy the relation

$$U_Z \geq \max_{l=\mathrm{E,W,N,S}} U_l.$$

Furthermore, the difference formula implies

$$\sum_{l=\mathrm{Z,E,S,W,N}} \alpha_l U_l = f_Z \leq 0.$$

It follows

$$\begin{aligned} \sum_{l=\mathrm{E,S,W,N}} \alpha_l (U_l - U_Z) &= \sum_{l=\mathrm{Z,E,S,W,N}} \alpha_l (U_l - U_Z) \\ = \left( \sum_{l=\mathrm{Z,E,S,W,N}} \alpha_l U_l \right) - \left( U_Z \sum_{l=\mathrm{Z,E,S,W,N}} \alpha_l \right) &= \sum_{l=\mathrm{Z,E,S,W,N}} \alpha_l U_l \leq 0. \end{aligned}$$

It holds $\alpha_l < 0$ and $U_l - U_Z \leq 0$ for all $l$ in the original sum. Thus all terms of the original sum are non-negative. It follows that each term must be equal to zero. Again $\alpha_l < 0$ yields

$$U_Z = U_E = U_W = U_N = U_S$$

i.e., the neighbours of $U_Z$ have the same value. We continue this conclusion proceeding from $U_Z$ to the boundary. Each neighbour of the original $U_Z$ fulfills the assumptions and thus their neighbours also have the same value. Hence $U_Z$ exhibits the same value for all $Z \in \Omega_h$ as well as $Z \in \partial\Omega_h$. Thereby, we assume that the grid $\Omega_h$ is connected, i.e., each two grid points of $\Omega_h \cup \partial\Omega_h$ are connected by difference formulas. □

**Remarks:**

- The discrete maximum principle also holds for operators $L_h$ based on a nine-point star with analogous conditions on the coefficients. Just near the boundary, five-point stars have to be applied.
- The discrete operator (2.14) is consistent of order two with respect to the differential operator (2.13). However this difference formula does not satisfy the crucial condition $\alpha_l < 0$ for all coefficients outside the center. Thus a further analysis of stability is necessary in this case.

Further conclusions from the discrete maximum principle in Theorem 3 are:

- **discrete minimum principle:** If $Lu = f$ with $f \geq 0$ holds, then the discrete solution is either constant or it exhibits the minimum not on $\Omega_h$ but on the boundary $\partial\Omega_h$.
- **discrete comparison:** If $L_h U_Z \leq L_h V_Z$ for all grid points $Z \in \Omega_h$ and $U_B \leq V_B$ for all $B \in \partial\Omega_h$, then it follows $U_Z \leq V_Z$ in all $Z \in \Omega_h$.

Now we can show that the linear system from our finite difference method has a unique solution.

**Theorem 4** *We consider an elliptic PDE $Lu = f$ on $\Omega$ with Dirichlet boundary conditions $u = g$ on $\partial\Omega$. Let $A_h U_h = F_h$ be the linear system from a finite difference operator satisfying the assumptions of the discrete maximum principle. It follows that the matrix $A_h$ is regular and thus has a unique solution.*

Proof:

The homogeneous linear system $A_h U_h = 0$ represents the discretisation of the PDE for $f \equiv 0$ and $g \equiv 0$. Let $U_h$ be a solution. The discrete maximum principle implies $U_Z \leq 0$ in each $Z \in \Omega_h$, whereas the discrete minimum principle yields $U_Z \geq 0$ in each $Z \in \Omega_h$. It follows $U_h = 0$. Hence the matrix is not singular. □

It remains to show that the finite difference method is convergent. In the following, we restrict to the Laplace operator $Lu = -\Delta u$. We apply the five-point star, which is consistent of order at least one and satisfies the discrete maximum principle.

**Lemma 1** *Let $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ be the solution of the Poisson equation $-\Delta u = f$ with Dirichlet boundary conditions $u = g$ on $\partial\Omega$. Let $L_h$ be the difference operator of the five-point star on a grid $\Omega_h$. Then the local error and global error of the finite difference method fulfill the estimate*

$$\max_{Z \in \Omega_h} |u(Z) - U_Z| \leq K \max_{Z \in \Omega_h} |\tau(Z)| \tag{2.15}$$

*with a constant $K > 0$, which is independent of the step size $h$.*

Proof:

We investigate the local and global errors

$$\tau(Z) = -\Delta u(Z) - L_h u(Z), \qquad \eta(Z) = u(Z) - U_Z \qquad \text{for } Z \in \Omega_h.$$

It follows due to the linearity of the operators

$$L_h \eta(Z) = L_h u(Z) - L_h U_Z = L_h u(Z) - f(Z) = L_h u(Z) + \Delta u(Z) = -\tau(Z).$$

The global error vanishes at the boundary $\partial\Omega_h$, since the solution is equal to the predetermined boundary values. We analyse the problem

$$L_h\eta = -\tau \quad \text{in } \Omega_h, \qquad \eta = 0 \quad \text{on } \partial\Omega_h.$$

The scaling

$$\tilde{\eta} := \frac{\eta}{\gamma}, \quad \tilde{\tau} := \frac{\tau}{\gamma} \qquad \text{with } \gamma := \max_{Z\in\Omega_h} |\tau(Z)|$$

yields the problem

$$L_h\tilde{\eta} = -\tilde{\tau} \qquad \text{with } -1 \le \tilde{\tau}(Z) \le 1 \qquad \text{for all } Z \in \Omega_h.$$

It still holds $\tilde{\eta} = 0$ on $\partial\Omega_h$. Let $\Omega \subset \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < R^2\}$. We define the auxiliary function

$$w(x,y) := \tfrac{1}{4}(R^2 - x^2 - y^2) \ge 0.$$

Since all third derivatives of $w$ are identical zero, the local errors of the five-point star vanish. It follows $L_h w = -\Delta w = 1$ in $\Omega_h$. Since the five-point star satisfies the discrete maximum principle, the discrete comparison implies

$$\tilde{\eta} \le w \le \tfrac{1}{4}R^2 \qquad \text{for all } Z \in \Omega_h.$$

Using $-w$ instead of $w$, the discrete comparison yields

$$-\tfrac{1}{4}R^2 \le -w \le \tilde{\eta} \qquad \text{for all } Z \in \Omega_h.$$

Hence it follows

$$\max_{Z\in\Omega_h} |\eta(Z)| \le \tfrac{1}{4}R^2 \max_{Z\in\Omega_h} |\tau(Z)|$$

and we can choose the constant $K := \frac{R^2}{4}$. □

In the proof of Lemma 1, both the consistency and the discrete maximum principle corresponding to the five-point star are applied. The discrete maximum principle guarantees the stability of the finite difference method. Now we can conclude the convergence.

**Theorem 5** *Let $u \in C^3(\bar{\Omega})$ be the solution satisfying the Poisson equation $-\Delta u = f$ with Dirichlet boundary conditions $u = g$ on $\partial\Omega$. Let $L_h$ be the difference operator of the five-point star. Then the numerical solution of the finite difference method converges to the exact solution and it holds*

$$\max_{Z \in \Omega_h} |u(Z) - U_Z| = \mathcal{O}(h).$$

*If $u \in C^4(\bar{\Omega})$ and all step sizes are equidistant, then it holds*

$$\max_{Z \in \Omega_h} |u(Z) - U_Z| = \mathcal{O}(h^2).$$

Proof:

The consistency of the five-point star yields $\tau(Z) = \mathcal{O}(h)$ for all $Z \in \Omega_h$. The estimate (2.15) from Lemma 1 implies

$$\max_{Z \in \Omega_h} |\eta(Z)| \leq \tfrac{1}{4} R^2 C h \qquad \text{for all } 0 < h < h_0$$

with some constant $C > 0$. In case of $u \in C^4(\bar{\Omega})$ and equidistant step sizes, it follows $\tau(Z) = \mathcal{O}(h^2)$ and thus convergence of order $p = 2$. □

A further analysis shows that a convergence of order $p = 2$ is also given for variable step sizes. Moreover, the convergence can also be shown in case of $u \in C^2(\bar{\Omega})$.

### Generalisations

We outline the generalisation of the above finite difference method to further problems.

- *von-Neumann boundary value problem*: We consider the Poisson equation $-\Delta u = f$ in $\Omega \subset \mathbb{R}^2$ with boundary conditions $\frac{\partial u}{\partial \nu} = g(x,y)$ on $\partial\Omega$. For example, let $\Omega = (0,1) \times (0,1)$. We apply the grid points (2.8) for $i,j = 0,1,\ldots,M,M+1$. In comparison to Dirichlet problems, $4M$ additional unknowns appear (the four edges of the square are not used). Hence we arrange $4M$ equations using difference formulas to replace the derivative $\frac{\partial u}{\partial \nu}$:

$$\begin{array}{rcrcll} g(x_i,0) & = & -\frac{\partial u}{\partial y}(x_i,0) & \doteq & \frac{1}{h}[u_{i,0} - u_{i,1}] & \text{for } i = 1,\ldots,M\,, \\ g(x_i,1) & = & \frac{\partial u}{\partial y}(x_i,1) & \doteq & \frac{1}{h}[u_{i,M+1} - u_{i,M}] & \text{for } i = 1,\ldots,M\,, \\ g(0,y_j) & = & -\frac{\partial u}{\partial x}(0,y_j) & \doteq & \frac{1}{h}[u_{0,j} - u_{1,j}] & \text{for } j = 1,\ldots,M\,, \\ g(1,y_j) & = & \frac{\partial u}{\partial x}(1,y_j) & \doteq & \frac{1}{h}[u_{M+1,j} - u_{M,j}] & \text{for } j = 1,\ldots,M\,. \end{array}$$

  Techniques for Neumann boundary conditions on arbitrary domains $\Omega$ also exist. Remark that a solution of a pure von-Neumann boundary value problem is not unique and thus requires an additional condition.

- *space-dependent coefficients*: Given the elliptic PDE

$$Lu := a(x,y)\,\frac{\partial^2 u}{\partial x^2} + 2b(x,y)\,\frac{\partial^2 u}{\partial x \partial y} + c(x,y)\,\frac{\partial^2 u}{\partial y^2} = f(x,y),$$

  the coefficients $a,b,c : \Omega \to \mathbb{R}$ depend on space. Appropriate finite difference methods can be constructed in this case. For proving the convergence, a uniformly elliptic operator, see Definition 2, has to be assumed.

- *right-hand side includes the solution*: We consider the (nonlinear) PDE $-\Delta u = f(x,y,u)$ with a nonlinear function $f$. Let $\Omega$ be the unit square. Homogeneous boundary condition $u = 0$ on $\partial\Omega$ are applied. The five-point star with equidistant step sizes yields the equations

$$\tfrac{1}{h^2}[4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}] = f(x_i, y_j, u_{i,j})$$

for $i, j = 1, \ldots, M$. We obtain a nonlinear system for the unknowns $u_{i,j}$. Newton's method yields an approximation of the corresponding discrete solution. In the special case $f(x, y, u) = b(x, y) + c(x, y)u$, it follows a linear system again.

- *three-dimensional space*: The Laplace operator in three space dimensions reads $\Delta u = u_{xx} + u_{yy} + u_{zz}$. We consider an open and bounded domain $\Omega \subset \mathbb{R}^3$ for the Poisson equation $-\Delta u = f$. The above theory of finite difference methods can be repeated in this case. The same results appear concerning consistency, stability and convergence.

### Iterative solution of linear systems

The finite difference methods yield large and sparse linear systems. Gaussian elimination becomes expensive, since many fill-ins appear in the factorisation. In contrast, iterative methods allow for efficient algorithms.

The types of iterative solvers are:

- *stationary methods*: Jacobi method, Gauss-Seidel method, SOR, etc.
- *instationary methods*: conjugate gradient method, GMRES, etc.
- *multigrid methods*.

An introduction to iterative methods for solving linear systems can be found in J. Stoer, R. Bulirsch: Introduction to Numerical Analysis. (2nd ed.) Springer, New York, 1993. (Chapter 8)

## 2.3 Sobolev Spaces and Variational Formulation

In this section, we introduce weak solutions of elliptic PDEs. Finite element methods represent convergent techniques for weak solutions, whereas finite difference methods fail.

### Classical solutions

Classical solutions of a PDE are sufficiently smooth in some sense.

**Definition 7 (classical solution)** *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. For an elliptic PDE $Lu = f$, a function $u : \bar{\Omega} \to \mathbb{R}$ is called a classical solution, if it holds*

- *for Dirichlet problems: $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$,*
- *for von-Neumann problems: $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$.*

As an example for $n = 2$, we consider the Laplace equation on three quarter of the unit disc as domain

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

We identify $\Omega \subset \mathbb{C}$ via $x = \mathrm{Re}(z)$, $y = \mathrm{Im}(z)$. The function $w : \Omega \to \mathbb{C}$, $w(z) := z^{2/3}$ is analytic. The imaginary part $u := \mathrm{Im}(w)$ satisfies

$$\begin{array}{rcll} \Delta u & = & 0 & \text{in } \Omega, \\ u(\mathrm{e}^{\mathrm{i}\varphi}) & = & \sin(\frac{2}{3}\varphi) & \text{for } 0 \le \varphi \le \frac{3\pi}{2}, \\ u & = & 0 & \text{elsewhere on } \partial\Omega. \end{array}$$

We obtain the representation

$$u(x, y) = \sqrt[3]{x^2 + y^2}\,\sin(\tfrac{2}{3}\arctan(\tfrac{y}{x})) \qquad \text{for } x > 0.$$

It holds $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$, i.e., $u$ is a classical solution of the Dirichlet problem. However, due to $w'(z) = \frac{2}{3}z^{-1/3}$ and $w''(z) = -\frac{2}{9}z^{-4/3}$, both the first and the second derivative of $u$ is not bounded in a neighbourhood

of $z = 0$. It follows that $u \notin C^2(\bar{\Omega})$. Hence we cannot guarantee the convergence of the finite difference method constructed in Sect. 2.2. An alternative numerical method is required.

### Weak Derivatives and Sobolev Spaces

We consider an open domain $\Omega \subseteq \mathbb{R}^n$. We define the space of test functions

$$C_0^\infty(\Omega) := \{\phi \in C^\infty(\Omega) : \text{supp}(\phi) \subset \Omega,\ \text{supp}(\phi) \text{ is compact}\},$$

where $\text{supp}(\phi) := \overline{\{x \in \Omega : \phi(x) \neq 0\}}$. If $\Omega$ is bounded, it follows that $\phi(x) = 0$ for $x \in \partial\Omega$. Furthermore, we apply the Hilbert space $L^2(\Omega)$, which has the inner product

$$\langle f, g\rangle_{L^2} := \int_\Omega f(x) \cdot g(x)\, \mathrm{d}x$$

for each $f, g \in L^2(\Omega)$. The corresponding norm reads

$$\|f\|_{L^2} = \sqrt{\int_\Omega f(x)^2\, \mathrm{d}x}\ .$$

The set $C_0^\infty(\Omega) \subset L^2(\Omega)$ is dense. A multi-index is given by

$$\alpha := (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n, \qquad |\alpha| := \sum_{i=1}^n \alpha_i.$$

For $u \in C^k(\Omega)$ with $k = |\alpha|$, an elementary differential operator can be defined via

$$\mathrm{D}^\alpha u := \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

For $u \in C^k(\Omega)$, $\mathrm{D}^\alpha u$ represents a usual (strong) derivative.

**Definition 8 (weak derivative)** *Given a function $f \in L^2(\Omega)$, a function $g \in L^2(\Omega)$ is called the weak derivative $\mathrm{D}^\alpha f$ of $f$, if it holds*

$$\int_\Omega g(x) \cdot \phi(x)\, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f(x) \cdot \mathrm{D}^\alpha \phi(x)\, \mathrm{d}x$$

*for all $\phi \in C_0^\infty(\Omega)$. We write $\mathrm{D}^\alpha f = g$.*

The property of this definition can also be written as

$$\langle g, \phi \rangle_{L^2} = (-1)^{|\alpha|} \langle f, \mathrm{D}^\alpha \phi \rangle_{L^2} \qquad \text{for all } \phi \in C_0^\infty(\Omega).$$

It can be shown that a weak derivative of a function is unique, since the space $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$.

**Example:** In the special case $n = 1$, a function $f \in C^1(a, b)$ leads to

$$\int_a^b f'(x)\phi(x)\,\mathrm{d}x = [f(x)\phi(x)]_{x=a}^{x=b} - \int_a^b f(x)\phi'(x)\,\mathrm{d}x = -\int_a^b f(x)\phi'(x)\,\mathrm{d}x$$

for each $\phi \in C_0^\infty(a, b)$. Hence $f'$ is a weak derivative of $f$.

In case of $n \geq 2$, we apply Green's formula

$$\int_\Omega v \cdot \frac{\partial w}{\partial x_i}\,\mathrm{d}x = \int_{\partial\Omega} v \cdot w \cdot \nu_i\,\mathrm{d}s - \int_\Omega \frac{\partial v}{\partial x_i} \cdot w\,\mathrm{d}x \tag{2.16}$$

for $i \in \{1, \ldots, n\}$ and $v, w \in C^1(\bar{\Omega})$, where $\nu_i$ is the $i$th component of the outer normal vector on the boundary $\partial\Omega$. The formula (2.16) does not hold for arbitrary domains $\Omega$. A domain with a (so-called) smooth boundary is sufficient for the application of the formula. For simplicity, we restrict to domains, where the formula (2.16) holds.

The concept of the weak derivative is used to define the Sobolev spaces.

**Definition 9 (Sobolev space)** *For $m \geq 0$, $H^m(\Omega)$ is the set of all functions $u \in L^2(\Omega)$, where a weak derivative $\mathrm{D}^\alpha u \in L^2(\Omega)$ exists for all $|\alpha| \leq m$. The space $H^m(\Omega)$ exhibits an inner product*

$$\langle u, v \rangle_{H^m} := \sum_{|\alpha| \leq m} \langle \mathrm{D}^\alpha u, \mathrm{D}^\alpha v \rangle_{L^2}.$$

*The set $H^m(\Omega)$ is called a Sobolev space. $\|\cdot\|_{H^m}$ is a Sobolev norm.*

Hence $H^m(\Omega)$ can be interpreted as a generalisation of the space $C^m(\Omega)$, which is not a Hilbert space. The spaces $(H^m(\Omega), \|\cdot\|_{H^m})$ are Hilbert spaces, i.e., they are complete. Remark that it holds $H^m(\Omega) \subset L^2(\Omega)$ for $m \geq 1$ and $H^0(\Omega) = L^2(\Omega)$.

To investigate PDE problems with homogeneous Dirichlet boundary conditions ($\Omega$ is bounded), we define the subsets $H_0^m(\Omega) :\equiv \overline{C_0^\infty(\Omega)}$ as the closure of $C_0^\infty(\Omega) \subset H^m(\Omega)$ with respect to the norm $\|\cdot\|_{H^m}$. More precisely, it holds

$$H_0^m(\Omega) = \left\{ u \in H^m(\Omega) : \exists (v_i)_{i\in\mathbb{N}} \subset C_0^\infty(\Omega) \text{ with } \lim_{i\to\infty} \|u - v_i\|_{H^m} = 0 \right\}.$$

It follows that the subspace $(H_0^m, \|\cdot\|_{H^m})$ is also a Hilbert space.

Furthermore, we obtain a semi-norm on $H^m(\Omega)$ via

$$|u|_{H^m} := \left( \sum_{|\alpha|=m} \|D^\alpha u\|_{L^2}^2 \right)^{1/2}.$$

If $\Omega$ is inside a cube of edge length $s$, then it holds the equivalence

$$|u|_{H^m} \le \|u\|_{H^m} \le (1+s)^m |u|_{H^m} \quad \text{for all } u \in H_0^m(\Omega). \tag{2.17}$$

The above construction can also be done using the spaces $L^p(\Omega)$ instead of $L^2(\Omega)$ with $1 \le p \le \infty$. It follows the Sobolev spaces $W_p^m(\Omega)$, where it holds $W_2^m(\Omega) = H^m(\Omega)$.

### Symmetric operators and bilinear forms

In the following, we assume homogeneous Dirichlet boundary conditions. Given an elliptic PDE $Lu = f$ in $\Omega$ and $u = g$ on $\partial\Omega$. Let $u_0$ be a sufficiently smooth function with $u_0 = g$ on $\partial\Omega$. The function $w := u - u_0$ satisfies $Lw = \tilde{f}$ in $\Omega$ with $\tilde{f} := f - Lu_0$ and $w = 0$ on $\partial\Omega$. Hence we have transformed the problem to homogeneous boundary conditions.

We consider a general linear differential operator of the form

$$Lu := -\left( \sum_{i,j=1}^n a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} \right) + \left( \sum_{j=1}^n a_j \frac{\partial u}{\partial x_j} \right) + a_0 u. \tag{2.18}$$

Thereby, we assume $a_{ij} \in C^2(\bar{\Omega})$, $a_j \in C^1(\bar{\Omega})$, $a_0 \in C^0(\bar{\Omega})$. The corresponding adjoint operator $L^*$ is defined by the property

$$\langle Lu, v\rangle_{L^2} = \langle u, L^* v\rangle_{L^2}$$

for $u, v \in C^2(\bar{\Omega})$ with $u = 0, v = 0$ on $\partial\Omega$. It follows

$$L^* v = -\left( \sum_{i,j=1}^{n} \frac{\partial^2 (a_{ij} v)}{\partial x_i \partial x_j} \right) - \left( \sum_{j=1}^{n} \frac{\partial (a_j v)}{\partial x_j} \right) + a_0 v.$$

A symmetric (also: self-adjoint) operator satisfies $L = L^*$. It can be shown that a symmetric operator exhibits the form

$$Lu := -\left( \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left( a_{ij} \frac{\partial u}{\partial x_j} \right) \right) + a_0 u. \tag{2.19}$$

In particular, each operator (2.18) is self-adjoint in case of constant coefficients $a_{ij}$ and $a_1 = \cdots = a_n = 0$, for example. Green's formula (2.16) implies

$$\langle Lu, v \rangle = \left( \sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} \, \mathrm{d}x \right) + \int_\Omega a_0 uv \, \mathrm{d}x. \tag{2.20}$$

We recognise that the right-hand side is symmetric in $u$ and $v$. Hence it holds $\langle Lu, v \rangle = \langle u, Lv \rangle$. Furthermore, just derivatives of first order appear.

**Definition 10** *Let $H$ be a Hilbert space. A bilinear form $a : H \times H \to \mathbb{R}$ is symmetric, if $a(u, v) = a(v, u)$ holds for all $u, v \in H$. A bilinear form $a$ is continuous , if a constant $C > 0$ exists such that*

$$|a(u, v)| \leq C \cdot \|u\| \cdot \|v\| \qquad \text{for all } u, v \in H.$$

*A symmetric, continuous bilinear form $a$ is called $H$-elliptic (also: coercive), if a constant $\beta > 0$ exists with*

$$a(u, u) \geq \beta \cdot \|u\|^2 \qquad \text{for all } u \in H.$$

In particular, each $H$-elliptic bilinear form is positive, i.e., $a(u, u) > 0$ for $u \neq 0$.

Each $H$-elliptic bilinear form $a$ induces a norm

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

which is called the energy norm. The energy norm is equivalent to the norm of the Hilbert space.

The relation (2.20) motivates the definition of a symmetric bilinear form corresponding to a uniformly elliptic differential operator.

**Theorem 6** *The bilinear form* $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ *given by*

$$a(u,v) := \left( \sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} \, \mathrm{d}x \right) + \int_\Omega a_0 uv \, \mathrm{d}x \tag{2.21}$$

*with* $a_{ij}, a_0 \in C^0(\bar{\Omega})$, $A = (a_{ij})$ *symmetric and positive definite,* $a_0 \geq 0$ *is continuous and* $H_0^1(\Omega)$*-elliptic provided that the underlying differential operator is uniformly elliptic.*

Proof:

We define $c := \sup\{|a_{ij}(x)| : x \in \Omega,\ 1 \leq i,j \leq n\}$. It follows using the Cauchy-Schwarz inequality

$$\begin{aligned} \left| \sum_{i,j=1}^{n} \int_\Omega a_{ij} u_{x_i} v_{x_j} \, \mathrm{d}x \right| &\leq c \sum_{i,j=1}^{n} \int_\Omega |u_{x_i} v_{x_j}| \, \mathrm{d}x \\ &\leq c \sum_{i,j=1}^{n} \|u_{x_i}\|_{L^2} \cdot \|v_{x_j}\|_{L^2} \\ &\leq c \sum_{i,j=1}^{n} |u|_{H^1} \cdot |v|_{H^1} \\ &= cn^2 \cdot |u|_{H^1} \cdot |v|_{H^1}. \end{aligned}$$

We arrange $b := \sup\{|a_0(x)| : x \in \Omega\}$. It follows

$$\left| \int_\Omega a_0 uv \, \mathrm{d}x \right| \leq b \int_\Omega |uv| \, \mathrm{d}x \leq b \cdot \|u\|_{L^2} \cdot \|v\|_{L^2}.$$

We obtain applying $\|u\|_{L^2} \leq \|u\|_{H^1}$ and $|u|_{H^1} \leq \|u\|_{H^1}$ with $C := b + cn^2$

$$|a(u,v)| \leq C \cdot \|u\|_{H^1} \cdot \|v\|_{H^1}.$$

Since the differential operator is uniformly elliptic, see (2.6), it holds using the monotonicity of the integral

$$\int_\Omega \sum_{i,j=1}^n a_{ij} v_{x_i} v_{x_j} \,\mathrm{d}x \geq \alpha \int_\Omega \sum_{i=1}^n (v_{x_i})^2 \,\mathrm{d}x$$

for $v \in H_0^1(\Omega)$. It follows due to $a_0 \geq 0$

$$a(v,v) \geq \alpha \sum_{i=1}^n \int_\Omega (v_{x_i})^2 \,\mathrm{d}x = \alpha |v|_{H^1}^2$$

for each $v \in H_0^1(\Omega)$. The equivalence (2.17) implies $a(v,v) \geq \alpha K \|v\|_{H^1}^2$ for $v \in H_0^1(\Omega)$ with some constant $K > 0$ depending just on $\Omega$. Thus the bilinear form $a$ is $H_0^1(\Omega)$-elliptic with $\beta := \alpha K$. □

### Variational formulation

Now we consider the PDE $Lu = f$ with a uniformly elliptic operator $L$. Let $u$ be a classical solution and $Lu, f \in L^2(\Omega)$. It follows

$$\begin{aligned} Lu - f &= 0 \\ (Lu - f)v &= 0 \\ \langle Lu - f, v \rangle_{L^2} &= 0 \\ \langle Lu, v \rangle_{L^2} - \langle f, v \rangle_{L^2} &= 0 \end{aligned}$$

for each $v \in L^2(\Omega)$. We define the linear mapping

$$\ell(v) := \langle f, v \rangle_{L^2} \tag{2.22}$$

for $v \in L^2(\Omega)$ or a corresponding subspace. The Cauchy-Schwarz inequality yields $|\ell(v)| \leq \|f\|_{L^2} \|v\|_{L^2}$. Hence $\ell$ is bounded on $L^2(\Omega)$ with $\|\ell\| \leq \|f\|_{L^2}$. It follows that $\ell$ is also bounded on $H^1(\Omega)$.

Let $\ell \in V'$, i.e., $\ell : V \to \mathbb{R}$ be an arbitrary linear mapping. We apply the notation $\langle \ell, v \rangle := \ell(v)$, which refers to the bilinear form $\langle \cdot, \cdot \rangle : V' \times V \to \mathbb{R}$.

The right-hand side $f$ yields the linear mapping (2.22), whereas the left-hand side $Lu$ corresponds to the bilinear form (2.21). It follows the concept of a weak solution.

**Definition 11 (weak solution)** *A function $u \in H_0^1(\Omega)$ is called a weak solution of the elliptic PDE problem*

$$\begin{aligned} Lu &= f \quad in\ \Omega \\ u &= 0 \quad on\ \partial\Omega, \end{aligned}$$

*if the corresponding bilinear form (2.21) and linear mapping (2.22) satisfy*

$$a(u,v) = \langle \ell, v \rangle \qquad for\ all\ v \in H_0^1(\Omega). \tag{2.23}$$

Now we show that a classical solution represents also a weak solution of the problem. For simplicity, we demand the property $u \in C^2(\bar{\Omega})$ for the classical solution.

**Theorem 7** *Let $u$ be a classical solution of*

$$\begin{aligned} -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + a_0 u &= f \quad in\ \Omega \\ u &= 0 \quad on\ \partial\Omega \end{aligned}$$

*with $u \in C^2(\bar{\Omega})$ and $a_{ij} \in C^1(\bar{\Omega})$, $a_0, f \in C^0(\Omega) \cap L^2(\Omega)$. Then $u$ represents also a weak solution of the problem.*

Proof:

We apply Green's formula

$$\int_\Omega v \cdot \frac{\partial w}{\partial x_i}\,\mathrm{d}x = \int_{\partial\Omega} v \cdot w \cdot \nu_i\,\mathrm{d}s - \int_\Omega \frac{\partial v}{\partial x_i} \cdot w\,\mathrm{d}x,$$

where we choose $w := a_{ij}u_{x_j} \in C^1(\bar{\Omega})$ and $v \in C_0^\infty(\Omega)$. It follows

$$\int_\Omega v\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right)\,\mathrm{d}x = -\int_\Omega a_{ij}\frac{\partial v}{\partial x_i}\frac{\partial u}{\partial x_j}\,\mathrm{d}x.$$

We apply the bilinear and linear form, respectively,

$$a(u,v) := \int_\Omega \sum_{i,j=1}^{n} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + a_0 uv\,\mathrm{d}x, \qquad \langle \ell, v\rangle := \int_\Omega fv\,\mathrm{d}x.$$

It follows

$$\begin{aligned} a(u,v) - \langle \ell, v \rangle &= \int_\Omega v \left[ -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left( a_{ij} \frac{\partial u}{\partial x_j} \right) + a_0 u - f \right] \mathrm{d}x \\ &= \int_\Omega v[Lu - f] \, \mathrm{d}x \;=\; 0 \end{aligned}$$

for all $v \in C_0^\infty(\Omega)$ due to $Lu = f$. The bilinear form $a$ as well as the linear form $\ell(v) = \langle f, v \rangle_{L^2}$ are continuous on $H_0^1(\Omega)$. Since $C_0^\infty(\Omega) \subset H_0^1(\Omega)$ is dense, it follows $a(u,v) - \langle \ell, v \rangle = 0$ for all $v \in H_0^1(\Omega)$. Furthermore, it holds $u \in H_0^1(\Omega)$ due to $u \in C^0(\bar{\Omega}) \cap H^1(\Omega)$ and $u = 0$ on $\partial\Omega$. □

Solutions satifying the assumptions of Theorem 7 can be computed by finite difference methods due to $u \in C^2(\bar{\Omega})$.

Now we show an important equivalence of our problem.

**Theorem 8** *Let $V$ be a linear space and $a : V \times V \to \mathbb{R}$ a symmetric, positive bilinear form and $\ell : V \to \mathbb{R}$ a linear mapping. The function*

$$J(v) := \tfrac{1}{2} a(v,v) - \langle \ell, v \rangle \tag{2.24}$$

*exhibits a minimum in $V$ at $u$ if and only if*

$$a(u,v) = \langle \ell, v \rangle \quad \textit{for all } \; v \in V. \tag{2.25}$$

*There exists at most one minimum.*

Proof:

A positive bilinear form fulfills $a(u,u) > 0$ for all $u \neq 0$. For $u, v \in V$ and $t \in \mathbb{R}$, we calculate

$$\begin{aligned} J(u+tv) &= \tfrac{1}{2} a(u+tv, u+tv) - \langle \ell, u+tv \rangle \\ &= J(u) + t\,[a(u,v) - \langle \ell, v \rangle] + \tfrac{1}{2} t^2 a(v,v). \end{aligned}$$

If $u \in V$ satisfies the condition (2.25), then it follows using $t = 1$

$$J(u+v) = J(u) + \tfrac{1}{2} a(v,v) > J(u)$$

for $v \in V$ with $v \neq 0$. Hence $u$ is the unique minimum.

Vice versa, let $u \in V$ be a minimum of the function (2.24). For each $v \in V$, it holds

$$\frac{\mathrm{d}}{\mathrm{d}t} J(u + tv)\Big|_{t=0} = 0.$$

Due to

$$\frac{\mathrm{d}}{\mathrm{d}t} J(u + tv) = a(u, v) - \langle \ell, v \rangle + ta(v, v),$$

it follows the condition (2.25). □

Theorem 8 yields the uniqueness of a weak solution. If a classical solution exists satisfying additional properties (like $u \in C^2(\bar{\Omega})$), then this function also represents the unique weak solution.

We obtain an additional characterisation of the weak solution by Theorem 8: The weak solution of the PDE also represents a solution of a minimisation problem

$$J(v) := \tfrac{1}{2} a(v, v) - \langle \ell, v \rangle \quad \longrightarrow \quad \min. \tag{2.26}$$

and vice versa. The task (2.26) is called a variational formulation (of the problem) or a variational problem.

**Theorem 9 (Lax-Milgram)** *Let $H$ be a Hilbert space and $V \subseteq H$ be a closed convex set. For an $H$-elliptic bilinear form $a : H \times H \to \mathbb{R}$ and $\ell \in H'$, the variational problem (2.26) exhibits a unique solution in $V$.*

Proof:

The mapping $J$ is bounded from below, sice it holds

$$J(v) \geq \tfrac{1}{2}\alpha \|v\|^2 - \|\ell\| \cdot \|v\| = \tfrac{1}{2\alpha}(\alpha\|v\| - \|\ell\|)^2 - \tfrac{1}{2\alpha}\|\ell\|^2 \geq -\tfrac{1}{2\alpha}\|\ell\|^2.$$

We define $c := \inf\{J(v) : v \in V\}$. Let $(v_n)_{n \in \mathbb{N}} \subset V$ be a sequence satisfying

$$\lim_{n \to \infty} J(v_n) = c.$$

It follows

$$\begin{aligned}\alpha\|v_n - v_m\|^2 &\leq a(v_n - v_m, v_n - v_m)\\ &= 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m)\\ &= 4J(v_n) + 4J(v_m) - 8J(\tfrac{1}{2}(v_n + v_m))\\ &\leq 4J(v_n) + 4J(v_m) - 8c,\end{aligned}$$

since $\frac{1}{2}(v_n + v_m) \in V$ holds due to $V$ convex. The upper bound converges to zero for $n, m \to \infty$. Thus $\|v_n - v_m\| \to 0$ for $n, m \to \infty$, i.e., $(v_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Since $V$ is a closed set, a limit $u \in V$ exists. It follows

$$J(u) = J\left(\lim_{n\to\infty} v_n\right) = \lim_{n\to\infty} J(v_n) = \inf\{J(v) : v \in V\}$$

due to the continuity of $J$. Hence $u$ represents a minimum.

Concerning the uniqueness, let $u_1, u_2 \in V$ be two solutions of the variational problem (2.26). Then it holds $J = c$ for each component of the sequence $(u_1, u_2, u_1, u_2, \ldots)$. Due to the above calculations, a Cauchy sequence is given. It follows $\|u_1 - u_2\| < \varepsilon$ for each $\varepsilon > 0$. Hence it holds $\|u_1 - u_2\| = 0$ and $u_1 = u_2$. □

We apply Theorem 9 in the special case $V = H = H_0^1(\Omega)$, since $H_0^1(\Omega)$ is a Hilbert space. It follows that the variational problem (2.26) has a unique solution $u \in H_0^1(\Omega)$. Due to Theorem 8, the solution $u$ of the variational problem is also a weak solution of the PDE problem according to Definition 11. We obtain directly the following result.

**Theorem 10** *Let $L$ be a uniformly elliptic, symmetric differential operator. Then the homogeneous Dirichlet boundary value problem for $Lu = f$ exhibits a unique weak solution in $H_0^1(\Omega)$.*

**Remark:** Not all open and bounded domains $\Omega \subset \mathbb{R}^n$ are feasible, since Green's formula has to be applicable. Nevertheless, Green's formula is valid for nearly all domains in practice.

In the one-dimensional case ($n = 1$), we obtain a homogeneous boundary value problem of a second-order ordinary differential equation

$$-(p(x)u'(x))' + q(x)u(x) = f(x) \quad \text{for } x \in (a, b)$$

with $u(a) = u(b) = 0$. Assuming $p(x) \geq p_0 > 0$ and $q(x) \geq 0$, the involved differential operator is uniformly elliptic. A corresponding variational formulation can be constructed as above. The detailed derivation for this special case can be found in, for example, Stoer/Bulirsch: Introduction to Numerical Analysis, Springer (Section 7.5).

### Von-Neumann boundary conditions

For the Poisson equation $-\Delta u = f$, we demand $\frac{\partial u}{\partial \nu} = g$ on $\partial\Omega$ in case of von-Neumann boundary conditions. Weak solutions are considered in the space $H^1(\Omega)$ now. For a broad class of domains $\Omega$, a bounded linear mapping

$$\gamma : H^1(\Omega) \to L^2(\partial\Omega), \quad \|\gamma(v)\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)} \tag{2.27}$$

exists satisfying $\gamma(v) = v|_{\partial\Omega}$ for all $v \in C^0(\bar{\Omega}) \cap H^1(\Omega)$. The linear mapping $\gamma$ is called the trace operator.

The trace operator allows for an alternative characterisation of the Hilbert space $H_0^1(\Omega)$. We defined $H_0^1(\Omega) :\equiv \overline{C_0^\infty(\Omega)}$, i.e., the closure of the test functions with respect to the Sobolev norm $\|\cdot\|_{H^1}$. It can be shown that it holds

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : \gamma(u) = 0\}.$$

This property has already been used in the proof of Theorem 7, since $u \in C^0(\bar{\Omega}) \cap H^1(\Omega)$ and $u = 0$ on $\partial\Omega$ implies $u \in H_0^1(\Omega)$ now.

A variational formulation can be derived also in the case of von-Neumann boundary conditions. For simplicity, let $u \in C^2(\bar{\Omega}), v \in C^1(\bar{\Omega})$. Green's formula yields for $\langle \Delta u, v \rangle_{L^2}$

$$\int_\Omega v \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} \, \mathrm{d}x = \sum_{i=1}^n \int_\Omega v \frac{\partial^2 u}{\partial x_i^2} \, \mathrm{d}x = \sum_{i=1}^n \int_{\partial\Omega} v \frac{\partial u}{\partial x_i} \nu_i \, \mathrm{d}s - \int_\Omega \frac{\partial v}{\partial x_i} \frac{\partial u}{\partial x_i} \, \mathrm{d}x.$$

The second term implies the definition of the bilinear form

$$a(u, v) = \int_\Omega \sum_{i=1}^n \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, \mathrm{d}x \tag{2.28}$$

as in the case of Dirichlet boundary conditions. The first terms are not identical to zero now. It follows

$$\sum_{i=1}^{n} \int_{\partial\Omega} v \frac{\partial u}{\partial x_i} \nu_i \, \mathrm{d}s = \int_{\partial\Omega} v \left( \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \nu_i \right) \, \mathrm{d}s = \int_{\partial\Omega} v (\underbrace{\nabla u \cdot \nu}_{=\frac{\partial u}{\partial \nu}}) \, \mathrm{d}s = \int_{\partial\Omega} vg \, \mathrm{d}s.$$

Hence the corresponding linear mapping $\ell$ reads

$$\langle \ell, v \rangle := \int_{\Omega} fv \, \mathrm{d}x + \int_{\partial\Omega} gv \, \mathrm{d}s \tag{2.29}$$

assuming $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$. Now functions $v \in H^1(\Omega)$ can be considered by applying the operator (2.27). More precisely, the linear mapping (2.29) changes into

$$\langle \ell, v \rangle := \int_{\Omega} fv \, \mathrm{d}x + \int_{\partial\Omega} g\gamma(v) \, \mathrm{d}s$$

for $v \in H^1(\Omega)$. Nevertheless, the notation (2.29) is applied in general.

The linear mapping (2.29) includes both the information of the right-hand side $f$ and the boundary conditions $g$. Numerical methods can be constructed for solving the variational problem or its equivalent conditions. Recall again that a solution of a pure von-Neumann boundary value problem is not unique ($u$ solution implies $u + c$ solution for arbitrary $c \in \mathbb{R}$). Hence an additional condition has to be included.

More details can be found in Braess: Finite Elements. (Chapter 3)

## 2.4 Finite Element Methods

Now we apply the theory of the previous section to construct numerical methods for the determination of weak solutions.

### Ritz-Galerkin approach

We consider a homogeneous Dirichlet boundary value problem including a uniformly elliptic, symmetric differential operator (2.19). It follows the existence of a unique weak solution. Theorem 8 shows two properties, which characterise the weak solution. Numerical methods can be based on each of these properties. Typically, finite-dimensional subspaces $S_h \subset H_0^1(\Omega)$ are chosen, where $h > 0$ represents a discretisation step size to be defined later. Typically, it holds $\dim(S_h) \to \infty$ for $h \to 0$.

We obtain three classes of numerical techniques:

- *Galerkin method*: The definition (2.23) is used. The approximation of the weak solution is determined in a finite dimensional subspace $S_h$. The condition (2.23) shall be satisfied for all $v \in S_h$.
- *Petrov-Galerkin method* (or: *method of weighted residuals*): The definition (2.23) is applied again. The approximation is situated in some space $S_h$. The condition (2.23) shall be satisfied for all $v \in T_h$ with another subspace $T_h$ of the same dimension. The special case $S_h = T_h$ yields the Galerkin method.
- *Rayleigh-Ritz method* (or: *Ritz method*): The solution of the variational problem (2.26) is computed approximately. Thereby, a minimum of $J$ is determined in a finite-dimensional subspace $S_h$.

We discuss the Galerkin method first. For some subspace $S_h$, we choose a basis $\{\phi_1, \ldots, \phi_N\}$. The approximation is inside this space, i.e.,

$$u_h(x) = \sum_{j=1}^{N} \alpha_j \phi_j(x) \tag{2.30}$$

with unknown coefficients $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$. Replacing the exact solution $u$ by the approximation $u_h$, the condition (2.23) demands that

$$a(u_h, v) = \langle \ell, v \rangle \tag{2.31}$$

for all $v \in H_0^1(\Omega)$. Since $u_h \neq u$ in general, this condition cannot be satisfied for all $v \in H_0^1(\Omega)$. Alternatively, we demand that the property (2.23) holds for all $v \in S_h$. Using the basis functions, this condition is equivalent to

$$a(u_h, \phi_i) = \langle \ell, \phi_i \rangle \qquad \text{for } \; i = 1, \ldots, N.$$

Inserting (2.30), it follows a linear system

$$\sum_{j=1}^{N} \alpha_j a(\phi_j, \phi_i) = \langle \ell, \phi_i \rangle \qquad \text{for } \; i = 1, \ldots, N$$

with the unknown coefficients $\alpha_1, \ldots, \alpha_N$. The matrix of this linear system reads $A := (a(\phi_j, \phi_i)) \in \mathbb{R}^{N \times N}$, which is obviously symmetric. Since the bilinear form is positive, it holds for $\xi = (\xi_1, \ldots, \xi_N)^\top \neq 0$

$$\xi^\top A \, \xi = \sum_{i,j=1}^{N} a(\phi_j, \phi_i) \xi_j \xi_i = a\left( \sum_{j=1}^{N} \xi_j \phi_j, \sum_{i=1}^{N} \xi_i \phi_i \right) > 0.$$

Hence the matrix $A$ is positive definite. It follows that a unique solution exists, which yields the approximation (2.30).

In the Petrov-Galerkin method, we demand that the condition (2.31) is satisfied for all $v \in T_h$ for some other subspace $T_h \subset H_0^1(\Omega)$ satisfying $\dim(S_h) = \dim(T_h)$. The elements of $T_h$ are often called test functions. (However, they do not belong to the set $C_0^\infty$ in general.) We select a basis $\{\psi_1, \ldots, \psi_N\}$ of $T_h$. Now the condition (2.31) for all $v \in T_h$ is equivalent to

$$a\left( \sum_{j=1}^{N} \alpha_j \phi_j, \psi_i \right) = \langle \ell, \psi_i \rangle \quad \text{for } \; i = 1, \ldots, N.$$

It follows the linear system

$$\sum_{j=1}^{N} \alpha_j a(\phi_j, \psi_i) = \langle \ell, \psi_i \rangle \quad \text{for } \; i = 1, \ldots, N$$

with the matrix $A := (a(\phi_j, \psi_i))$. This matrix is not symmetric in general. It depends on the choice of the subspaces and the bases if the matrix is regular. In the special case $S_h = T_h$ and using the same basis, the approach coincides with the Galerkin method due to $\phi_i = \psi_i$ for all $i$.

For the Rayleigh-Ritz method, we insert the approximation (2.30) into the function $J$ from (2.26). It follows ($\alpha = (\alpha_1, \dots, \alpha_N)^\top$)

$$J\left(\sum_{j=1}^{N} \alpha_j \phi_j\right) = \tfrac{1}{2} \sum_{i,j=1}^{N} \alpha_j \alpha_i a(\phi_j, \phi_i) - \sum_{j=1}^{N} \alpha_j \langle \ell, \phi_j \rangle = \tfrac{1}{2} \alpha^\top A \alpha - \alpha^\top b$$

with the same matrix $A$ and right-hand side $b$ as in the Galerkin method. The minimisation in $S_h$ only demands

$$\frac{\partial J}{\partial \alpha_k} = 0 \quad \text{for } \; k = 1, \dots, N.$$

The gradient of $J$ is $\nabla J = A\alpha - b$. It follows the linear system $A\alpha = b$. Thus the technique coincides with the Galerkin method in this case. Different approaches may appear if the underlying bilinear form $a$ is not symmetric or not positive. For problems, where the Rayleigh-Ritz method and the Galerkin method are the same, the technique is called the *Ritz-Galerkin method*. The involved matrix $A_h$ is also calles *stiffness matrix*.

The method of weighted residuals can be motivated also in case of smooth solutions. Let $u_h \in S_h \subset C^2(\Omega) \cap C^0(\bar{\Omega})$ be an approximation of a classical solution. For a finite-dimensional space $S_h$, we choose a basis $\phi_1, \dots, \phi_N$ (of ansatz functions) and consider an approximation (2.30). It follows the residual $\rho : \Omega \to \mathbb{R}$

$$\rho := Lu_h - f = \left(\sum_{i=1}^{N} \alpha_i L\phi_i\right) - f.$$

We want to determine the coefficients $\alpha_1, \dots, \alpha_N \in \mathbb{R}$ such that the residual $\rho$ becomes small in some sense. In the method of weighted residuals, a space $T_h$ of test functions with dimension $N$ is chosen. We demand that the residual $\rho$ is orthogonal to the space $T_h$ with respect to the inner product of $L^2$, i.e.,

$$\langle Lu_h - f, v \rangle_{L^2} = 0 \quad \text{for all } \; v \in T_h.$$

Selecting a basis $\{\psi_1, \dots, \psi_N\}$ of $T_h$, this property can be written as

$$\int_\Omega \psi_j(x) \cdot \rho(x) \, \mathrm{d}x = 0 \quad \text{or} \quad \langle \rho, \psi_j \rangle_{L^2} = 0 \qquad \text{for } \; j = 1, \dots, N.$$

This expression can be seen as weighted integrals of the residual $\rho$, where the functions $\psi_j$ represent the weights. It follows the linear system

$$\sum_{i=1}^{N} \alpha_i \langle L\phi_i, \psi_j \rangle_{L^2} = \langle f, \psi_j \rangle_{L^2} \quad \text{for } j = 1, \ldots, N$$

for the unknown coefficients. In the special case $S_h = T_h$ and choosing the same basis in each space, it follows the Galerkin method.

**Remark:** A significant advantage of the Galerkin method is that the matrix of the linear system is symmetric and positive definite for an arbitrary domain $\Omega$. Hence iterative solvers can be applied efficiently. In the finite difference method, see Sect. 2.2, the matrix of the linear system is symmetric and positive definite in case of the unit square ($\Omega = (0,1)^2$). The matrix becomes unsymmetric for other domains like the unit disc, for example.

Concerning the stability of the Ritz-Galerkin method, we obtain the following result.

**Theorem 11 (stability)** *Let $a : H_0^m(\Omega) \times H_0^m(\Omega) \to \mathbb{R}$ be a symmetric, continuous and $H_0^m(\Omega)$-elliptic bilinear form. Let $\ell : H_0^m(\Omega) \to \mathbb{R}$ be a linear, continuous mapping. Then the solution of the Ritz-Galerkin method satisfies*

$$\|u_h\|_{H^m} \leq \tfrac{1}{\alpha} \|\ell\|. \tag{2.32}$$

*independent of the choice $S_h \subset H_0^m(\Omega)$.*

Proof:

Since $\ell$ is continuous, it holds $|\ell(v)| \leq \|\ell\| \cdot \|v\|_{H^m}$. The $H_0^m(\Omega)$-ellipticity yields

$$0 \leq \alpha \|u_h\|_{H^m}^2 \leq a(u_h, u_h) = \langle \ell, u_h \rangle \leq \|\ell\| \cdot \|u_h\|_{H^m}.$$

For $u_h = 0$, the inequality (2.32) is trivial. For $u_h \neq 0$, we divide by $\|u_h\|_{H^m}$ and obtain (2.32). $\square$

The stability implies the Lipschitz-continuous dependence of the approximation on the input data. For example, we consider a perturbation in the right-hand side $f$. It holds $\|\ell\| \leq \|f\|_{L^2}$. Consider two right-hand sides $f_1, f_2$ with corresponding weak solutions $u_1, u_2$. The difference $u_1 - u_2$ is a weak solution for the right-hand side $f_1 - f_2$. The approximations following from the Galerkin method satisfy according to (2.32)

$$\|u_h^1 - u_h^2\|_{H^m} \leq \frac{1}{\alpha}\|f_1 - f_2\|_{L^2}$$

due to the linearity. This estimate is independent of the choice of $S_h$, i.e., it is uniform in $h > 0$ for a discretisation step size to be defined later.

Concerning the quality of the approximation resulting from the Ritz-Galerkin method, the following important theorem holds.

**Theorem 12 (Lemma of Céa)** *Let $H$ be a Hilbert space, $l : H \to \mathbb{R}$ be a linear continuous form and $a : H \times H \to \mathbb{R}$ be a bilinear form, which is symmetric, continuous and $H$-elliptic. Then the function $u$ defined by $a(u, v) = \langle l, v\rangle$ for all $v \in H$ and the approximation $u_h$ of the corresponding Ritz-Galerkin method using some $S_h \subset H$ satisfy the estimate*

$$\|u - u_h\| \leq \frac{C}{\alpha} \inf_{v_h \in S_h} \|u - v_h\|. \tag{2.33}$$

Proof:

It holds

$$a(u, v) = \langle \ell, v\rangle \quad \text{for } \ v \in H, \quad a(u_h, v) = \langle \ell, v\rangle \quad \text{for } \ v \in S_h \subset H.$$

By subtraction, we obtain

$$a(u - u_h, v) = 0 \quad \text{for all } \ v \in S_h.$$

For an arbitrary $v_h \in S_h$, we conclude

$$\begin{aligned} \alpha\|u - u_h\|^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &= a(u - u_h, u - v_h) \\ &\leq C \cdot \|u - u_h\| \cdot \|u - v_h\| \end{aligned}$$

due to $v_h - u_h \in S_h$. Dividing this inequality by $\|u - u_h\| \neq 0$ yields

$$\|u - u_h\| \leq \frac{C}{\alpha}\|u - v_h\|.$$

Since $v_h \in S_h$ is arbitrary, it follows the relation (2.33). □

Theorem 12 implies already the convergence of the Galerkin method provided that

$$\lim_{h \to 0} \inf_{v_h \in S_h} \|u - v_h\| = 0.$$

Hence the subspaces have to be chosen such that the distance to the exact solution decreases. However, this is more a question in approximation theory in our case $H = H_0^1(\Omega)$. We apply the Ritz-Galerkin method in the following. It remains to choose the spaces $S_h$ appropriately.

In a finite difference method, the matrix of a linear system is typically sparse or even a band matrix. Thus the computational effort is significantly lower than in case of a dense matrix with the same size. We want to achieve also a sparse matrix or a band matrix in the Ritz-Galerkin method. We apply spaces $S_h$ consisting of piecewise polynomial functions. However, it turns out that the matrix will be sparse just for specific choices of basis functions.

Let $\operatorname{supp}(\phi) := \overline{\{x \in \Omega : \phi(x) \neq 0\}}$. The bilinear form (2.21) satifies

$$a(\phi, \psi) = 0 \quad \text{if } \mu(\operatorname{supp}(\phi) \cap \operatorname{supp}(\psi)) = 0$$

with the Lebesgue measure $\mu$, since the bilinear form represents an integral in $\Omega$. Hence we will construct a basis such that the supports of the basis functions overlap only rarely. Of course, it should still hold

$$\bigcup_{j=1}^{N} \operatorname{supp}(\phi_j) = \bar{\Omega}$$

for a basis $\{\phi_1, \ldots, \phi_N\}$. The domain $\Omega$ will be decomposed into smaller subdomains for the construction of the space $S_h$ as well as the choice of the basis functions.

**Triangulations**

We consider the two-dimensional case ($n = 2$). Let $\Omega \subset \mathbb{R}^2$ be an open polygonal domain. Hence we can divide the domain $\Omega$ into triangles.

**Definition 12 (triangulation)** *Let $\Omega \subset \mathbb{R}^2$ be a domain with a polygonal boundary. A set $\mathcal{T} = \{T_1, \ldots, T_Q\}$, where the $T_j$ are non-empty closed triangles, is an admissible triangulation if it holds*

*(i)* $\bar{\Omega} = \bigcup_{j=1}^{Q} T_j$,

*(ii)* $\mathrm{int}(T_i) \cap \mathrm{int}(T_j) = \emptyset$ *for* $i \neq j$ (int*: interior),*

*(iii)* $T_i \cap T_j$ *for* $i \neq j$ *is either an empty set or a corner of both triangles or a complete edge of both triangles.*

For each $T \in \mathcal{T}$, we define

$$h_T := \tfrac{1}{2}\mathrm{diam}(T) = \tfrac{1}{2}\max\{\|x - y\|_2 : x, y \in T\}$$

(diam: diameter). For a triangulation $\mathcal{T}$, the (global) step size reads

$$h := \max\{h_T : T \in \mathcal{T}_h\}.$$

Each triangle $T$ contains a (maximal) circle of radius $\rho_T$. Given a family $\mathcal{T}_h$ of triangulations for $0 < h < h_0$, we assume $\max\{h_T : T \in \mathcal{T}_h\} \leq h$. A family $\mathcal{T}_h$ of triangulations is called uniform, if a constant $\kappa > 0$ exists such that $\rho_T \geq \frac{h}{\kappa}$ for all $T$. The family $\mathcal{T}_h$ is called quasi-uniform, if $\rho_T \geq \frac{h_T}{\kappa}$ for each $T$. Remark that it always holds $\rho_T \leq h_T \leq h$. Both properties exclude that the angles of the triangles become arbitrarily small. For uniform triangulations, the size of the triangles is similar for fixed $h$.

Given an arbitray open and bounded domain $\Omega \subset \mathbb{R}^2$, the boundary is approximated by a polygon first. Then the triangulation is applied to the polygonal domain. For $\Omega \subset \mathbb{R}^2$, also quadrangles can be used to decompose the domain. However, triangulations allow for more flexibility.

**Basis functions**

We consider an admissible triangulation $\mathcal{T}_h$ of an open and polygonal domain $\Omega \subset \mathbb{R}^2$. We define finite-dimensional function spaces $S_h$ consisting of all functions $v : \bar{\Omega} \to \mathbb{R}$ satisfying the properties

(i) $v \in C^k(\bar{\Omega})$ for some $k \geq 0$,

(ii) $v|_{\partial\Omega} = 0$,

(iii) $v|_T$ is a polynomial of degree (at most) $l \geq 1$ for each $T \in \mathcal{T}_h$.

Thereby, the choice of the integers $k, l$ is independent of $h > 0$.

Hence piecewise polynomial functions appear. We apply the case $k = 0$ (globally continuous functions) and $l = 1$ (piecewise linear functions). It holds

$$v|_T = \alpha_T + \beta_T x + \gamma_T y \quad \text{for each } \ T \in \mathcal{T}_h$$

with coefficients $\alpha_T, \beta_T, \gamma_T \in \mathbb{R}$.

Let $R = \{(x_i, y_i) : i = 1, \ldots, N\}$ be the set of inner nodes, i.e., the corners of the triangles inside $\Omega$. Let $\partial R = \{(x_i, y_i) : i = N+1, \ldots, N+K\}$ be the set of boundary nodes, i.e., the corners of the triangles on $\partial\Omega$. We define piecewise linear basis function $\phi_i$ via

$$\phi_i(x_j, y_j) = \begin{cases} 1 & \text{if } \ i = j \\ 0 & \text{if } \ i \neq j \end{cases} \tag{2.34}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, N+K$. It holds $\dim(S_h) = N$.

We have to evaluate the bilinear form (2.21), which can be decomposed into

$$\begin{aligned} a(\phi_i, \phi_j) &= \int_\Omega \sum_{k,l=1}^{2} a_{kl} \frac{\partial \phi_i}{\partial x_k} \frac{\partial \phi_j}{\partial x_l} + a_0 \phi_i \phi_j \, \mathrm{d}x \\ &= \sum_{T \in \mathcal{T}_h} \int_T \sum_{k,l=1}^{2} a_{kl} \frac{\partial \phi_i}{\partial x_k} \frac{\partial \phi_j}{\partial x_l} + a_0 \phi_i \phi_j \, \mathrm{d}x \end{aligned}$$

for $i, j = 1, \ldots, N$. Likewise, the information of the right-hand side is evaluated via

$$\langle \ell, \phi_i \rangle = \int_\Omega f(x)\phi_i(x)\ \mathrm{d}x = \sum_{T \in \mathcal{T}_h} \int_T f(x)\phi_i(x)\ \mathrm{d}x$$

for $i = 1, \ldots, N$.

In the case $\Omega \subset \mathbb{R}^n$, the general definition of finite elements following Ciarlet is given now.

**Definition 13 (finite elements)**
*A finite element is a triple $(T, \Pi, \Sigma)$ with the properties:*

*(i) $T \subset \mathbb{R}^n$ is a polyhedron (it follows that $T$ is bounded),*

*(ii) $\Pi \subset C^0(T)$ is a linear space of finite dimension $s$,*

*(iii) $\Sigma$ is a set of $s$ linear independent mappings $\sigma : \Pi \to \mathbb{R}$, which define each $\pi \in \Pi$ uniquely (generalised interpolation).*

Sometimes, just the subdomains $T \subset \Omega$ are called the finite elements. In case of $\Omega \subset \mathbb{R}^2$, a triangulation implies a corresponding set of finite elements, where $T$ is a triangle.

### Benchmark problem

Given a uniform grid in the square $\Omega = (0, 1) \times (0, 1)$, cf. Figure 2, it is straightforward to generate a triangulation, see Figure 6. The defined step size $h$ is not half of the diameter in this case. We consider the Poisson equation $-\Delta u = f$ with homogeneous Dirichlet boundary conditions. The corresponding bilinear form is given in (2.28).

We apply the piecewise linear basis functions (2.34). For $\phi_i$, let $Z = (x_i, y_i)$ be the central node. The neighbouring nodes are labelled as shown in Figure 7 (left). We calculate the stiffness matrix $A_h = (a(\phi_i, \phi_j))$ in the

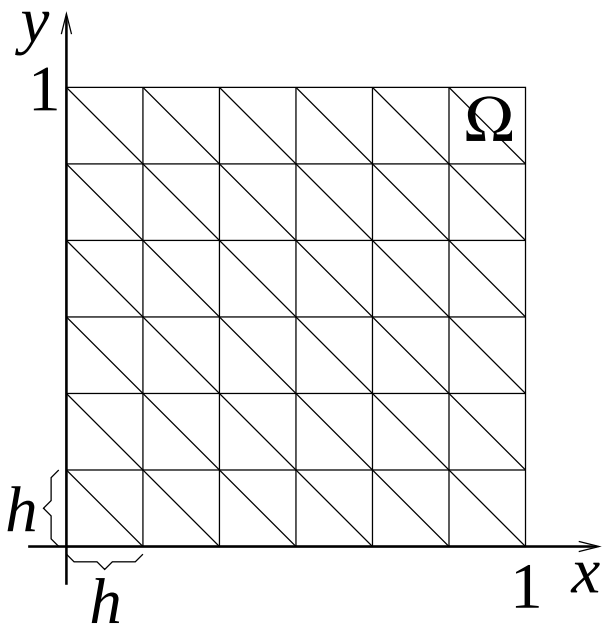


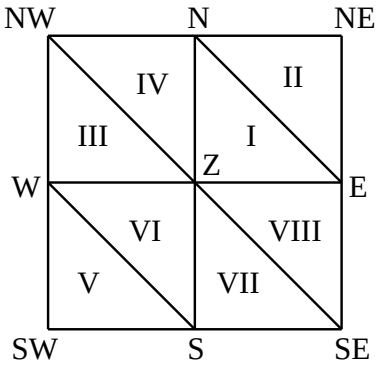

Figure 6: Uniform grid with corresponding triangulation.

Ritz-Galerkin method. Considering (2.28), it follows

$$\begin{aligned} a(\phi_Z,\phi_Z) &= \int_\Omega (\nabla\phi_Z)^2 \,\mathrm{d}x\mathrm{d}y = 2\int_{\mathrm{I,III,IV}} \left(\frac{\partial\phi_Z}{\partial x}\right)^2 + \left(\frac{\partial\phi_Z}{\partial y}\right)^2 \,\mathrm{d}x\mathrm{d}y \\ &= 2\int_{\mathrm{I,III}} \left(\frac{\partial\phi_Z}{\partial x}\right)^2 \,\mathrm{d}x\mathrm{d}y + 2\int_{\mathrm{I,IV}} \left(\frac{\partial\phi_Z}{\partial y}\right)^2 \,\mathrm{d}x\mathrm{d}y \\ &= \frac{2}{h^2}\int_{\mathrm{I,III}} \mathrm{d}x\mathrm{d}y + \frac{2}{h^2}\int_{\mathrm{I,IV}} \mathrm{d}x\mathrm{d}y = \frac{2}{h^2}\cdot 4\cdot\frac{h^2}{2} = 4 \end{aligned}$$

due to the values of the first derivative, see Figure 7 (right). Furthoremore, we obtain

$$\begin{aligned} a(\phi_Z,\phi_N) &= \int_\Omega (\nabla\phi_Z)\cdot(\nabla\phi_N) \,\mathrm{d}x\mathrm{d}y \\ &= \int_{\mathrm{I,IV}} \frac{\partial\phi_Z}{\partial x}\frac{\partial\phi_N}{\partial x} + \frac{\partial\phi_Z}{\partial y}\frac{\partial\phi_N}{\partial y} \,\mathrm{d}x\mathrm{d}y \\ &= \int_{\mathrm{I,IV}} \left(-\frac{1}{h}\right)\frac{1}{h} \,\mathrm{d}x\mathrm{d}y = -\frac{1}{h^2}\int_{\mathrm{I,IV}} \mathrm{d}x\mathrm{d}y \\ &= -\frac{1}{h^2}\cdot 2\cdot\frac{h^2}{2} = -1 \end{aligned}$$




| | I | II | III | IV | V | VI | VII | VIII |
|---|---|---|---|---|---|---|---|---|
| $\frac{\partial \phi_Z}{\partial x}$ | $-h^{-1}$ | 0 | $h^{-1}$ | 0 | 0 | $h^{-1}$ | 0 | $-h^{-1}$ |
| $\frac{\partial \phi_Z}{\partial y}$ | $-h^{-1}$ | 0 | 0 | $-h^{-1}$ | 0 | $h^{-1}$ | $h^{-1}$ | 0 |

Figure 7: Basic cell in benchmark problem.

and due to a symmetry also $a(\phi_Z, \phi_S) = a(\phi_Z, \phi_W) = a(\phi_Z, \phi_E) = -1$. It is straightforward to verify

$$a(\phi_Z, \phi_{NW}) = a(\phi_Z, \phi_{NE}) = a(\phi_Z, \phi_{SW}) = a(\phi_Z, \phi_{SE}) = 0$$

by observing the supports of the basis functions.

For the right-hand side, we apply an approximation

$$\langle \ell, \phi_i \rangle = \int_\Omega f(x,y)\phi_i(x,y)\,\mathrm{d}x\mathrm{d}y \doteq h^2 f(x_i, y_i),$$

since it holds $f(x_j, y_j)\phi_i(x_j, y_j) = f(x_i, y_i)\delta_{ij}$ for all $j$ and

$$\int_\Omega \phi_i(x,y)\,\mathrm{d}x\mathrm{d}y = h^2.$$

It follows just the five-point star from the finite difference method, cf. (2.9). Each finite difference method corresponds to some finite element method. However, not each finite element method is equivalent to a finite difference method. Hence finite element techniques allow for more flexibility.

### Computation of stiffness matrix

We outline the efficient computation of the stiffness matrix in the Ritz-Galerkin method, where a general admissible triangulation is considered, see Def. 12. The structure of $A_h = (a(\phi_i, \phi_j)) \in \mathbb{R}^{N\times N}$ suggests to use

a loop over the inner nodes $i = 1, \ldots, N$ to evaluate the bilinear form (node-oriented form). However, it can be shown that this procedure is inefficient. Alternatively, the loop is arranged over the triangles (element-oriented form).

We consider a polygonal domain $\Omega \subset \mathbb{R}^2$ with an arbitrary admissible triangulation $\mathcal{T}_h = \{T_1, \ldots, T_Q\}$. The Poisson equation $-\Delta u = f$ with homogeneous Dirichlet boundary conditions is used as benchmark again. The corresponding bilinear form, cf. (2.28),

$$\begin{aligned} a_{\mu\nu} \;:=\; a(\phi_\mu, \phi_\nu) \;&=\; \int_\Omega \frac{\partial \phi_\mu}{\partial x}\frac{\partial \phi_\nu}{\partial x} + \frac{\partial \phi_\mu}{\partial y}\frac{\partial \phi_\nu}{\partial y}\,\mathrm{d}x\mathrm{d}y \\ &=\; \sum_{q=1}^{Q} \int_{T_q} \frac{\partial \phi_\mu}{\partial x}\frac{\partial \phi_\nu}{\partial x} + \frac{\partial \phi_\mu}{\partial y}\frac{\partial \phi_\nu}{\partial y}\,\mathrm{d}x\mathrm{d}y \end{aligned}$$

has to be evaluated for $\mu, \nu = 1, \ldots, N$. We define

$$a^q_{\mu\nu} := \int_{T_q} \frac{\partial \phi_\mu}{\partial x}\frac{\partial \phi_\nu}{\partial x} + \frac{\partial \phi_\mu}{\partial y}\frac{\partial \phi_\nu}{\partial y}\,\mathrm{d}x\mathrm{d}y. \tag{2.35}$$

It follows for $A_h \in \mathbb{R}^{N\times N}$.

$$a_{\mu\nu} = \sum_{q=1}^{Q} a^q_{\mu\nu} \quad \text{and} \quad A_h = \sum_{q=1}^{Q} A^q_h \quad \text{with} \;\; A^q_h := (a^q_{\mu\nu}).$$

Let $i, j, k$ be the index of the corners of the triangle $T_q$. Hence just $\phi_i, \phi_j, \phi_k$ are non-zero in $T_q$ and give a contribution to the integral over $T_q$. We obtain the structure

$$A^q_h = \begin{pmatrix} & \vdots & & \vdots & & \vdots & \\ \cdots & a^q_{ii} & \cdots & a^q_{ij} & \cdots & a^q_{ik} & \cdots \\ & \vdots & & \vdots & & \vdots & \\ \cdots & a^q_{ji} & \cdots & a^q_{jj} & \cdots & a^q_{jk} & \cdots \\ & \vdots & & \vdots & & \vdots & \\ \cdots & a^q_{ki} & \cdots & a^q_{kj} & \cdots & a^q_{kk} & \cdots \\ & \vdots & & \vdots & & \vdots & \end{pmatrix} \in \mathbb{R}^{M\times M},$$

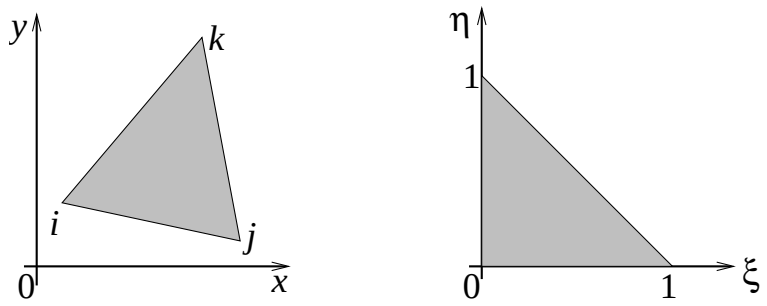



Figure 8: Transformation to reference triangle.

where (at most) nine entries are non-zero. The matrix can be written in condensed form

$$\tilde{A}_h^q = \begin{pmatrix} a_{ii}^q & a_{ij}^q & a_{ik}^q \\ a_{ji}^q & a_{jj}^q & a_{jk}^q \\ a_{ki}^q & a_{kj}^q & a_{kk}^q \end{pmatrix} \in \mathbb{R}^{3\times 3}. \tag{2.36}$$

To compute (2.35), we transform each triangle $T_q$ to a reference triangle $\hat{T} = \{(\xi, \eta) \in \mathbb{R}^2 : 0 \le \xi, \eta,\ \xi + \eta \le 1\}$, see Figure 8. It follows the formula

$$\tilde{A}_h^q = \frac{1}{4|T_q|} E_q E_q^\top \quad \text{with} \quad E_q := \begin{pmatrix} y_j - y_k & x_k - x_j \\ y_k - y_i & x_i - x_k \\ y_i - y_j & x_j - x_i \end{pmatrix},$$

where $|T_q|$ represents the area of the triangle. Recall that the indices $i, j, k$ depend on $q$. Thus the entries of $A_h$ follow directly from the coordinates of the corners of the triangles.

If one corner of $T_q$ does not belong to the inner nodes but to the boundary, say index $i$, then a corresponding basis function $\phi_i$ is not defined. It follows that the first row as well as the first column in (2.36) are omitted. Accordingly, two rows and two columns are deleted if two corners are situated on the boundary. This strategy is in agreement to the homogeneous boundary conditions.

The matrix $A_h \in \mathbb{R}^{N\times N}$ includes $N^2$ entries. Since $A_h$ is the sum of $A_h^q$ for $q = 1, \ldots, Q$, we obtain a rough estimate of the non-zero entries in $A_h$: at most 9Q entries are non-zero.

### Approximations of higher order

In the previous subsections, we applied piecewise linear polynomials corresponding to a triangulation $\mathcal{T}_h = \{T_1, \ldots, T_Q\}$ of a polygonal domain $\Omega$. We are able to construct piecewise polynomials of higher degrees. Let $\mathcal{P}_l$ be the set of all polynomials up to degree $l$, i.e.,

$$\mathcal{P}_l := \left\{ p(x,y) = \sum_{i,j \geq 0, i+j \leq l} c_{ij} x^i y^j \right\}.$$

It holds $\dim(\mathcal{P}_l) = \frac{(l+1)(l+2)}{2}$, which is also the number of coefficients $c_{ij}$.

On each triangle $T \in \mathcal{T}_h$, we choose $\frac{(l+1)(l+2)}{2}$ points $z_s = (x_s, y_s)$ for an interpolation. Figure 9 illustrates the construction of the points within the reference triangle $\hat{T}$. It follows a unique interpolation operator

$$I_T : C^0(T) \to \mathcal{P}_l, \quad (I_T u)(z_s) = u(z_s) \text{ for } s = 1, \ldots, \tfrac{(l+1)(l+2)}{2}.$$

We obtain a global interpolation operator

$$I_h : C^0(\bar{\Omega}) \to C^0(\bar{\Omega}), \quad I_h|_T = I_T.$$

Hence $I_h u$ is a piecewise polynomial of degree up to $l$ for $u \in C^0(\bar{\Omega})$. Moreover, $I_h u$ is a globally continuous function. The resticition of the polynomial $I_T u$ to the edge of the triangle $T$ represents a univariate polynomial of (at most) degree $l$. Since each edge includes $l+1$ nodes, the univariate polynomials on the boundary of two neighbouring triangles coincide.

We want to apply the Sobolev spaces $H^m(\Omega)$. The theorem of Sobolev implies $H^m(\Omega) \subset C^0(\Omega)$ for $m \geq 2$, i.e., each $u \in H^m(\Omega)$ exhibits a continuous representative. It follows that the interpolation operator can be extended to an operator $I_h : H_0^m(\Omega) \to C^0(\bar{\Omega})$ provided that $m \geq 2$. We demand $(I_h u)(z_s) = 0$ for a node $z_s \in \partial\Omega$ due to the homogeneous boundary conditions.

If the degree $l$ of the piecewise polynomial functions is sufficiently large, then also global interpolants $I_h u \in C^k(\bar{\Omega})$ for $k \geq 1$ can be defined. However, the construction becomes much more complicated. The choice of the

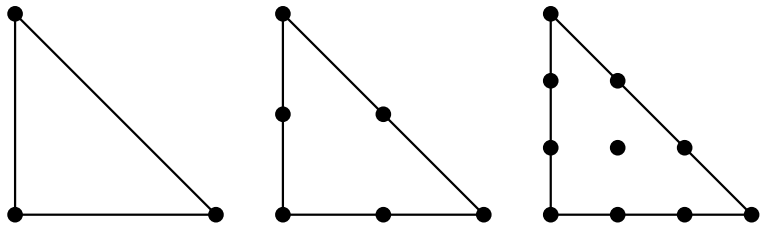


Figure 9: Nodes for linear (left), quadratic (center) and cubic (right) polynomial interpolation in the reference triangle.

interpolation is related to the selection of the finite-dimensional spaces $S_h$ in the Ritz-Galerkin method.

**Remark:** On a triangulation, we already obtain functions globally $C^k(\bar{\Omega})$ for arbitrary $k \geq 0$ provided that the degree $l$ of the local polynomials is sufficiently large. Thus we do not require more complicated subdomains of $\Omega$ to achieve an approximation of higher order.

### Convergence of finite element method

We consider the finite element method for the general problem $Lu = f$ with a uniformly elliptic differential operator and homogeneous Dirichlet boundary conditions in a polygonal domain $\Omega \subset \mathbb{R}^2$. Let an admissible triangulation $\mathcal{T}_h = \{T_1, \ldots, T_Q\}$ be given. The convergence of the method follows from Theorem 12, where we have to discuss approximations resulting from interpolation schemes.

A finite element method based on the triangulation $\mathcal{T}_h$ applies a space

$$S_h := \left\{ v \in C^0(\bar{\Omega}) : v|_{\partial\Omega} = 0, v|_T \in \mathcal{P}_l \text{ for each } T \in \mathcal{T}_h \right\}. \tag{2.37}$$

We apply the global interpolation operator

$$I_h : H_0^m(\Omega) \to C^0(\bar{\Omega}), \quad I_h u|_T \in \mathcal{P}_l \text{ for each } T \in \mathcal{T}_h$$

assuming $m \geq 2$, which has been introduced in the previous subsection. It holds $I_h u \in S_h$ for $u \in H_0^m(\Omega)$.

We define the norm

$$\|v\|_{m,h} := \sqrt{\sum_{T \in \mathcal{T}_h} \|v\|^2_{H^m(T)}}$$

for functions $v : \Omega \to \mathbb{R}$ satisfying $v|_T \in H^m(T)$ for each $T \in \mathcal{T}_h$. The functions $v_h \in S_h$ from (2.37) exhibit this property. Remark that $v \in C^k(\bar{\Omega})$ implies just $v \in H^{k+1}(\Omega)$ even for piecewise polynomial functions $v$. It holds $\|v\|_{m,h} = \|v\|_{H^m}$ for $v \in H^m(\Omega)$. However, the subspace (2.37) fulfills just $S_h \subset H^1_0(\Omega)$.

The following theorem holds for general functions, i.e., they are not necessarily the solution of some PDE.

**Theorem 13** *Let $t \geq 2$ and $\mathcal{T}_h$ be a quasi-uniform triangulation of $\Omega$. The corresponding interpolation by piecewise polynomials of degree $t-1$ satisfies*

$$\|u - I_h u\|_{m,h} \leq c \cdot h^{t-m} \cdot |u|_{H^t} \quad \textit{for } u \in H^t(\Omega)$$

*and $0 \leq m \leq t$. The constant $c \geq 0$ depends on $\Omega$, the constant $\kappa$ of the quasi-uniform triangulation $\mathcal{T}_h$ and the integer $t$.*

For the proof, see D. Braess: Finite Elements.

Since the weak solution of our elliptic PDE is defined in $H^1_0(\Omega)$, we apply the case $m = 1$ only. Thereby, $I_h u \in H^1_0(\Omega)$ is guaranteed. We assume that the unique weak solution satisfies $u \in H^t_0(\Omega) \subset H^1_0(\Omega)$ for some $t \geq 2$. Now we achieve the convergence by means of Theorem 12. Due to $I_h u \in S_h$, it holds

$$\inf_{v_h \in S_h} \|u - v_h\|_{H^1} \leq \|u - I_h u\|_{H^1} \leq c \cdot h^{t-1} \cdot |u|_{H^t}.$$

We conclude the convergence of order $p \geq 1$ for the approximation $u_h \in S_h$ resulting from the Ritz-Galerkin approach in the norm of $H^1(\Omega)$ due to

$$\|u - u_h\|_{H^1} \leq K \cdot h^p \cdot |u|_{H^{p+1}} \quad \text{for } u \in H^{p+1}(\Omega) \tag{2.38}$$

with $K := \frac{Cc}{\alpha}$ depending on $p$. For $t = 2$, piecewise linear polynomials are applied. For $t > 2$, we can achieve higher orders of convergence just by

choosing polynomials of higher degrees in each triangle. The approximation $u_h$ is still just continuous globally.

Due to (2.38), we require at least $u \in H^2(\Omega)$ to obtain convergence of order $p \geq 1$. It can be shown that the weak solution satisfies $u \in H^2(\Omega)$ provided that $f \in L^2(\Omega)$ holds and the domain $\Omega$ satisfies some basic assumptions.

Theorem 13 also yields an estimate corresponding to the norm of $L^2(\Omega)$, i.e., in the case of $m = 0$. We expect a convergence of order $t$ in the norm of $L^2(\Omega)$. Unfortunately, Theorem 12 cannot be applied in this case, since the underlying bilinear form is not continuous with respect to the norm of $L^2(\Omega)$. Nevertheless, the strategy of Aubin and Nitsche yields the estimates

$$\|u - u_h\|_{L^2} \leq \tilde{K} \cdot h^{p+1} \cdot |u|_{H^{p+1}} \quad \text{for } u \in H^{p+1}(\Omega)$$

with constants $\tilde{K} > 0$ depending on $p$.

For some problems, a uniform convergence can be shown like

$$\sup_{x \in \Omega} |u(x) - u_h(x)| \leq c \cdot h \cdot \|f\|_{L^2}$$

with a constant $c > 0$. Such estimates correspond to the space $L^\infty(\Omega)$.

We have shown the convergence in a Sobolev norm or the $L^2$-norm, respectively. Further estimates can be construced in the energy norm $\|\cdot\|_a$.

# Chapter 3

# Parabolic PDEs

Now we consider parabolic PDEs, which are time-dependent problems. The heat equation represents the benchmark for this class of PDEs. Numerical methods for initial-boundary value problems of parabolic PDEs will be derived and analysed.

## 3.1 Initial-boundary value problems

Time-dependent parabolic PDEs often exhibit the form

$$\frac{\partial u}{\partial t} + Lu = f(x_1, \ldots, x_n)$$

with solution $u : D \times [t_0, t_{\text{end}}] \to \mathbb{R}$ using some domain $D \subseteq \mathbb{R}^n$ in space. The linear differential operator $L$ includes second-order derivatives of $u$ with respect to space (no derivatives in time) and is often of elliptic type. We restrict to one space dimension ($n = 1$) in this chapter.

The heat equation reads

$$\frac{\partial v}{\partial t} = \lambda(x)\frac{\partial^2 v}{\partial x^2} \tag{3.1}$$

with a coefficient function $\lambda : D \to \mathbb{R}$ ($D \subseteq \mathbb{R}$) and $\lambda(x) > 0$ for each $x$. Without loss of generality, we choose $\lambda(x) \equiv 1$, i.e.,

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}. \tag{3.2}$$

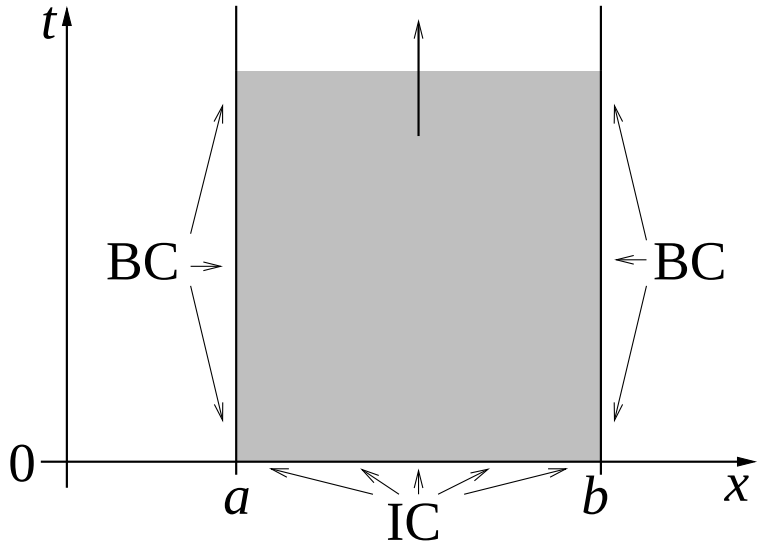



Figure 10: Initial-boundary value problem.

Given a solution $u$ of (3.2), we obtain a solution of (3.1) for constant $\lambda$ via the transformation $v(x,t) = u(x,\lambda t)$.

We choose a finite interval $[a,b]$ in space ($a < b$). Boundary conditions (BCs) will be specified at $x = a$ and $x = b$. Initial conditions (ICs) will be given in the form

$$u(x,t_0) = u_0(x) \quad \text{for} \;\; x \in [a,b] \tag{3.3}$$

with a predetermined function $u_0 : [a,b] \to \mathbb{R}$. Without loss of generality, we define $t_0 := 0$. The initial-boundary value problem is sketched in Figure 10.

We distinguish three different types of boundary value problems:

(i) Boundary conditions of *Dirichlet type* read

$$u(a,t) = \alpha(t), \quad u(b,t) = \beta(t) \quad \text{for all} \;\; t \geq 0 \tag{3.4}$$

with predetermined functions $\alpha, \beta : [0,\infty) \to \mathbb{R}$.

(ii) Boundary conditions of *von-Neumann type* demand

$$\left.\frac{\partial u}{\partial x}\right|_{x=a} = \alpha(t), \quad \left.\frac{\partial u}{\partial x}\right|_{x=b} = \beta(t) \quad \text{for all} \;\; t \geq 0 \tag{3.5}$$

with predetermined functions $\alpha, \beta : [0,\infty) \to \mathbb{R}$.

(iii) Boundary problems of *Robin type*, i.e., a mixed problem of the types (i) and (ii), namely

$$\gamma_a(t)u(a,t) + \delta_a(t)\left.\frac{\partial u}{\partial x}\right|_{x=a} = \alpha(t), \quad \gamma_b(t)u(b,t) + \delta_b(t)\left.\frac{\partial u}{\partial x}\right|_{x=b} = \beta(t)$$

for all $t \geq 0$ with predetermined functions $\alpha, \beta, \gamma_a, \gamma_b, \delta_a, \delta_b$.

The initial values (3.3) have to be compatible with the boundary conditions. For example, $u_0(a) = \alpha(0)$ and $u_0(b) = \beta(0)$ is required in case of Dirichlet boundary conditions.

Let $u$ be a solution of (3.2) according to homogeneous Dirichlet boundary conditions ($\alpha, \beta \equiv 0$). The function

$$\hat{u}(x) := \tfrac{b-x}{b-a}\alpha + \tfrac{x-a}{b-a}\beta$$

satisfies the inhomogeneous boundary conditions (3.4) for constant values $\alpha, \beta \neq 0$. It follows that $v := u + \hat{u}$ is a solution of (3.2), which fulfills the inhomogeneous Dirichlet problem. Hence we consider homogeneous Dirichlet conditions without loss of generality.

We solve the heat equation (3.2) with homogeneous Dirichlet boundary conditions analytically for $a = 0, b = 1$. We assume a separation

$$u(x,t) = \phi(t)\psi(x).$$

Inserting this relation into (3.2) gives

$$\phi'(t)\psi(x) = \phi(t)\psi''(x) \quad \Leftrightarrow \quad \frac{\phi'(t)}{\phi(t)} = \frac{\psi''(x)}{\psi(x)} =: \kappa.$$

Thereby, $\kappa \in \mathbb{R}$ represents the separation constant. Solving the two ordinary differential equations

$$\phi'(t) = \kappa\phi(t), \qquad \psi''(x) = \kappa\psi(x)$$

yields

$$\phi(t) = C\mathrm{e}^{\kappa t}, \qquad \psi(x) = A\mathrm{e}^{\sqrt{\kappa}x} + B\mathrm{e}^{-\sqrt{\kappa}x}.$$

We obtain the general solution

$$u(x,t) = \mathrm{e}^{\kappa t}\left[\tilde{A}\mathrm{e}^{\sqrt{\kappa}x} + \tilde{B}\mathrm{e}^{-\sqrt{\kappa}x}\right]$$

with arbitrary constants $\tilde{A}, \tilde{B} \in \mathbb{C}$. The homogeneous boundary conditions are satisfied if and only if

$$\kappa = -k^2\pi^2 \quad \text{for } \; k = 1, 2, 3, \dots \; .$$

It follows the family of solutions

$$v_k(x,t) = \tilde{A}_k \mathrm{e}^{-k^2\pi^2 t}\sin(k\pi x)$$

for $k \in \mathbb{N}$ with new coefficients $\tilde{A}_k \in \mathbb{R}$. We use these solutions to construct a single solution satisfying the predetermined initial conditions (3.3). It holds $u_0(0) = u_0(1) = 0$ due to the homogeneous boundary conditions. We can extend the function $u_0$ to an odd function $\hat{u} : [-1,1] \to \mathbb{R}$ by the definition $\hat{u}(x) = u_0(x)$ for $x \geq 0$ and $\hat{u}(x) = -u_0(-x)$ for $x < 0$. Assuming $u_0 \in L^2([0,1])$, it exists the Fourier expansion

$$u_0(x) = \sum_{k=1}^{\infty} a_k \sin(k\pi x).$$

It follows $\tilde{A}_k = a_k$. Thus the solution of the initial-boundary value problem reads

$$u(x,t) = \sum_{k=1}^{\infty} a_k \mathrm{e}^{-k^2\pi^2 t}\sin(k\pi x). \tag{3.6}$$

However, to evaluate the formula (3.6), the series has to be truncated and the Fourier coefficients $a_k$ have to be computed numerically.

The formula (3.6) also characterises the condition of the initial-boundary value problem. Let $u_0, \tilde{u}_0 \in L^2([0,1])$ be two initial conditions with corresponding Fourier coefficients $a_k$ and $\tilde{a}_k$, respectively. The resulting solutions satisfy

$$u(x,t) - \tilde{u}(x,t) = \sum_{k=1}^{\infty}(a_k - \tilde{a}_k)\mathrm{e}^{-k^2\pi^2 t}\sin(k\pi x).$$

Hence we obtain

$$|u(x,t) - \tilde{u}(x,t)| \leq \sum_{k=1}^{\infty} |a_k - \tilde{a}_k| \cdot \mathrm{e}^{-k^2\pi^2 t}.$$

The Cauchy-Schwarz inequality in $\ell^2$ and Parseval's theorem yield

$$\begin{aligned} \sum_{k=1}^{\infty} |a_k - \tilde{a}_k| \cdot \mathrm{e}^{-k^2\pi^2 t} &\leq \sqrt{\sum_{k=1}^{\infty} |a_k - \tilde{a}_k|^2} \sqrt{\sum_{k=1}^{\infty} |\mathrm{e}^{-k^2\pi^2 t}|^2} \\ &= \|u_0 - \tilde{u}_0\|_{L^2([0,1])} \sqrt{\sum_{k=1}^{\infty} \mathrm{e}^{-2k^2\pi^2 t}}. \end{aligned}$$

Thereby, we employ that the extensions $\hat{u}, \hat{\tilde{u}}$ of $u_0, \tilde{u}_0$ exhibit the period 2 in Parseval's theorem and that $\|\hat{u} - \hat{\tilde{u}}\|^2_{L^2([-1,1])} = 2\|u_0 - \tilde{u}_0\|^2_{L^2([0,1])}$ due to the symmetry.

We apply the formula of the limit of a geometric series

$$\sum_{k=1}^{\infty} \mathrm{e}^{-2k^2\pi^2 t} = \sum_{k=1}^{\infty} \left(\mathrm{e}^{-2\pi^2 t}\right)^{k^2} < \frac{1}{1 - \mathrm{e}^{-2\pi^2 t}} - 1 = \frac{\mathrm{e}^{-2\pi^2 t}}{1 - \mathrm{e}^{-2\pi^2 t}}.$$

It follows

$$|u(x,t) - \tilde{u}(x,t)| \leq \|u_0 - \tilde{u}_0\|_{L^2([0,1])} \frac{\mathrm{e}^{-\pi^2 t}}{\sqrt{1 - \mathrm{e}^{-2\pi^2 t}}}$$

for all $x \in [0,1]$ and $t > 0$. Hence differences in the initial values are damped out exponentially in time. The condition of this initial-boundary value problem is excellent. Vice versa, an initial-boundary value problem backwards in time (from $t_0 = 0$ to some $t_{\mathrm{end}} < 0$) is drastically ill-conditioned, since small differences are amplified exponentially.

For the heat equation (3.1), the condition of initial-boundary value problems depends on the constant $\lambda \in \mathbb{R}\backslash\{0\}$ as follows:

| | forward in time | backward in time |
|---|---|---|
| $\lambda > 0$: | well-conditioned | ill-conditioned |
| $\lambda < 0$: | ill-conditioned | well-conditioned |

Initial value problems backwards in time are also called final value problems (the values at the earlier time $t_{\text{end}} < 0$ are unknown, whereas the state at the final time $t_0 = 0$ is given).

We achieve a solution of an initial value problem in the case $x \in (-\infty, +\infty)$, where no boundary appears. It follows

$$u(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{(x-y)^2}{4t}} u_0(y)\, \mathrm{d}y. \tag{3.7}$$

The integrals exist for a bounded measurable function $u_0$ or in the case of $u_0 \in L^2(\mathbb{R})$, for example. Otherwise, integrability conditions have to be imposed. The formula (3.7) cannot be evaluated at $t = 0$. The initial conditions are satisfied in the sense

$$\lim_{t \to 0+} u(x,t) = u_0(x) \qquad \text{for each } \ x \in \mathbb{R}.$$

Moreover, this convergence is uniform in compact domains $D \subset \mathbb{R}$.

Let $u_0$ be continuous, $u_0 \geq 0$ and $u_0 \not\equiv 0$. Even if $u_0$ exhibits a compact support, it follows $u(x,t) > 0$ for all $x \in \mathbb{R}$ and each $t > 0$. Hence the transport of information proceeds with infinite speed. This also holds in case of initial-boundary value problems within a finite domain $x \in [a,b]$.

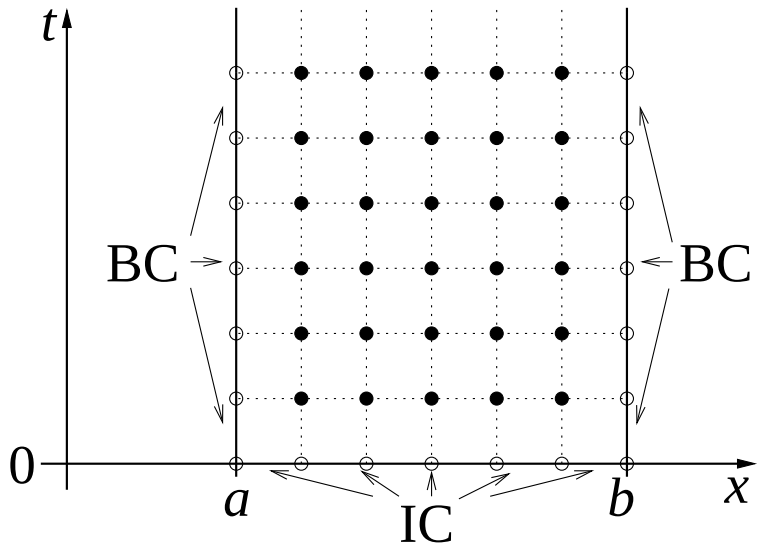



Figure 11: Grid in finite difference method.

### 3.2 Finite difference methods

We want to apply a finite difference method to solve the initial-boundary value problem of the heat equation (3.2) introduced in Sect. 3.1. A grid is constructed in the $(x,t)$-domain for $x \in [a,b]$ and $t \in [0,T]$, see Figure 11. Without loss of generality, we assume $x \in [0,1]$. The grid points are defined by

$$x_j := jh \quad \text{for} \quad j = 0, 1, \ldots, M-1, M, \quad h := \tfrac{1}{M},$$
$$t_n := nk \quad \text{for} \quad n = 0, 1, \ldots, N-1, N, \quad k := \tfrac{T}{N}.$$

The corresponding step sizes are $h = \Delta x$ and $k = \Delta t$ in space and time, respectively. Let $u_j^n := u(x_j, t_n)$ be the values of the exact solution and $U_j^n$ the corresponding approximations in the grid points. We replace the partial derivatives in the heat equation (3.2) by difference formulas now.

### Classical explicit method

We substitute the time derivative by the common difference formula of first order and the space derivative by the symmetric difference formula of second order, i.e.,

$$\begin{aligned} u_t(x_j,t_n) &= \tfrac{1}{k}(u(x_j,t_{n+1})-u(x_j,t_n)) + \tfrac{k}{2}u_{tt}(x_j,t_n+\vartheta k) \\ u_{xx}(x_j,t_n) &= \tfrac{1}{h^2}(u(x_{j-1},t_n)-2u(x_j,t_n)+u(x_{j+1},t_n)) \\ &\quad + \tfrac{h^2}{12}u_{xxxx}(x_j+\theta h,t_n) \end{aligned}$$

with intermediate values $\vartheta \in (0,1)$, $\theta \in (-1,1)$. Thereby, we assume that $u$ is sufficiently smooth. The heat equation yields $u_t(x_j,t_n) = u_{xx}(x_j,t_n)$. It follows

$$\tfrac{1}{k}(u_j^{n+1}-u_j^n) + \mathcal{O}(k) = \tfrac{1}{h^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n) + \mathcal{O}(h^2).$$

Thus the finite difference method reads

$$\begin{aligned} \tfrac{1}{k}(U_{j+1}^n - U_j^n) &= \tfrac{1}{h^2}(U_{j-1}^n - 2U_j^n + U_{j+1}^n), \\ U_j^{n+1} &= U_j^n + \tfrac{k}{h^2}(U_{j-1}^n - 2U_j^n + U_{j+1}^n). \end{aligned}$$

We define the ratio $r := \frac{k}{h^2}$. The formula of the technique becomes

$$U_j^{n+1} = rU_{j-1}^n + (1-2r)U_j^n + rU_{j+1}^n \tag{3.8}$$

for $j = 1,\ldots,M-1$. The scheme is an explicit (one-stage) method. The time layers can be calculated subsequently. The initial values follow from (3.3), i.e.,

$$U_j^0 = u_0(x_j) \quad \text{for } j = 0,1,\ldots,M.$$

In the subsequent layers, the boundary conditions have to be included. Dirichlet boundary conditions yield

$$U_0^n = \alpha(t_n), \quad U_M^n = \beta(t_n) \quad \text{for each } n.$$

Von-Neumann boundary conditions will be discussed later.

The local discretisation error reads

$$\tau(k,h) := \tfrac{k}{2}u_{tt}(x_j,t_n+\vartheta k) - \tfrac{h^2}{12}u_{xxxx}(x_j+\theta h,t_n).$$

We assume that $u_{tt}$ and $u_{xxxx}$ exist and are continuous on $[0,1] \times [0,T]$. Let

$$C_1 := \max_{x\in[0,1],t\in[0,T]} |u_{tt}(x,t)|\,, \quad C_2 := \max_{x\in[0,1],t\in[0,T]} |u_{xxxx}(x,t)|\,.$$

It follows

$$|\tau(k,h)| \le (k+h^2)(\max\{\tfrac{1}{2}C_1, \tfrac{1}{12}C_2\}) =: C(k+h^2) \tag{3.9}$$

uniformly for all grid points $(x_j, t_n)$ in $[0,1] \times [0,T]$. Hence the finite difference method is consistent. For $k, h \to 0$, the local discretisation error converges uniformly to zero. We obtain consistency of order one in time and consistency of order two in space.

### Classical implicit method

Now the same difference formulas are applied with respect to the point $(x_j, t_{n+1})$ and the discretisation in time is done backwards, i.e.,

$$\begin{aligned} u_t(x_j,t_{n+1}) &= \tfrac{1}{k}(u(x_j,t_{n+1}) - u(x_j,t_n)) + \tfrac{k}{2}u_{tt}(x_j, t_n + \vartheta k) \\ u_{xx}(x_j,t_{n+1}) &= \tfrac{1}{h^2}(u(x_{j-1},t_{n+1}) - 2u(x_j,t_{n+1}) + u(x_{j+1},t_{n+1})) \\ &\quad + \tfrac{h^2}{12}u_{xxxx}(x_j + \theta h, t_{n+1}) \end{aligned}$$

with intermediate values $\vartheta \in (-1,0)$, $\theta \in (-1,1)$. The heat equation yields $u_t(x_j,t_{n+1}) = u_{xx}(x_j,t_{n+1})$. We obtain

$$\tfrac{1}{k}(U_j^{n+1} - U_j^n) = \tfrac{1}{h^2}(U_{j-1}^{n+1} - 2U_j^{n+1} + U_{j+1}^{n+1}).$$

It follows the method (using $r := \frac{k}{h^2}$ again)

$$-rU_{j-1}^{n+1} + (1+2r)U_j^{n+1} - rU_{j+1}^{n+1} = U_j^n \tag{3.10}$$

for $j = 1, \ldots, M-1$. The scheme represents an implicit (one-stage) method. To compute the approximations, a linear system has to be solved in each time step. The corresponding matrix reads

$$B := r \begin{pmatrix} 2+\frac{1}{r} & -1 & & & \\ -1 & 2+\frac{1}{r} & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+\frac{1}{r} & -1 \\ & & & -1 & 2+\frac{1}{r} \end{pmatrix} \in \mathbb{R}^{(M-1)\times(M-1)}. \tag{3.11}$$

The matrix is symmetric and tridiagonal. Moreover, the matrix is strict diagonal dominant. An $LU$-decomposition can be done without pivoting, where the computational effort is $\sim M$.

We arrange the approximations in the vector

$$U^n := (U_1^n, U_2^n, \ldots, U_{M-2}^n, U_{M-1}^n)^\top \in \mathbb{R}^{M-1}.$$

Inhomogeneous Dirichlet boundary conditions have to be included in the right-hand side via

$$b^n := \left(rU_0^{n+1}, 0, \ldots, 0, rU_M^{n+1}\right)^\top \in \mathbb{R}^{M-1}.$$

It follows the linear system $BU^{n+1} = U^n + b^n$. For homogeneous Dirichlet boundary conditions, we obtain simply $BU^{n+1} = U^n$.

### Leapfrog method

We want to achieve methods of higher order in time now. We apply the symmetric difference formula of second order for the first-order time derivative, i.e.,

$$\begin{aligned} u_t(x_j, t_n) &= \tfrac{1}{2k}(u(x_j, t_{n+1}) - u(x_j, t_{n-1})) + \mathcal{O}(k^2) \\ u_{xx}(x_j, t_n) &= \tfrac{1}{h^2}(u(x_{j-1}, t_n) - 2u(x_j, t_n) + u(x_{j+1}, t_n)) + \mathcal{O}(h^2). \end{aligned}$$

Due to $u_t(x_j, t_n) = u_{xx}(x_j, t_n)$, it follows the scheme

$$U_j^{n+1} = U_j^{n-1} + \tfrac{2k}{h^2}(U_{j-1}^n - 2U_j^n + U_{j+1}^n)$$

and with $r := \frac{k}{h^2}$

$$U_j^{n+1} = U_j^{n-1} + 2r(U_{j-1}^n + U_{j+1}^n) - 4rU_j^n \tag{3.12}$$

for $j = 1, \ldots, M-1$. We obtain an explicit (two-stage) method, which is called the leapfrog method. The scheme is consistent of order two in both time and space. However, the leapfrog method is unstable for all $r > 0$ as we will show in the next section. Thus this technique is useless in practice.

**Crank-Nicolson method**

We achieve a one-stage scheme of second order in both time and space via the following construction using $t_{n+\frac{1}{2}} := t_n + \frac{k}{2}$

$$\begin{aligned} u_t(x_j, t_{n+\frac{1}{2}}) &= \tfrac{1}{k}(u(x_j, t_{n+1}) - u(x_j, t_n)) + \mathcal{O}(k^2) \\ u_{xx}(x_j, t_{n+\frac{1}{2}}) &= \tfrac{1}{2}(u_{xx}(x_j, t_n) + u_{xx}(x_j, t_{n+1})) + \mathcal{O}(k^2) \\ &= \tfrac{1}{2h^2}(u(x_{j-1}, t_n) - 2u(x_j, t_n) + u(x_{j+1}, t_n)) + \mathcal{O}(h^2) \\ &\quad + \tfrac{1}{2h^2}(u(x_{j-1}, t_{n+1}) - 2u(x_j, t_{n+1}) + u(x_{j+1}, t_{n+1})) + \mathcal{O}(h^2) \\ &\quad + \mathcal{O}(k^2). \end{aligned}$$

We can see this discretisation as an averaging of the symmetric difference formula in space. The heat equation $u_t(x_j, t_{n+\frac{1}{2}}) = u_{xx}(x_j, t_{n+\frac{1}{2}})$ implies with $r := \frac{k}{h^2}$

$$-rU_{j-1}^{n+1} + 2(1+r)U_j^{n+1} - rU_{j+1}^{n+1} = rU_{j-1}^n + 2(1-r)U_j^n + rU_{j+1}^n \qquad (3.13)$$

for $j = 1, \ldots, M-1$. The technique represents an implicit (one-stage) method called the Crank-Nicolson method. In each step, a linear system has to be solved with the matrix

$$B := r \begin{pmatrix} -2(1+\frac{1}{r}) & 1 & & & & \\ 1 & -2(1+\frac{1}{r}) & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -2(1+\frac{1}{r}) & 1 & \\ & & & 1 & -2(1+\frac{1}{r}) & \end{pmatrix}.$$

The matrix is again symmetric and strict diagonal dominant.

Although the Crank-Nicolson method is consistent of order two in space and time, the computational effort is nearly the same as in the classical implicit method (3.10), which is just of order one in time.

### Von-Neumann boundary conditions

In case of von-Neumann boundary conditions (3.5), the values $U_j^n$ are unknown for $j = 0, M$ a priori. Thus we have to add two equations in each step of the finite difference method. For example, we apply simply the common difference formula of first order to discretise the derivatives in (3.5). It follows

$$\begin{aligned} \alpha(t_n) &= \frac{\partial u}{\partial x}(x_0, t_n) &&= \frac{1}{h}(u(x_1, t_n) - u(x_0, t_n)) + \mathcal{O}(h) \\ \beta(t_n) &= \frac{\partial u}{\partial x}(x_M, t_n) &&= \frac{1}{h}(u(x_M, t_n) - u(x_{M-1}, t_n)) + \mathcal{O}(h) \end{aligned}$$

for each $n$. We obtain the two equations

$$U_0^n = U_1^n - \alpha(t_n)h, \qquad U_M^n = U_{M-1}^n + \beta(t_n)h,$$

which can be used to eliminate the unknowns $U_0^n, U_M^n$ in each time layer. Consequently, the finite difference methods are applied as described above.

### Source terms

The finite difference methods can be generalised directly to a heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t, u)$$

including a source term $f$. For example, the classical explicit method (3.8) becomes

$$U_j^{n+1} = rU_{j-1}^n + (1 - 2r)U_j^n + rU_{j+1}^n + kf(jh, nk, U_j^n)$$

for $j = 1, \ldots, M-1$. Just a function evaluation of $f$ is necessary to achieve the approximation. In case of the classical implicit method (3.10), we obtain

$$-rU_{j-1}^{n+1} + (1 + 2r)U_j^{n+1} - rU_{j+1}^{n+1} - kf(jh, (n+1)k, U_j^{n+1}) = U_j^n$$

for $j = 1, \ldots, M-1$. If the source term $f$ depends nonlinearly on $u$, then a nonlinear system has to be solved to obtain the approximations in each time layer.

## 3.3 Stability analysis

Now we investigate the stability of the finite difference methods. We analyse the amplification (or damping) of errors in the initial values.

### Direct estimation

Let $u_j^n := u(x_j, t_n)$ be the values of the exact solution of the heat equation (3.2) and $U_j^n$ the approximations in the finite difference method. We define the global errors

$$z_j^n := u_j^n - U_j^n.$$

Assuming that the boundary conditions are given exactly, it holds $z_j^n = 0$ for $j = 0, M$. For the classical explicit method (3.8), we obtain

$$z_j^{n+1} = r z_{j-1}^n + (1-2r) z_j^n + r z_{j+1}^n + \mathcal{O}(k^2 + kh^2).$$

We assume $r \leq \frac{1}{2}$ now. It follows the estimate

$$|z_j^{n+1}| \leq r|z_{j-1}^n| + (1-2r)|z_j^n| + r|z_{j+1}^n| + C(k^2 + kh^2)$$

with a constant $C \geq 0$, see (3.9). We define

$$\|z^n\| := \max_{j=0,\ldots,M} |z_j^n|,$$

which represents the maximum error in each time step. It follows

$$\|z^{n+1}\| \leq r\|z^n\| + (1-2r)\|z^n\| + r\|z^n\| + C(k^2 + kh^2)$$

and thus

$$\|z^{n+1}\| \leq \|z^n\| + C(k^2 + kh^2).$$

We obtain successively due to $nk \leq T$

$$\|z^n\| \leq \|z^0\| + nC(k^2 + kh^2) \leq \|z^0\| + CT(k + h^2)$$

for all $n = 1, \ldots, N$. If $k, h \to 0$ and $\|z^0\| \to 0$, then the global error converges to zero provided that $r \leq \frac{1}{2}$ holds.

**Matrix stability analysis**

We analyse the classical implicit method (3.10) now. Let

$$U^n := (U_1^n, \ldots, U_{M-1}^n)^\top, \quad z^n := (z_1^n, \ldots, z_{M-1}^n)^\top$$

be the approximations and the corresponding global errors, respectively. We assume exact boundary conditions in the finite difference method again. The global error satisfies the linear system

$$Bz^{n+1} = z^n + k\tau^n$$

with the matrix (3.11) and the local discretisation errors

$$\tau^n := (\tau_1^n, \ldots, \tau_{M-1}^n)^\top.$$

It holds $\tau_j^n = \mathcal{O}(k + h^2)$. Subsequently, we obtain

$$z^n = (B^{-1})^n z^0 + k \sum_{i=1}^{n} (B^{-1})^i \tau^{n-i}. \tag{3.14}$$

It holds $B = I + r\hat{B}$ with the tridiagonal matrix

$$\hat{B} := \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{(M-1)\times(M-1)}.$$

Let $\lambda_i$ for $i = 1, \ldots, M-1$ be the eigenvalues of $\hat{B}$. The theorem of Gerschgorin implies $0 \leq \lambda_i \leq 4$. For the inverse matrix, it holds

$$B^{-1} = \left(I + r\hat{B}\right)^{-1}.$$

Let $\mu_i$ be the eigenvalues of the matrix $B^{-1}$. It follows

$$\mu_i = \frac{1}{1 + r\lambda_i}$$

for all $i$. Hence we obtain $0 < \mu_i \leq 1$ for all $i$ and all $r > 0$. Since the matrix $B^{-1}$ is symmetric, it follows $\|B^{-1}\|_2 = \rho(B^{-1}) \leq 1$ ($\rho$: spectral radius).

The formula (3.14) yields an estimate in the Euclidean norm, where we use $\|(B^{-1})^n\|_2 \leq \|B^{-1}\|_2^n$

$$\begin{aligned}
\|z^n\|_2 &\leq \|B^{-1}\|_2^n \cdot \|z^0\|_2 + k\sum_{i=1}^{n}\|B^{-1}\|_2^i \cdot \|\tau^{n-i}\|_2 \\
&\leq \|z^0\|_2 + k\sum_{i=1}^{n}\|\tau^{n-i}\|_2 \quad \leq \quad \|z^0\|_2 + nMC(k^2 + kh^2) \\
&\leq \|z^0\|_2 + CT(\tfrac{k}{h} + h) \quad = \quad \|z^0\|_2 + CT(r+1)h.
\end{aligned}$$

Hence the method is convergent for each constant $r > 0$ in case of $h \to 0$. Thereby, we assume $\|z^0\|_2 = \mathcal{O}(h)$. Remark that it holds

$$\|z^0\|_2 \leq M \max_{j=0,\ldots,M} |z_j^0| = \tfrac{1}{h} \max_{j=0,\ldots,M} |z_j^0|.$$

The above derivation just implies convergence of order one in space and of order $\frac{1}{2}$ in time for constant ratio $r$. Nevertheless, estimates in other norms can also be achieved, which confirm the convergence of order two in space and of order one in time.

The case $C = 0$ (for example, choose $u(x,t) \equiv 0$) yields $\|z^n\|_2 \leq \|z^0\|_2$ for all $n$, which demonstrates that errors in the initial values do not increase in time. It follows the stability of the finite difference method, since this estimate is independent of the choice of $k$ and $h$.

The stability alone can also be obtained as follows. Let initial values $U^0, V^0$ be given. In the classical implicit method, the corresponding approximations are defined by $BU^{n+1} = U^n$ and $BV^{n+1} = V^n$ for homogeneous boundary conditions. Defining $Z^n := U^n - V^n$, it follows $BZ^{n+1} = Z^n$. We obtain

$$Z^n = (B^{-1})^n Z^0 \quad \Rightarrow \quad \|Z^n\|_2 \leq \|(B^{-1})^n\|_2 \cdot \|Z^0\|_2 \leq \|Z^0\|_2.$$

Thereby, the estimate is independent of the used step sizes, which determine the dimension of the vectors. This relation indicates the stability of the finite difference method.

**Von-Neumann stability**

Let $\lambda \in \mathbb{R}$ be an arbitrary constant. The functions

$$u(x,t) = \mathrm{e}^{\alpha t}\mathrm{e}^{\mathrm{i}\lambda x} \tag{3.15}$$

satisfy the heat equation (3.2) provided that $\alpha = -\lambda^2$. In particular, it holds $\alpha \leq 0$ for all $\lambda$, which corresponds to the stability of initial value problems of the PDE. For initial values $u_0 \equiv 0$, the solution of (3.2) is just $u \equiv 0$. For perturbed initial values $\tilde{u}_0(x) = \mathrm{e}^{\mathrm{i}\lambda x}$, the solution (3.15) converges to original solution $u \equiv 0$. We consider pure initial value problems with $x \in (-\infty, +\infty)$ now, i.e., no boundaries appear.

Given a grid $(x_j, t_n) = (jh, nk)$, we make an ansatz for the approximation resulting from a finite difference method via

$$U_j^n = \mathrm{e}^{\alpha t_n}\mathrm{e}^{\mathrm{i}\lambda x_j} = \mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh} \tag{3.16}$$

with $\lambda \in \mathbb{R}$ and $\alpha \in \mathbb{C}$. At $t_n = 0$, the initial values

$$U_j^0 = \mathrm{e}^{\mathrm{i}\lambda jh}$$

represent harmonic oscillations, where the frequency is determined by the constant $\lambda \in \mathbb{R}$. We see these initial values as a perturbation of the initial values $u_0(x) \equiv 0$ again.

If $\lambda \in \mathbb{R}$ is given, then the corresponding $\alpha \in \mathbb{C}$ satisfying (3.16) is determined by the finite difference method. We distinguish the following cases:

- $\mathrm{Re}(\alpha) > 0$ ($\Leftrightarrow |\mathrm{e}^{\alpha k}| > 1$): The initial perturbation $U_j^0$ is amplified for increasing time $t > 0$.
- $\mathrm{Re}(\alpha) < 0$ ($\Leftrightarrow |\mathrm{e}^{\alpha k}| < 1$): The initial perturbation $U_j^0$ is damped for increasing time $t > 0$.
- $\mathrm{Re}(\alpha) = 0$ ($\Leftrightarrow |\mathrm{e}^{\alpha k}| = 1$): The magnitude of the initial perturbation $U_j^0$ remains constant in time.

The growth of the perturbations in dependence on the coefficients $\alpha$ motivates the following definition.

**Definition 14 (von-Neumann stability)** *A finite difference method is called stable with respect to the concept of von-Neumann, if* $Re(\alpha) \leq 0$ *holds for each* $\lambda \in \mathbb{R}$. *The method is unstable, if* $Re(\alpha) > 0$ *appears for some* $\lambda \in \mathbb{R}$.

If the method is stable with respect to the criterion of von-Neumann, then initial errors are not amplified in time. To analyse the von-Neumann stability, it is sufficient to check the term $|\mathrm{e}^{\alpha k}|$.

For the classical explicit method (3.8), the ansatz (3.16) yields

$$\mathrm{e}^{\alpha(n+1)k}\mathrm{e}^{\mathrm{i}\lambda jh} = r\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda(j-1)h} + (1-2r)\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh} + r\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda(j+1)h}$$

Dividing by $\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh}$ implies

$$\begin{aligned} \mathrm{e}^{\alpha k} &= r\mathrm{e}^{\mathrm{i}\lambda(-h)} + 1 - 2r + r\mathrm{e}^{\mathrm{i}\lambda h} = 1 - 2r + 2r\cos(\lambda h) \\ &= 1 + 2r(\cos(\lambda h) - 1) \in [1-4r, 1]. \end{aligned}$$

For $r \leq \frac{1}{2}$, it holds $-1 \leq \mathrm{e}^{\alpha k} \leq 1$ for all $\lambda \in \mathbb{R}$. Hence the method is stable for $r \leq \frac{1}{2}$. For each $r > \frac{1}{2}$, a constant $\lambda \in \mathbb{R}$ exists such that $|\mathrm{e}^{\alpha k}| > 1$. Thus the method is unstable for $r > \frac{1}{2}$. Furthermore, it holds $0 \leq \mathrm{e}^{\alpha k} \leq 1$ for $r \leq \frac{1}{4}$. The criterion of von-Neumann is in agreement to the direct estimate for the classical explicit method, where $r \leq \frac{1}{2}$ was assumed. Moreover, instability is proved for $r > \frac{1}{2}$, which we were not able to show by direct estimation.

For the classical implicit method (3.10), we apply the ansatz (3.16) and obtain

$$-r\mathrm{e}^{\alpha(n+1)k}\mathrm{e}^{\mathrm{i}\lambda(j-1)h} + (1+2r)\mathrm{e}^{\alpha(n+1)k}\mathrm{e}^{\mathrm{i}\lambda jh} - r\mathrm{e}^{\alpha(n+1)k}\mathrm{e}^{\mathrm{i}\lambda(j+1)h} = \mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh}.$$

Dividing by $\mathrm{e}^{\alpha(n+1)k}\mathrm{e}^{\mathrm{i}\lambda jh}$ yields

$$\mathrm{e}^{-\alpha k} = -r\mathrm{e}^{\mathrm{i}\lambda(-h)} + 1 + 2r - r\mathrm{e}^{\mathrm{i}\lambda h} = 1 + 2r(1 - \cos(\lambda h))$$

and thus (using $1 - \cos(\gamma) = 2\sin^2(\frac{\gamma}{2})$)

$$\mathrm{e}^{\alpha k} = \frac{1}{1 + 4r\sin^2\left(\frac{\lambda h}{2}\right)} \in [0, 1].$$

It follows that the classical implicit method is stable for all $r > 0$. This criterion is in agreement to the matrix stability analysis applied to the classical implicit method.

The leapfrog method (3.12) implies the equation

$$e^{\alpha(n+1)k}e^{i\lambda jh} = e^{\alpha(n-1)k}e^{i\lambda jh} + 2r(e^{\alpha nk}e^{i\lambda(j-1)h} + e^{\alpha nk}e^{i\lambda(j+1)h}) - 4re^{\alpha nk}e^{i\lambda jh}.$$

Dividing by $e^{\alpha nk}e^{i\lambda jh}$ results in

$$e^{\alpha k} = e^{-\alpha k} + 2r(e^{i\lambda(-h)} + e^{i\lambda h}) - 4r = e^{-\alpha k} + 4r(\cos(\lambda h) - 1).$$

It follows the quadratic equation (using $1 - \cos(\gamma) = 2\sin^2(\frac{\gamma}{2})$)

$$\left(e^{\alpha k}\right)^2 + 8r\sin^2\left(\tfrac{\lambda h}{2}\right)e^{\alpha k} - 1 = 0. \tag{3.17}$$

Let $\xi := e^{\alpha k}$ and $b := 8r\sin^2\left(\frac{\lambda h}{2}\right)$. The roots of the quadratic equation are

$$\xi_{1/2} = \tfrac{1}{2}\left[-b \pm \sqrt{b^2+4}\right] \in \mathbb{R}.$$

We deduce

$$\xi_1 \cdot \xi_2 = \tfrac{1}{4}\left((-b)^2 - \sqrt{b^2+4}^2\right) = -1.$$

It follows $\xi_1 \neq \xi_2$, $\xi_1, \xi_2 \neq 0$ and

$$|\xi_1| = \frac{1}{|\xi_2|}, \quad |\xi_2| = \frac{1}{|\xi_1|}.$$

If $|\xi_1| < 1$, then $|\xi_2| > 1$ and vice versa. The case $\xi_1 = 1$, $\xi_2 = -1$ can be excluded by inserting $\xi_{1/2}$ in (3.17). Hence one root satisfies $|e^{\alpha k}| > 1$. The leapfrog method is unstable for all $r > 0$.

Furthermore, it can be shown that the Crank-Nicolson method (3.13) is stable for each $r > 0$. This criterion is in agreement to the matrix stability analysis applied to the Crank-Nicolson scheme.

**Remark:** It can be shown again that the stability is necessary and sufficient for the convergence in case of a consistent finite difference method.

### Heat equation with coefficient

For the heat equation $v_t = \lambda v_{xx}$ with a constant $\lambda > 0$, the linear transformation $v(x,t) = u(x, \lambda t)$ yields the standardised heat equation $u_t = u_{xx}$, which has been discussed above. We consider a finite difference method. Let $r := \frac{k}{h^2}$. If the stability implies a restriction like $r \leq c$ for some constant $c > 0$ in case of $u_t = u_{xx}$, then it follows the condition $r \leq \frac{c}{\lambda}$ in case of $v_t = \lambda v_{xx}$. Hence a disadvantageous restriction occurs on the time step size ($k \leq \frac{c}{\lambda} h^2$) in case of large constants $\lambda$.

## 3.4 Semidiscretisation

The idea of semidiscretisation is to replace just one partial derivative in the PDE by a difference formula. It follows a system of ordinary differential equations (ODEs). Consequently, the system of ODEs can be solved by standard numerical algorithms.

### Method of lines

Let the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t,u) \tag{3.18}$$

be given including a source term $f$. We consider initial-boundary value problems in the space domain $x \in [0,1]$. Let Dirichlet boundary conditions (3.4) be specified. We apply a discretisation in space using the grid points $x_j = jh$ for $j = 0, 1, \ldots, M$ with $h := \frac{1}{M}$. In the domain of dependence, the sets $\{(x_j, t) \in \mathbb{R}^2 : t \geq 0\}$ are called the lines. We define as approximations the time-dependent functions $U_j(t) \doteq u(x_j, t)$ for $j = 1, \ldots, M-1$. Figure 12 illustrates this construction.

Now the derivative with respect to space in (3.18) is substituted by the symmetric difference formula of second order. It follows

$$\frac{\partial u}{\partial t}(x_j,t) = \frac{1}{h^2}\big[u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)\big] + f(x_j,t,u(x_j,t)) + \mathcal{O}(h^2)$$

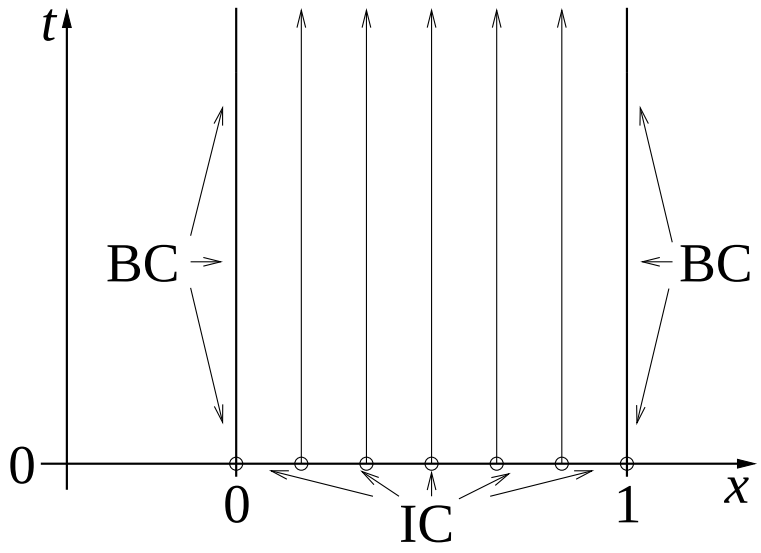



Figure 12: Method of lines.

for $j = 1, \ldots, M-1$. We rewrite these equations as a system of ODEs

$$U_j'(t) = \tfrac{1}{h^2}\Big[U_{j-1}(t) - 2U_j(t) + U_{j+1}(t)\Big] + f(x_j, t, U_j(t)) \tag{3.19}$$

for $j = 1, \ldots, M-1$. We define the abbreviations

$$B := \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{(M-1)\times(M-1)},$$

$$U(t) := \begin{pmatrix} U_1(t) \\ \vdots \\ U_{M-1}(t) \end{pmatrix},\ F(t,U) := \begin{pmatrix} f(x_1, t, U_1) \\ \vdots \\ f(x_{M-1}, t, U_{M-1}) \end{pmatrix},\ b(t) := \begin{pmatrix} \alpha(t)/h^2 \\ 0 \\ \vdots \\ 0 \\ \beta(t)/h^2 \end{pmatrix}.$$

Now the system of ODEs exhibits the compact form

$$U'(t) = BU(t) + F(t, U(t)) + b(t). \tag{3.20}$$

The initial values follow from (3.3), i.e.,

$$U(0) = (U_1(0), \ldots, U_{M-1}(0))^\top = (u_0(x_1), \ldots, u_0(x_{M-1}))^\top. \tag{3.21}$$

If no source term appears in (3.18), then it follows $F \equiv 0$ and the system of ODEs (3.20) is linear. The eigenvalues $\mu_l$ of the matrix $B$ can be calculated explicitly and it holds

$$\mu_l = -\tfrac{4}{h^2} \sin^2\left(\tfrac{\pi}{2} l h\right) < 0 \qquad \text{for } \; l = 1, 2, \ldots, M-1.$$

The magnitude of the eigenvalues is

$$\mu_{\max} \approx -\pi^2 \quad (l = 1), \qquad \mu_{\min} \approx -\frac{4}{h^2} \quad (l = M-1).$$

If $h$ is small, then we obtain $\mu_{\min} \ll \mu_{\max} < 0$. Hence the system of ODEs (3.20) becomes stiff. Implicit methods are required to solve the initial value problem of ODEs.

Now software packages for solving systems of ODEs can be applied for the initial value problem (3.20),(3.21). The explicit Euler scheme and the implicit Euler scheme yield the classical explicit method (3.8) and the classical implicit method (3.10), respectively. More sophisticated integrators can be applied like Runge-Kutta methods or multistep schemes.

Let $\tilde{U}_j(\tau_i)$ be approximations of the exact solutions $U_j(t)$ of the ODE problem (3.20),(3.21), which are computed by an ODE method with order $p$ of convergence. The error can be estimated as

$$\left|\tilde{U}_j(\tau_i) - u(x_j, \tau_i)\right| \le \left|\tilde{U}_j(\tau_i) - U_j(\tau_i)\right| + \left|U_j(\tau_i) - u(x_j, \tau_i)\right|.$$

Since the space discretisation is consistent of order two, we expect an error

$$\left|\tilde{U}_j(\tau_i) - u(x_j, \tau_i)\right| \le C(\Delta t)^p + D(\Delta x)^2 \tag{3.22}$$

with $\Delta x = h$ and $\tau_{i+1} - \tau_i \le \Delta t$ for all $i$. The error consists of two parts: the error of the space discretisation and the error of the following time discretisation. Unfortunately, the constant $C$ of the time discretisation depends on the system of ODEs (3.20) and thus on the step size $h$ in space, i.e., $C = C(h)$ In particular, the dimension $M - 1$ of the system (3.20)

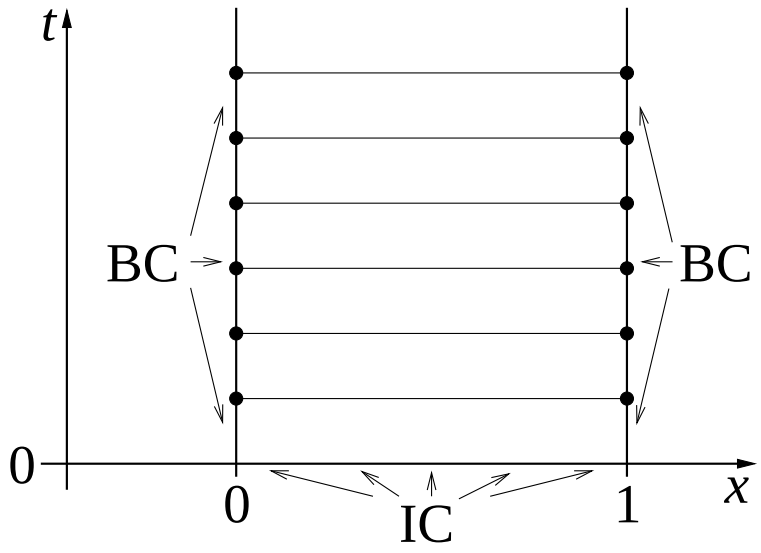



Figure 13: Rothe method.

depends on the step size $h = \frac{1}{M}$. The convergence cannot be concluded directly, since the case $C = \mathcal{O}(\frac{1}{h})$ is given in general. The two terms in the estimate (3.22) are not independent of each other.

Von-Neumann boundary conditions (3.5) can be included in the method of lines by the same strategy as in the finite difference methods, see Sect. 3.2.

### Rothe method

We consider the heat equation (3.18) including a source term again. Let Dirichlet boundary conditions (3.4) be given. In the method of Rothe, we discretise the time derivative first using the time points $t_n = kn$ for $n = 0, 1, \ldots, N$ with $k := \frac{T}{N}$. It follows

$$\tfrac{1}{k}\left[u(x, t_{n+1}) - u(x, t_n)\right] = \frac{\partial^2 u}{\partial x^2}(x, t_{n+1}) + f(x, t_{n+1}, u(x, t_{n+1}))$$

for $n = 1, \ldots, N$. Figure 13 demonstrates this semidiscretisation.

We define the approximations $z_n(x) := u(x, t_n)$ for $n = 1, \ldots, N$. It follows

a two-point boundary value problem of an ODE of second order

$$\begin{aligned} z''_{n+1}(x) &= \tfrac{1}{k}\left[z_{n+1}(x) - z_n(x)\right] - f(x, t_{n+1}, z_{n+1}(x)), \\ z_{n+1}(0) &= \alpha(t_{n+1}), \qquad z_{n+1}(1) = \beta(t_{n+1}). \end{aligned} \tag{3.23}$$

The initial conditions (3.3) yield $z_0(x) = u_0(x)$. Hence the unknown functions $z_n$ can be calculated subsequently. Thereby, numerical methods for boundary value problems of ODEs are applied. Often the equivalent system of first order corresponding to (3.23) is applied. Using $v_j := z_j$ and $w_j := z'_j$, the two-point boundary value problem reads

$$\begin{aligned} v'_{n+1}(x) &= w_{n+1}(x), \\ w'_{n+1}(x) &= \tfrac{1}{k}\left[v_{n+1}(x) - v_n(x)\right] - f(x, t_{n+1}, v_{n+1}(x)), \\ v_{n+1}(0) &= \alpha(t_{n+1}), \\ v_{n+1}(1) &= \beta(t_{n+1}). \end{aligned}$$

Typically, the method for solving the ODE problem yields approximations $z_n(x_j)$ in grid points $0 < x_1 < \cdots < x_R < 1$. Hence an interpolation scheme has to be applied to evaluate the right-hand side of the ODE (3.23) for arbitrary $x \in [0, 1]$.

Let $\tilde{z}_n(x_j)$ be the approximations obtained from an ODE solver. We achieve again an error estimate

$$|\tilde{z}_n(x_j) - u(x_j, t_n)| \leq |\tilde{z}_n(x_j) - z_n(x_j)| + |z_n(x_j) - u(x_j, t_n)|.$$

If the method for solving the ODEs is convergent of order $q$, then we expect an error

$$|\tilde{z}_n(x_j) - u(x_j, t_n)| \leq C\Delta t + D(\Delta x)^q \tag{3.24}$$

with $x_{j+1} - x_j \leq \Delta x$. Now the two terms in the estimate (3.24) are independent of each other, since the ODEs (3.23) are qualitatively the same for each $k = \Delta t$ (just the right-hand sides differ slightly). Hence we achieve good convergence properties in the Rothe method. Furthermore, it is easy to apply adaptivity with respect to the time step size $\Delta t$ as well as the space step size $\Delta x$ in the Rothe method. In contrast, changing the step size $\Delta x$ yields an ODE system of a different dimension in the method of lines.

In the method of lines, an initial value problem of a (relatively large) stiff system of ODEs has to be solved. In the Rothe method, boundary value problems of just a single ODE of second order (or system of first order with two equations) have to be resolved subsequently. However, the computational effort for solving a boundary value problem is much higher than for an initial value problem (say about 20 times for same dimensions).

**Multidimensional space domain**

We outline the application of a method of lines in case of the heat equation

$$\frac{\partial u}{\partial t} = \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \tag{3.25}$$

with two dimensions in space. We consider the domain $\Omega = (0,1)^2$ and homogeneous Dirichlet boundary conditions on $\partial\Omega$. We apply a discretisation in space according to the finite difference method from Sect. 2.2. Let $x_i := ih$ and $y_j := jh$ for $i, j = 0, 1, \ldots, M+1$ with the step size $h = \frac{1}{M+1}$. The approximations are $U_{i,j}(t) \doteq u(x_i, y_j, t)$. It follows the system of ODEs

$$U'_{i,j}(t) = \tfrac{1}{h^2}\left[U_{i-1,j}(t) + U_{i+1,j}(t) + U_{i,j-1}(t) + U_{i,j+1}(t) - 4U_{i,j}(t)\right]$$

for $i, j = 1, \ldots, M$. The homogeneous boundary conditions imply

$$U_{i,j}(t) = 0 \qquad \text{for } \; i = 0, M+1 \; \text{ or } \; j = 0, M+1$$

and all $t \geq 0$. The initial conditions $u_0 : \Omega \to \mathbb{R}$ yield

$$U_{i,j}(0) = u_0(x_i, y_j) \qquad \text{for } \; i, j = 1, \ldots, M.$$

Again an initial value problem of a system of ODEs is achieved in the method of lines. The case of three space dimensions can be handled in the same form.

Likewise, we can apply finite element methods for the discretisation in space, see Sect. 2.4. According to the Ritz-Galerkin approach, the approximation reads

$$u_h(x, y, t) = \sum_{j=1}^{N} \alpha_j(t)\phi_j(x, y)$$

with time-dependent coefficients $\alpha_j$ and space-dependent basis functions $\phi_j$. We obtain the equations

$$\sum_{i=1}^{N} \alpha_i'(t)\langle \phi_i, \phi_j \rangle_{L^2(\Omega)} = -\sum_{i=1}^{N} \alpha(t) a(\phi_i, \phi_j) \quad \text{for } j = 1, \ldots, N$$

with the bilinear form $a$. Hence an implicit system of ODEs

$$M\alpha'(t) = A\alpha(t) \tag{3.26}$$

for the coefficients $\alpha := (\alpha_1, \ldots, \alpha_N)^\top$ is achieved. The entries $M = (m_{ij})$, $A = (a_{ij})$ of the constant matrices are

$$m_{ij} = \langle \phi_i, \phi_j \rangle_{L^2(\Omega)}, \qquad a_{ij} = -a(\phi_i, \phi_j).$$

In particular, both matrices are symmetric. Again standard methods for systems of ODEs can be used to solve initial value problems of (3.26).

Methods of Rothe type can also be constructed. Given the PDE (3.25) on $\Omega = (0,1)^2$, we discretise the time derivative simply via the first-order difference formula, i.e.,

$$\tfrac{1}{k}\left[u(x, y, t_{n+1}) - u(x, y, t_n)\right] + \mathcal{O}(k) = \Delta u|_{t=t_{n+1}}.$$

Let $z_n(x, y) \doteq u(x, y, t_n)$. It follows the appraoch

$$\Delta z_{n+1} = \tfrac{1}{k}\left[z_{n+1} - z_n\right] \tag{3.27}$$

with $z_n$ given and $z_{n+1}$ unknown. The semidiscretisation (3.27) represents a Poisson equation with source term. The corresponding boundary value problems can be solved by standard numerical algorithms.

Chapter 4

# Hyperbolic PDEs of Second Order

We discuss numerical methods for hyperbolic PDEs of second order now. The benchmark is the wave equation. The speed of the transport of information is finite in hyperbolic models.

## 4.1 Wave equation

The wave equation for one space dimension is given by

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2} \tag{4.1}$$

with the wave speed $c > 0$. Using an arbitrary function $\Phi : \mathbb{R} \to \mathbb{R}$ with $\Phi \in C^2$, the functions

$$u(x,t) := \Phi(x + ct) \qquad \text{and} \qquad u(x,t) := \Phi(x - ct)$$

are both solutions of (4.1).

A pure initial value problem is called Cauchy problem, which reads

$$u(x,0) = u_0(x), \qquad \tfrac{\partial u}{\partial t}(x,0) = u_1(x) \tag{4.2}$$

with predetermined functions $u_0, u_1 : \mathbb{R} \to \mathbb{R}$ at time $t_0 = 0$ without loss of generality. We assume $u_0 \in C^2$ and $u_1 \in C^1$. The solution of the Cauchy

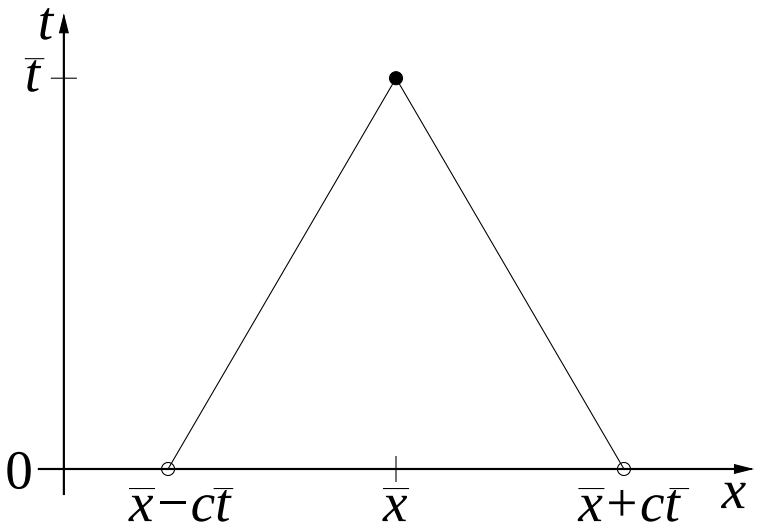



Figure 14: Domain of dependence and transport of information in case of the wave equation $u_{tt} = c^2 u_{xx}$ with one space dimension.

problem (4.1),(4.2) is given by, cf. (1.3),

$$u(x,t) = \frac{1}{2}\left( u_0(x+ct) + u_0(x-ct) + \frac{1}{c}\int_{x-ct}^{x+ct} u_1(s)\,\mathrm{d}s \right). \tag{4.3}$$

It is straightforward to verify this formula by differentiation. We recognise the finite speed of the information transport from the initial values. The solution $u$ in a point $(\bar{x}, \bar{t})$ depends on the initial values in the interval $x \in [\bar{x} - c\bar{t}, \bar{x} + c\bar{t}]$ at time $t_0 = 0$ only, see Figure 14.

The wave equation for three space dimensions reads

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u = c^2 \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \tag{4.4}$$

with the wave speed $c > 0$. Let $r := (x, y, z)$. Particular solutions of (4.4) are given by

$$u(x,y,z,t) = \mathrm{e}^{\mathrm{i}(r\cdot k - \omega t)} = \mathrm{e}^{\mathrm{i}(k_x x + k_y y + k_z z - \omega t)}$$

with the frequency $\omega > 0$ and the wave vector $k := (k_x, k_y, k_z)$ provided that

$$\omega^2 = c^2(k_x^2 + k_y^2 + k_z^2) \quad \Rightarrow \quad \omega = c\|k\|_2.$$

The Cauchy problem is given by (4.2) with initial functions $u_0, u_1 : \mathbb{R}^3 \to \mathbb{R}$.

**Theorem 14** *The solution of the Cauchy problem (4.4), (4.2) is given by*

$$u(x,t) = \frac{1}{4\pi c^2 t^2} \iint_{\|y-x\|_2=ct} u_0(y) + t u_1(y) + (y-x)^\top \nabla u_0(y) \, \mathrm{d}y \qquad (4.5)$$

*for* $x \in \mathbb{R}^3$.

Proof:

We define the spherical means

$$w(x,\theta,t) := \frac{1}{4\pi} \iint_{\|z\|_2=1} u(x+\theta z, t) \, \mathrm{d}z.$$

For each continuous function $u$, it holds

$$\lim_{\theta \to 0} w(x,\theta,t) = u(x,t).$$

We show that the spherical means satisfy the wave equation $(\theta w)_{tt} = c^2 (\theta w)_{rr}$ of the one-dimensional case. It holds

$$\Delta_x w = \frac{1}{4\pi} \iint_{\|z\|_2=1} \Delta_x u(x+\theta z, t) \, \mathrm{d}z = \iint_{\|z\|_2=1} \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}(x+\theta z, t) \, \mathrm{d}z = \frac{1}{c^2} w_{tt}.$$

The formula of Darboux for the spherical means yields $(\theta^2 w_\theta)_\theta = \Delta_x (\theta^2 w)$. It follows

$$\theta w_{tt} = \theta c^2 \Delta_x w = \tfrac{1}{\theta} c^2 (\Delta_x (\theta^2 w)) = \tfrac{1}{\theta} c^2 (\theta^2 w_\theta)_\theta = c^2 (\theta w)_{\theta\theta}.$$

For the one-dimensional wave equation, we obtain the solution (4.3), i.e.,

$$\theta w(x,\theta,t) = \tfrac{1}{2} \left( (\theta + ct) w(x, \theta + ct, 0) + (\theta - ct) w(x, \theta - ct, 0) + \frac{1}{c} \int_{\theta - ct}^{\theta + ct} s w_t(x, s, 0) \, \mathrm{d}s \right).$$

Applying the symmetry $w(x, \theta - ct, 0) = w(x, ct - \theta)$, it follows

$$w(x,\theta,t) = \tfrac{1}{2\theta} \left[ (ct+\theta) w(x, \theta + ct, 0) - (ct - \theta) w(x, ct - \theta, 0) \right] + \tfrac{1}{2c\theta} \int_{\theta - ct}^{\theta + ct} s w_t(x, s, 0) \, \mathrm{d}s.$$

For a general function $f \in C^1$, it holds

$$\lim_{\theta \to 0} \tfrac{1}{2\theta} \left[ f(ct+\theta) - f(ct-\theta) \right] = f'(ct) = \tfrac{1}{c} \cdot \tfrac{\mathrm{d}}{\mathrm{d}t} f(ct).$$

It follows

$$\lim_{\theta \to 0} \tfrac{1}{2\theta} \left[ (ct+\theta) w(x, \theta + ct, 0) - (ct - \theta) w(x, ct - \theta, 0) \right] = \tfrac{1}{c} \cdot \tfrac{\mathrm{d}}{\mathrm{d}t} \left[ ct \cdot w(x, ct, 0) \right] =: A.$$

The function $w$ and thus $w_t$ are symmetric with respect to $\theta$. It follows

$$\begin{aligned}\int_{\theta-ct}^{\theta+ct} s w_t(x,s,0)\,\mathrm{d}s &= \int_{ct-\theta}^{ct+\theta} s w_t(x,s,0)\,\mathrm{d}s + \int_{\theta-ct}^{ct-\theta} s w_t(x,s,0)\,\mathrm{d}s \\ &= \int_{ct-\theta}^{ct+\theta} s w_t(x,s,0)\,\mathrm{d}s.\end{aligned}$$

We obtain

$$\lim_{\theta\to 0}\frac{1}{2c\theta}\int_{ct-\theta}^{ct+\theta} s w_t(x,s,0)\,\mathrm{d}s = t w_t(x,ct,0) =: B.$$

It follows

$$u(x,t) = A + B = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t}{4\pi}\iint_{\|z\|_2=1} u_0(x+ctz)\,\mathrm{d}z\right) + \frac{t}{4\pi}\iint_{\|z\|_2=1} u_1(x+ctz)\,\mathrm{d}z.$$

We calculate using the product rule of differentiation

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t}{4\pi}\iint_{\|z\|_2=1} u_0(x+ctz)\,\mathrm{d}z\right) \\ =\ &\frac{1}{4\pi}\iint_{\|z\|_2=1} u_0(x+ctz)\,\mathrm{d}z + \frac{t}{4\pi}\iint_{\|z\|_2=1} (cz)^\top \nabla u_0(x+ctz)\,\mathrm{d}z.\end{aligned}$$

The substitution

$$y := x + ctz, \qquad \mathrm{d}y = (ct)^2 \mathrm{d}z$$

yields the formula (4.5). The area of the surface of the sphere $\{y : \|y-x\|_2 = ct\}$ is just $4\pi(ct)^2$. Thus it holds $y - x = \mathcal{O}(t)$. For $t \approx 0$, the formula (4.5) implies

$$u(x,t) \approx u_0(x) + t u_1(x).$$

Hence the initial conditions are satisfied. □

We encounter again the finite speed $c$ for the transport of information. Given a point $(\bar{x},\bar{t})$ with $\bar{x} \in \mathbb{R}^3$ and $\bar{t} > 0$, the solution $u(\bar{x},\bar{t})$ depends on the initial values in the set $\{x \in \mathbb{R}^3 : \|x-\bar{x}\|_2 = c\bar{t}\}$.

## 4.2 Finite difference methods

Firstly, we discuss finite difference methods in the case of one space dimension. Secondly, we generalise the strategy to the multidimensional case.

### One space dimension

We apply finite difference methods to a wave equation with source term

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2} + f(x,t,u) \tag{4.6}$$

including a single space dimension. The step sizes $k, h > 0$ are introduced in time and space, respectively. Let the grid points be $x_j = jh$ and $t_n = nh$. Typically, the partial derivatives are replaced by the difference formulas

$$\begin{aligned}
\frac{\partial^2 u}{\partial x^2}(x,t) &= \frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\vartheta h, t) \\
\frac{\partial^2 u}{\partial t^2}(x,t) &= \frac{u(x,t+k) - 2u(x,t) + u(x,t-k)}{k^2} + \frac{k^2}{12}\frac{\partial^4 u}{\partial t^4}(x, t+\eta k)
\end{aligned}$$

with $-1 < \vartheta, \eta < 1$. It follows the finite difference method

$$\tfrac{1}{k^2}\left[U_j^{n+1} - 2U_j^n + U_j^{n-1}\right] = c^2 \tfrac{1}{h^2}\left[U_{j-1}^n - 2U_j^n + U_{j+1}^n\right] + f(x_j, t_n, U_j^n)$$

or, equivalently,

$$\begin{aligned}
U_j^{n+1} &= -U_j^{n-1} + 2\left(1 - c^2\tfrac{k^2}{h^2}\right) U_j^n + c^2 \tfrac{k^2}{h^2}\left[U_{j-1}^n + U_{j+1}^n\right] \\
&\quad + k^2 f(x_j, t_n, U_j^n).
\end{aligned} \tag{4.7}$$

Hence we achieve an explicit two-stage method. The discretisation applies a five-point star. The local discretisation error of this scheme reads

$$\tau(k,h) := \tfrac{k^2}{12}\tfrac{\partial^4 u}{\partial t^4}(x, t+\eta k) - c^2 \tfrac{h^2}{12}\tfrac{\partial^4 u}{\partial x^4}(x+\vartheta h, t).$$

For $u \in C^4$, the consistency of order two follows from the uniform estimate

$$|\tau(k,h)| \leq k^2 \tfrac{1}{12} \max_{x\in[a,b], t\in[0,T]} \left|\tfrac{\partial^4 u}{\partial t^4}\right| + h^2 \tfrac{c^2}{12} \max_{x\in[a,b], t\in[0,T]} \left|\tfrac{\partial^4 u}{\partial x^4}\right|$$

for $x \in (a,b)$ and $t \in (0,T)$ for arbitrary $a, b \in \mathbb{R}$ and $T > 0$.

We consider the Cauchy problem (4.2). For the given finite difference method (4.7), we require the initial values $U_j^0$ and $U_j^1$ for each $j$. The predetermined initial values imply

$$U_j^0 = u_0(x_j), \qquad U_j^1 = u_0(x_j) + k u_1(x_j).$$

However, this discretisation is just consistent of order one. To achieve an overall method of second order, we apply the discretisation

$$\tfrac{1}{2k}[u(x_j,t_1) - u(x_j,t_{-1})] = u_t(x_j,t_0) + \mathcal{O}(k^2)$$

using the auxiliary time layer $t_{-1} = -k$. It follows

$$U_j^1 = U_j^{-1} + 2k u_1(x_j).$$

The finite difference method (4.7) yields for $n = 0$

$$U_j^1 = -U_j^{-1} + 2(1 - c^2\tfrac{k^2}{h^2})U_j^0 + c^2\tfrac{k^2}{h^2}\left[U_{j-1}^n + U_{j+1}^n\right] + k^2 f(x_j, 0, U_j^0).$$

and thus

$$\begin{aligned} U_j^1 &= -U_j^{-1} + 2(1 - c^2\tfrac{k^2}{h^2})u_0(x_j) + c^2\tfrac{k^2}{h^2}\left[u_0(x_{j-1}) + u_0(x_{j+1})\right] \\ &\quad + k^2 f(x_j, 0, u_0(x_j)). \end{aligned}$$

It follows the approximation

$$\begin{aligned} U_j^1 &= k u_1(x_j) + (1 - c^2\tfrac{k^2}{h^2})u_0(x_j) + c^2\tfrac{k^2}{2h^2}\left[u_0(x_{j-1}) + u_0(x_{j+1})\right] \\ &\quad + \tfrac{k^2}{2} f(x_j, 0, u_0(x_j)), \end{aligned}$$

where all terms on the right-hand side are predetermined.

We discuss the Cauchy problem (4.1), (4.2), i.e., no boundary conditions appear. The finite difference method (4.7) is applied (with $f \equiv 0$). Let $r := \frac{k}{h}$ be constant. We choose a finite interval $x \in [a,b]$ and $h = \frac{b-a}{2M}$ for some integer $M$. Let $x_j = a + jh$. If $R$ grid points are given in the time layer $t_n$, then just $R - 2$ new grid points can be used in the calculations within the time layer $t_{n+1}$. This proceeding is sketched in Figure 15. It

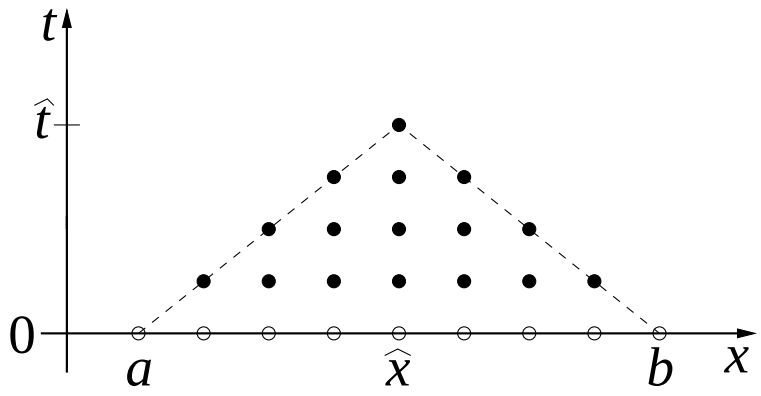



Figure 15: Grid in finite difference method for pure initial value problem.

follows that $M$ time steps can be done. We achieve an approximation in the final point

$$\hat{x} := \tfrac{a+b}{2}, \qquad \hat{t} := Mk = Mrh = r\tfrac{b-a}{2},$$

which is independent of $M$ assuming constant $r > 0$. The interval $[a, b]$ represents the domain of dependence for the numerical method (dependence on initial values).

According to (4.3), the exact solution $u(\hat{x}, \hat{t})$ depends on the initial values for $x \in [\hat{x} - c\hat{t}, \hat{x} + c\hat{t}]$, cf. Figure 14. Hence the method can only be convergent if it holds

$$\mathcal{D}(\hat{x}, \hat{t}) := [\hat{x} - c\hat{t}, \hat{x} + c\hat{t}] \subseteq [a, b] =: \mathcal{D}_0(\hat{x}, \hat{t}).$$

Otherwise, we can change the initial values for $x \notin [a, b]$ such that $u(\hat{x}, \hat{t})$ becomes different, whereas the numerical approximation remains the same. In this context, $\mathcal{D}$ and $\mathcal{D}_0$ are called the analytical domain of dependence and the numerical domain of dependence, respectively. It follows the necessary condition

$$c\hat{t} \le \tfrac{b-a}{2} \qquad \Rightarrow \qquad r \le \tfrac{1}{c}.$$

If the step size $h$ is given in space, then we obtain a restriction on the step size $k$ in time due to $r = \frac{k}{h}$. However, this restriction is not as severe as in explicit methods for parabolic problems, where $r = \frac{k}{h^2}$ holds.

Furthermore, boundary conditions can be applied at $x = a$ and/or $x = b$ for $t \ge 0$, see (3.4) and (3.5).

We analyse the stability of the method via the concept of von-Neumann. Typical solutions of $u_{tt} = c^2 u_{xx}$ read

$$u(x,t) = \mathrm{e}^{\mathrm{i}(\lambda x - \omega t)} = \mathrm{e}^{-\mathrm{i}\omega t}\mathrm{e}^{\mathrm{i}\lambda x} \qquad \text{for } \lambda, \omega \in \mathbb{R}.$$

We obtain the factor $\alpha := -\mathrm{i}\omega$ in this exact solution. It holds $\mathrm{Re}(\alpha) = 0$ and thus $|\mathrm{e}^{\alpha t}| = 1$. Perturbations in initial values are neither amplified nor damped, since they are transported in time. In contrast, the heat equation $u_t = u_{xx}$ with solution $u = \mathrm{e}^{\alpha t}\mathrm{e}^{\mathrm{i}\lambda x}$ yields $\alpha = -\lambda^2 \leq 0$. It follows $|\mathrm{e}^{\alpha k}| < 1$ for each $\lambda \neq 0$.

Now we apply the ansatz $U_j^n = \mathrm{e}^{\alpha n k}\mathrm{e}^{\mathrm{i}\lambda j h}$ in the finite difference method (4.7) without source term ($f \equiv 0$). It follows

$$\begin{aligned} \mathrm{e}^{\alpha(n+1)k}\mathrm{e}^{\mathrm{i}\lambda j h} &= -\mathrm{e}^{\alpha(n-1)k}\mathrm{e}^{\mathrm{i}\lambda j h} + 2\left(1 - c^2 r^2\right)\mathrm{e}^{\alpha n k}\mathrm{e}^{\mathrm{i}\lambda j h} \\ &\quad + c^2 r^2 \left[\mathrm{e}^{\alpha n k}\mathrm{e}^{\mathrm{i}\lambda(j-1)h} + \mathrm{e}^{\alpha n k}\mathrm{e}^{\mathrm{i}\lambda(j+1)h}\right]. \end{aligned}$$

We divide by $\mathrm{e}^{\alpha n k}\mathrm{e}^{\mathrm{i}\lambda j h}$ and obtain

$$\mathrm{e}^{\alpha k} = -\mathrm{e}^{-\alpha k} + 2\left(1 - c^2 r^2\right) + c^2 r^2 \left[\mathrm{e}^{\mathrm{i}\lambda(-h)} + \mathrm{e}^{\mathrm{i}\lambda h}\right].$$

For $\xi := \mathrm{e}^{\alpha k}$, it follows the quadratic equation

$$\xi^2 + \left(4 r^2 c^2 \sin^2\left(\tfrac{\lambda h}{2}\right) - 2\right)\xi + 1 = 0.$$

We use the abbreviation $b := 4 r^2 c^2 \sin^2\left(\tfrac{\lambda h}{2}\right) - 2$. It holds $b \in [-2, 4r^2c^2 - 2]$. The roots are

$$\xi_{1/2} = \tfrac{1}{2}\left[-b \pm \sqrt{b^2 - 4}\right].$$

We assume the necessary condition $r \leq \frac{1}{c}$. It follows $b \in [-2, 2]$ due to $r^2 c^2 \leq 1$. Thus the roots become

$$\xi_{1/2} = \tfrac{1}{2}\left[-b \pm \mathrm{i}\sqrt{4 - b^2}\right]$$

with $4 - b^2 \geq 0$. It follows

$$|\xi_{1/2}|^2 = \tfrac{1}{4}\left[(-b)^2 + \sqrt{4 - b^2}^2\right] = 1.$$

Since $|\xi_1| = 1$ and $|\xi_2| = 1$ holds, the finite difference method (4.7) is stable with respect to the criterion of von-Neumann provided that $r \leq \frac{1}{c}$

is satisfied. Moreover, the magnitude of the terms $\xi = \mathrm{e}^{\alpha k}$ agrees to the structure of the exact solutions of $u_{tt} = c^2 u_{xx}$. Due to the consistency of the method, it follows the convergence of order two for $r \leq \frac{1}{c}$. The technique (4.7) is not convergent in case of $r > \frac{1}{c}$.

Furthermore, the speed of the transport of information is finite in an explicit method – both for parabolic and hyperbolic PDEs. In contrast, the speed of the transport of information is unbounded in an implicit technique – both for parabolic and hyperbolic PDEs. Thus explicit methods fit better to the structure of hyperbolic PDEs, whereas implicit methods are more appropriate for parabolic PDEs.

### Multidimensional space

We consider the Cauchy problem (4.2) of the three-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) + f(x, y, z, t, u) \tag{4.8}$$

including a source term $f$. The derivative in time is discretised by the symmetric difference formula using the step size $k$ again. We use identical step sizes $h$ for the discretisations in the space variables. However, the difference formulas may be different. Let the grid points be $x_j = jh$, $y_p = ph$, $z_q = qh$, $t_n = nk$ and $U^n_{j,p,q} \doteq u(x_j, y_p, z_q, t_n)$ the corresponding approximations. We apply discretisations of the form

$$\begin{aligned}
\frac{\partial^2 u}{\partial x^2}(x_j, y_p, z_q, t_n) &\doteq \frac{1}{h^2} \sum_{\nu=-N}^{N} w^x_\nu U^n_{j+\nu,p,q} \\
\frac{\partial^2 u}{\partial y^2}(x_j, y_p, z_q, t_n) &\doteq \frac{1}{h^2} \sum_{\nu=-N}^{N} w^y_\nu U^n_{j,p+\nu,q} \\
\frac{\partial^2 u}{\partial z^2}(x_j, y_p, z_q, t_n) &\doteq \frac{1}{h^2} \sum_{\nu=-N}^{N} w^z_\nu U^n_{j,p,q+\nu}
\end{aligned}$$

with the coefficients $w^x_\nu, w^y_\nu, w^z_\nu \in \mathbb{R}$. The symmetric difference formula of second order exhibits the coefficients $w_0 = -2$, $w_1 = w_{-1} = 1$. The

symmetric difference formula of fourth order yields the set of coefficients $w_0 = -\frac{30}{12}$, $w_1 = w_{-1} = \frac{16}{12}$, $w_2 = w_{-2} = -\frac{1}{12}$.

The resulting finite difference method reads

$$\begin{aligned} & U_{j,p,q}^{n+1} - 2U_{j,p,q}^{n} + U_{j,p,q}^{n-1} \\ = \; & c^2 r^2 \left( \sum_{\nu=-N}^{N} w_\nu^x U_{j+\nu,p,q}^n + \sum_{\nu=-N}^{N} w_\nu^y U_{j,p+\nu,q}^n + \sum_{\nu=-N}^{N} w_\nu^z U_{j,p,q+\nu}^n \right) \\ & + k^2 f(x_j, y_p, z_q, t_n, U_{j,p,q}^n) \end{aligned}$$

with $r := \frac{k}{h}$

We analyse the stability criterion of von-Neumann in case of the wave equation (4.8) without source term ($f \equiv 0$). The ansatz

$$U_{j,p,q}^n = \mathrm{e}^{\alpha n k} \mathrm{e}^{\mathrm{i}(\lambda_x j h + \lambda_y p h + \lambda_z q h)}$$

with arbitrary constants $\lambda_x, \lambda_y, \lambda_z \in \mathbb{R}$ is inserted in the formula of the finite difference method. A division by $U_{j,p,q}^n$ yields with the abbreviation $\xi := \mathrm{e}^{\alpha k}$

$$\xi - 2 + \xi^{-1} = c^2 r^2 \left( \sum_{\nu=-N}^{N} w_\nu^x \mathrm{e}^{\mathrm{i}\lambda_x \nu h} + w_\nu^y \mathrm{e}^{\mathrm{i}\lambda_y \nu h} + w_\nu^z \mathrm{e}^{\mathrm{i}\lambda_z \nu h} \right).$$

We obtain the quadratic equation $\xi^2 + b\xi + 1 = 0$ with $b := -2 - c^2 r^2 A$ and

$$A := \sum_{\nu=-N}^{N} w_\nu^x \mathrm{e}^{\mathrm{i}\lambda_x \nu h} + w_\nu^y \mathrm{e}^{\mathrm{i}\lambda_y \nu h} + w_\nu^z \mathrm{e}^{\mathrm{i}\lambda_z \nu h}.$$

We assume $A \in \mathbb{R}$ and $A < 0$ in the following, which is satisfied by the symmetric difference formulas. The roots $\xi_1, \xi_2$ of the quadratic equation satisfy $|\xi_{1/2}| = 1$ in case of $b^2 - 4 \leq 0$. It follows the demand

$$(-2 - c^2 r^2 A)^2 \leq 4 \quad \Leftrightarrow \quad 4A + c^2 r^2 A^2 \leq 0 \quad \Leftrightarrow \quad 4 + c^2 r^2 A \geq 0$$

and thus (using $A = -|A|$ due to $A < 0$)

$$r \leq \frac{2}{c\sqrt{|A|}}.$$

We determine an upper bound for $|A|$. The triangle inequality yields successively

$$|A| \leq \sum_{\nu=-N}^{N} |w_\nu^x| + |w_\nu^y| + |w_\nu^z| =: B. \tag{4.9}$$

Thus the stability criterion of von-Neumann is satisfied in the case

$$r \leq \frac{2}{c\sqrt{B}} \leq \frac{2}{c\sqrt{|A|}}.$$

We recover the one-dimensional and two-dimensional case of the wave equation by omitting the coefficients $w_\nu^y$ or $w_\nu^z$. Assuming $c = 1$, the following table illustrates the restrictions $\frac{2}{\sqrt{B}}$ on the step sizes in the symmetric finite difference schemes of order 2 and order 4 (in space):

| | one-dim. | two-dim. | three-dim. |
|---|---|---|---|
| order 2 | 1 | $\frac{1}{\sqrt{2}} \doteq 0.707$ | $\frac{1}{\sqrt{3}} \doteq 0.577$ |
| order 4 | $\frac{\sqrt{3}}{2} \doteq 0.866$ | $\sqrt{\frac{3}{8}} \doteq 0.612$ | $\frac{1}{2} = 0.5$ |

A more detailed analysis shows that these bounds on $r$ are also necessary for the stability concept of von-Neumann. It follows a restriction on the selection of the step size $k$ in time for given step size $h$ in space.

## 4.3 Methods of Characteristics

We introduce the characteristic curves of a general PDE of second order now. In case of hyperbolic PDEs, we construct a corresponding numerical method.

### Motivation

We consider a semi-linear PDE of second order

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} = f(x, y, u, u_x, u_y) \tag{4.10}$$

with solution $u : \mathbb{R}^2 \to \mathbb{R}$ and constant coefficients $A, B, C \in \mathbb{R}$. According to the classification given in Chapter 1, the PDE (4.10) is

$$\begin{array}{ll} \text{elliptic} & \text{for } AC - B^2 > 0, \\ \text{parabolic} & \text{for } AC - B^2 = 0, \\ \text{hyperbolic} & \text{for } AC - B^2 < 0. \end{array}$$

We are looking for a coordinate transformation $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ with $w(\xi, \eta) = u(x, y)$ such that the tranformed equation

$$A^* w_{\xi\xi} + 2B^* w_{\xi\eta} + C^* w_{\eta\eta} = \tilde{f}(\xi, \eta, w, w_\xi, w_\eta)$$

satisfies $A^* = C^* = 0$. Consequently, we assume $A \neq 0$ or $C \neq 0$ in (4.10). Without loss of generality, let $A \neq 0$. The transformation is bijective, if and only if it holds

$$\det \begin{pmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{pmatrix} = \xi_x \eta_y - \xi_y \eta_x \neq 0. \tag{4.11}$$

We obtain

$$\begin{aligned} u_x &= w_\xi \xi_x + w_\eta \eta_x \\ u_{xx} &= (w_{\xi\xi}\xi_x + w_{\xi\eta}\eta_x)\xi_x + w_\xi \xi_{xx} + (w_{\eta\xi}\xi_x + w_{\eta\eta}\eta_x)\eta_x + w_\eta \eta_{xx} \\ &= w_{\xi\xi}\xi_x^2 + 2w_{\eta\xi}\xi_x\eta_x + w_{\eta\eta}\eta_x^2 + w_\xi \xi_{xx} + w_\eta \eta_{xx} \\ &\quad \text{etc.} \end{aligned}$$

It follows the transformed system

$$\begin{aligned} &\left(A\xi_x^2 + 2B\xi_x\xi_y + C\xi_y^2\right) w_{\xi\xi} \\ +\, 2\left(A\xi_x\eta_x + B(\xi_x\eta_y + \xi_y\eta_x) + C\xi_y\eta_y\right) &w_{\xi\eta} \\ &+ \left(A\eta_x^2 + 2B\eta_x\eta_y + C\eta_y^2\right) w_{\eta\eta} \;=\; \tilde{f}(\xi, \eta, w, w_\xi, w_\eta). \end{aligned}$$

We want to achieve

$$\begin{aligned} A^* &:= A\xi_x^2 + 2B\xi_x\xi_y + C\xi_y^2 &&= 0, \\ C^* &:= A\eta_x^2 + 2B\eta_x\eta_y + C\eta_y^2 &&= 0. \end{aligned}$$

We obtain two quadratic equations for $\frac{\xi_x}{\xi_y}$ and $\frac{\eta_x}{\eta_y}$, respectively. However, the two quadratic equations are identical. To ensure that the coordinate transformation is bijective, we need $\frac{\xi_x}{\xi_y} \neq \frac{\eta_x}{\eta_y}$ due to (4.11), i.e., two different solutions of the quadratic equation. For $A \neq 0$, the solutions of the quadratic equation $A\mu^2 + 2B\mu + C = 0$ are

$$\mu_{1/2} = \frac{-B \pm \sqrt{B^2 - AC}}{A}.$$

The condition $B^2 - AC > 0$ is equivalent to the existence of two different solutions $\mu_1, \mu_2 \in \mathbb{R}$. Only for hyperbolic PDEs, we achieve a transformed equation

$$2B^* w_{\xi\eta} = \tilde{f}(\xi, \eta, w, w_\xi, w_\eta).$$

The involved coefficient satisfies

$$B^* = A\xi_x\eta_x + B(\xi_x\eta_y + \xi_y\eta_x) + C\xi_y\eta_y = \cdots = -\tfrac{2}{A}(B^2 - AC)\xi_y\eta_y \neq 0$$

for $\xi_y, \eta_y \neq 0$.

Since $A, B, C$ are constant, the relations $\xi_x = \mu_1\xi_y$ and $\eta_x = \mu_2\eta_y$ yield the transformation

$$\xi = \mu_1 x + y, \qquad \eta = \mu_2 x + y.$$

For $\xi = \text{const.}$ or $\eta = \text{const.}$, we obtain straight lines in the domain of dependence $(x, y)$. These straight lines are the characterisitic curves. Due to $\mu_1 \neq \mu_2$, we obtain two families of characteristic curves.

### Characteristic curves

We consider the semi-linear PDE

$$A(x,y)u_{xx} + 2B(x,y)u_{xy} + C(x,y)u_{yy} = f(x,y,u,u_x,u_y) \tag{4.12}$$

with non-constant coefficients $A, B, C$. We want to obtain a well-posed initial value problem. In the domain of dependence, let a curve

$$\mathcal{K} := \{(x(\tau), y(\tau)) : \tau \in [\tau_0, \tau_{\text{end}}]\}$$

be given with $x, y \in C^1$ and $\dot{x}(\tau)^2 + \dot{y}(\tau)^2 > 0$ for all $\tau$. In a Cauchy problem, initial values are specified on the curve, i.e.,

$$u(x(\tau), y(\tau)) = u_0(\tau), \qquad \tfrac{\partial u}{\partial n}\big|_{x=x(\tau),y=y(\tau)} = u_1(\tau) \tag{4.13}$$

with predetermined functions $u_0, u_1 : [\tau_0, \tau_{\text{end}}] \to \mathbb{R}$. Thereby, $n = (n_1, n_2)$ is a vector perpendicular to the curve $\mathcal{K}$ with $\|n\|_2 = 1$. Let $u_0 \in C^1$. The derivative of $u$ in tangential direction $s = (s_1, s_2)$ is given by

$$\tfrac{\partial u}{\partial s}\big|_{x=x(\tau),y=y(\tau)} = u_x(x(\tau), y(\tau))\dot{x}(\tau) + u_y(x(\tau), y(\tau))\dot{y}(\tau) = \dot{u}_0(\tau).$$

Since $s$ and $n$ are linearly independent, the Cauchy problem specifies all first-order derivatives $u_x, u_y$ along the curve $\mathcal{K}$. A further differentiation yields second-order derivatives

$$\dot{u}_x = \tfrac{\mathrm{d}}{\mathrm{d}\tau}u_x = u_{xx}\dot{x} + u_{xy}\dot{y}, \qquad \dot{u}_y = \tfrac{\mathrm{d}}{\mathrm{d}\tau}u_y = u_{yx}\dot{x} + u_{yy}\dot{y}. \tag{4.14}$$

For $u \in C^2$, it holds $u_{xy} = u_{yx}$. Let $\tilde{f}(\tau) := f(x(\tau), y(\tau), u(\tau), u_x(\tau), u_y(\tau))$. We write the relations (4.14) together with the PDE (4.12) as a linear system

$$\begin{pmatrix} A & 2B & C \\ \dot{x} & \dot{y} & 0 \\ 0 & \dot{x} & \dot{y} \end{pmatrix} \begin{pmatrix} u_{xx} \\ u_{xy} \\ u_{yy} \end{pmatrix} = \begin{pmatrix} \tilde{f} \\ \dot{u}_x \\ \dot{u}_y \end{pmatrix}. \tag{4.15}$$

We want that the data $u, u_x, u_y$ on $\mathcal{K}$ specifies a unique solution of the PDE (4.12). It can be shown that each Cauchy problem (4.13) has a unique solution if and only if the linear system (4.15) is uniquely solvable. Equivalently, we demand that the determinant of the matrix in (4.15) is non-zero, i.e.,

$$A\dot{y}^2 - 2B\dot{x}\dot{y} + C\dot{x}^2 \neq 0.$$

For simplicity, we assume $\dot{x} \neq 0$. Due to $y' = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\dot{y}}{\dot{x}}$, the opposite condition yields the quadratic equation

$$A(y')^2 - 2By' + C = 0.$$

It follows the definition of characteristic curves.

**Definition 15 (characteristics)** *The characteristic curves (or: characteristics) of a second-order PDE (4.12) are the real-valued solutions $y(x)$ of the ordinary differential equation*

$$y'(x) = \frac{B(x,y) \pm \sqrt{B(x,y)^2 - A(x,y)C(x,y)}}{A(x,y)} \tag{4.16}$$

*assuming $A(x,y) \neq 0$.*

It follows that the existence and uniqueness of a solution of the PDE (4.12) is not fulfilled in the Cauchy problem (4.13), if the initial curve $\mathcal{K}$ is tangential to a characteristic curve in some point. Vice versa, a unique solution exists, if the initial curve $\mathcal{K}$ is never tangential to a characterstic curve. The ODE (4.16) describes a family of characteristic curves.

For an elliptic PDE, it holds $B^2 - AC < 0$. Consequently, characteristic curves do not exist. A unique solution of the Cauchy problem exists for an arbitrary curve $\mathcal{K}$. However, the initial value problem of an elliptic PDE is not well-posed, since the solutions do not depend continuously on the initial data.

For a parabolic PDE, it holds $B^2 - AC = 0$ and thus $y' = \frac{B}{A}$. A family of characteristic curves exists. However, Cauchy problems of the form (4.13) are often not considered. For example, a pure initial value problem of the heat equation demands just the specification of the initial values $u$ at $t = 0$ and not of the normal derivative $u_t$.

For a hyperbolic PDE, it holds $B^2 - AC > 0$. Hence two families of characteristic curves exist. The initial curve $\mathcal{K}$ must never be tangential to one of these characteristics. For example, this demand is satisfied for the wave

equation $u_{tt} = c^2 u_{xx}$ in case of initial values $u, u_t$ specified at $t = 0$. Characterisitic curves are only interesting in case of hyperbolic PDEs, since the Cauchy problems (4.13) are irrelevant for elliptic PDEs or parabolic PDEs.

Definition 15 remains the same in the quasi-linear case $A = A(x, y, u, u_x, u_y)$, $B = B(x, y, u, u_x, u_y)$, $C = C(x, y, u, u_x, u_y)$. However, the characteristic curves depend on the a priori unknown solution in this case.

### Numerical method

We consider a Cauchy problem (4.13) for a hyperbolic PDE (4.12). Two families of characteristic curves exist, see Definition 15. The transport of information proceeds along the characteristic curves. We can use this property to construct a numerical method for the determination of the solution.

Along a characteristic curve, the linear system (4.15) does not exhibit a unique solution, since the involved matrix is singular. In particular, it holds

$$\operatorname{rank} \begin{pmatrix} A & 2B & C \\ \dot{x} & \dot{y} & 0 \\ 0 & \dot{x} & \dot{y} \end{pmatrix} = 2$$

for $\dot{x} \neq 0$. Nevertheless, we assume that a unique solution of some Cauchy problem exists, where the initial curve $\mathcal{K}$ is not tangential to some characteristic curve. It follows that the linear system (4.15) has a solution along the characteristic curve, which implies

$$\operatorname{rank} \begin{pmatrix} A & 2B & C & \tilde{f} \\ \dot{x} & \dot{y} & 0 & \dot{u}_x \\ 0 & \dot{x} & \dot{y} & \dot{u}_y \end{pmatrix} = 2.$$

If we choose three out of the four column vectors, the corresponding determinant is zero. In particular, it holds

$$\det \begin{pmatrix} A & C & \tilde{f} \\ \dot{x} & 0 & \dot{u}_x \\ 0 & \dot{y} & \dot{u}_y \end{pmatrix} = 0,$$

which is equivalent to

$$A\dot{u}_x\dot{y} + C\dot{u}_y\dot{x} - \tilde{f}\dot{x}\dot{y} = 0. \tag{4.17}$$

Another equivalent formulation is

$$A\frac{\dot{u}_x}{\dot{x}} + C\frac{\dot{u}_y}{\dot{y}} = \tilde{f} \qquad \text{for } \dot{y}, \dot{x} \neq 0. \tag{4.18}$$

Hence we obtain an information on the change of $u_x, u_y$ along the characteristic curves. We introduce the abbreviations

$$\alpha := \frac{B + \sqrt{B^2 - AC}}{A}, \qquad \beta := \frac{B - \sqrt{B^2 - AC}}{A}, \tag{4.19}$$

where $\alpha$ and $\beta$ depend on $x, y$. It holds $\alpha \neq \beta$ for hyperbolic PDEs. The relation (4.16) implies $\dot{y} = \alpha\dot{x}$ and $\dot{y} = \beta\dot{x}$. The two families of characteristic curves can be written as

$$\mathcal{K}_\alpha = \{(x(\tau), y(\tau)) : \dot{y} = \alpha\dot{x}\}, \qquad \mathcal{K}_\beta = \{(x(\tau), y(\tau)) : \dot{y} = \beta\dot{x}\}.$$

The equation (4.17) yields

$$\begin{aligned} A\alpha\dot{u}_x + C\dot{u}_y &= \tilde{f}\dot{y} = \alpha\tilde{f}\dot{x}, \\ A\beta\dot{u}_x + C\dot{u}_y &= \tilde{f}\dot{y} = \beta\tilde{f}\dot{x}. \end{aligned} \tag{4.20}$$

These two equations can be used to determine $u_x$ and $u_y$.

**Example:** For the hyperbolic PDE $u_{xx} - u_{yy} = 2(y^2 - x^2)$, we solve the initial value problem $u(0, y) = y^2$, $u_x(0, y) = 0$ analytically using the characteristic curves.

It holds $A = 1$, $B = 0$, $C = -1$. It follows $y' = \pm 1$ in (4.16), i.e., $\alpha = 1$, $\beta = -1$. The characteristic curves can be written as

$$\mathcal{K}_\alpha : y = C_\alpha + x, \qquad \mathcal{K}_\beta : y = C_\beta - x$$

with constants $C_\alpha, C_\beta \in \mathbb{R}$. The equation (4.18) yields

$$\begin{aligned} \frac{\dot{u}_x}{\dot{x}}(x, C_\alpha + x) - \frac{\dot{u}_y}{\dot{y}}(x, C_\alpha + x) &= 2((C_\alpha + x)^2 - x^2) = 4C_\alpha x + 2C_\alpha^2, \\ \frac{\dot{u}_x}{\dot{x}}(x, C_\beta - x) - \frac{\dot{u}_y}{\dot{y}}(x, C_\beta - x) &= 2((C_\beta - x)^2 - x^2) = -4C_\beta x + 2C_\beta^2. \end{aligned}$$

The initial values imply $u_x(0, y) = 0$, $u_y(0, y) = 2y$. It holds

$$\frac{\dot{u}}{\dot{x}} = \frac{\mathrm{d}u_x}{\mathrm{d}x}, \quad \frac{\dot{u}}{\dot{y}} = \frac{\mathrm{d}u_y}{\mathrm{d}y} = \frac{\mathrm{d}u_y}{\mathrm{d}x} \cdot \frac{\mathrm{d}x}{\mathrm{d}y} = \frac{\mathrm{d}u_y}{\mathrm{d}x} \cdot \frac{1}{y'}.$$

Integration with respect to $x$ yields

$$\begin{aligned} u_x(x, C_\alpha + x) - u_y(x, C_\alpha + x) &= 2C_\alpha x^2 + 2C_\alpha^2 x - 2C_\alpha, \\ u_x(x, C_\beta - x) + u_y(x, C_\beta - x) &= -2C_\beta x^2 + 2C_\beta^2 x + 2C_\beta. \end{aligned}$$

Two characteristic curves (for fixed $C_\alpha, C_\beta$) intersect in a point with the coordinates

$$\bar{x} = \tfrac{1}{2}(C_\beta - C_\alpha), \qquad \bar{y} = \tfrac{1}{2}(C_\alpha + C_\beta).$$

We obtain the equations

$$\begin{aligned} u_x(\bar{x}, \bar{y}) - u_y(\bar{x}, \bar{y}) &= 2(\bar{y} - \bar{x})\bar{x}^2 + 2(\bar{y} - \bar{x})^2\bar{x} - 2(\bar{y} - \bar{x}), \\ u_x(\bar{x}, \bar{y}) + u_y(\bar{x}, \bar{y}) &= -2(\bar{y} + \bar{x})\bar{x}^2 + 2(\bar{y} + \bar{x})^2\bar{x} + 2(\bar{y} + \bar{x}). \end{aligned}$$

We can solve this linear system for $u_x, u_y$ directly and obtain

$$u_x(x, y) = 2x(1 + y^2), \qquad u_y(x, y) = 2y(1 + x^2).$$

We achieve the solution $u$ via the integration

$$u(x, y) = u(x_0, y_0) + \int_{\mathcal{J}} u_x \, \mathrm{d}x + u_y \, \mathrm{d}y$$

along an arbitrary curve $\mathcal{J}$ interconnecting $(x_0, y_0)$ and $(x, y)$. We apply

$$\begin{aligned} u(x, y) &= u(0, y) + \int_0^x u_x(s, y) \, \mathrm{d}s \\ &= y^2 + \int_0^x 2s(1 + y^2) \, \mathrm{d}s \\ &= y^2 + x^2(1 + y^2). \end{aligned}$$

We have chosen a particular curve $\mathcal{J}$, which yields simple calculation. Remark that also a curve $\mathcal{K}_\alpha$ or $\mathcal{K}_\beta$ can be applied. It is straightforward to verify that the function $u$ satisfies the Cauchy problem of the PDE.

Now we construct a numerical method to achieve an approximative solution automatically. We consider a curve $\mathcal{K}$, where a Cauchy problem (4.13) is specified. We assume that the curve $\mathcal{K}$ is never tangential to a characteristic curve of the PDE (4.12). On the initial curve, we choose the points $P_1, \ldots, P_n$. Let $x(P_j), y(P_j)$ be the coordinates of the points. The values $u(P_j)$, $u_x(P_j)$, $u_y(P_j)$ are also predetermined for all $j = 1, \ldots, n$.

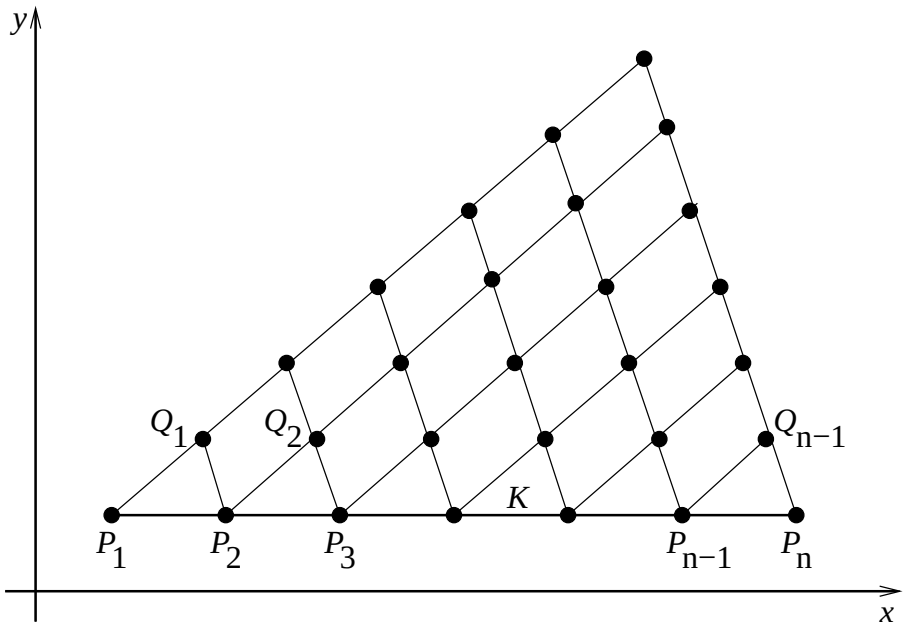



Figure 16: Grid in method of characteristics for PDE with constant coefficients.

In case of constant coefficients $A, B, C$, each family of characteristic curves is a continuum of parallel straight lines, see Figure 16. For non-constant coefficients $A, B, C$, the characteristics represent general curves, see Figure 17. Let $\mathcal{K}_\alpha^{(j)}$ be the characteristic curve of the first family and $\mathcal{K}_\beta^{(j)}$ be the characteristic curve of the second family, which are both running through the point $P_j$. The intersection of $\mathcal{K}_\alpha^{(j)}$ through $P_j$ and $\mathcal{K}_\beta^{(j+1)}$ through $P_{j+1}$ yields a new point $Q_j$ for $j = 1, \ldots, n-1$. We describe how the data $x(Q_1)$, $y(Q_1)$, $u(Q_1)$, $u_x(Q_1)$, $u_y(Q_1)$ can be computed by the corresponding data in $P_1$ and $P_2$. Successively, the other points of the characteristic grid are determined.

We discretise the corresponding ODEs $\dot{y} = \alpha\dot{x}$ and $\dot{y} = \beta\dot{x}$ of the two families of the characteristic curves, see (4.19). According to the explicit Euler method, it follows

$$\begin{aligned} y(Q_1) - y(P_1) &= \alpha(P_1)(x(Q_1) - x(P_1)), \\ y(Q_1) - y(P_2) &= \beta(P_2)(x(Q_1) - x(P_2)). \end{aligned} \tag{4.21}$$

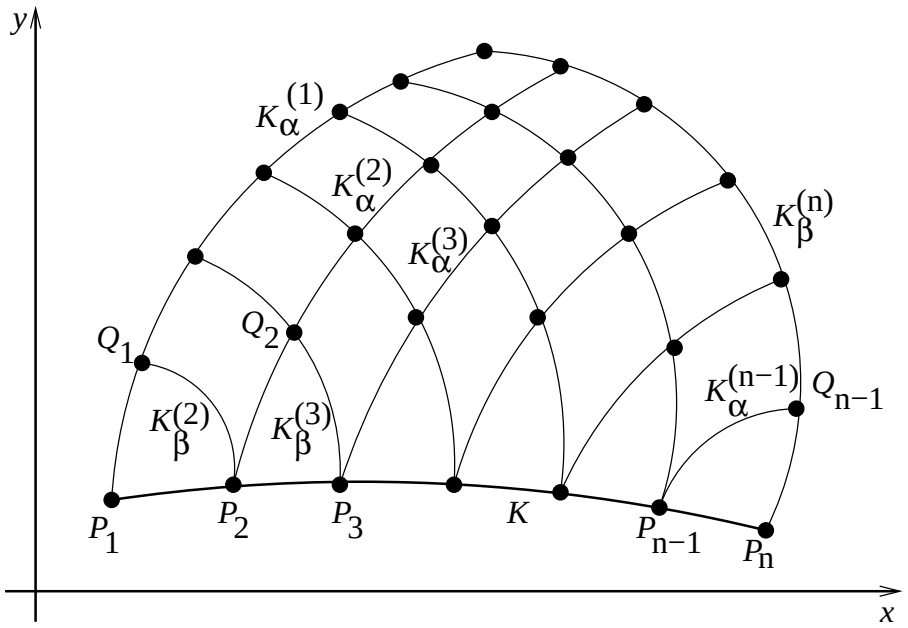



Figure 17: Grid in method of characteristics for PDE with non-constant coefficients.

Due to $\alpha = \alpha(A, B, C)$, it holds

$$\alpha(P_1) = \alpha(A(x(P_1), y(P_1)), B(x(P_1), y(P_1)), C(x(P_1), y(P_1)))$$

or, in a shorter form, $\alpha(P_1) = \alpha(x(P_1), y(P_1))$ and likewise for $\beta$. Hence we can evaluate $\alpha(P_1), \beta(P_2)$ directly. We obtain a linear system (4.21) for $x(Q_1), y(Q_1)$, which can be solved directly

$$\begin{aligned} x(Q_1) &= \frac{y(P_2)-y(P_1)+\alpha(P_1)x(P_1)-\beta(P_2)x(P_2)}{\alpha(P_1)-\beta(P_2)}, \\ y(Q_1) &= \frac{\alpha(P_1)y(P_2)-\beta(P_2)y(P_1)+\alpha(P_1)\beta(P_2)(x(P_1)-x(P_2))}{\alpha(P_1)-\beta(P_2)}. \end{aligned}$$

It holds $\alpha(P_j) \neq \beta(P_j)$ for all $j$. Hence $\alpha(P_1) \neq \beta(P_2)$ is satisfied for $P_2$ sufficiently close to $P_1$ due to the continuity of the functions. We achieve $u(Q_1)$ by a first-order approximation according to a Taylor expansion

$$u(Q_1) = u(P_1) + u_x(P_1)(x(Q_1) - x(P_1)) + u_y(P_1)(y(Q_1) - y(P_1)).$$

To be able to continue the method in other grid points, we also require approximations of $u_x(Q_1), u_y(Q_1)$. The equations (4.20) allow for the determination of $u_x, y_y$. We apply a discretisation like in the explicit Euler

scheme again

$$\begin{aligned} &A(P_1)\alpha(P_1)(u_x(Q_1) - u_x(P_1)) + C(P_1)(u_y(Q_1) - u_y(P_1)) \\ &= f(P_1)(y(Q_1) - y(P_1)), \\ &A(P_2)\beta(P_2)(u_x(Q_1) - u_x(P_2)) + C(P_2)(u_y(Q_1) - u_y(P_2)) \\ &= f(P_2)(y(Q_1) - y(P_2)), \end{aligned} \tag{4.22}$$

where

$$f(P_j) := f(x(P_j), y(P_j), u(P_j), u_x(P_j), u_y(P_j)).$$

A linear system appears for the unknowns $u_x(Q_1), u_y(Q_1)$, whose coefficient matrix is

$$G := \begin{pmatrix} A(P_1)\alpha(P_1) & C(P_1) \\ A(P_2)\beta(P_2) & C(P_2) \end{pmatrix}.$$

If holds

$$\det G = A(P_1)C(P_2)\alpha(P_1) - A(P_2)C(P_1)\beta(P_2).$$

Thus $\det G \neq 0$ is guaranteed for $P_1, P_2$ sufficiently close to each other. Hence we obtain $u_x(Q_1), u_y(Q_1)$ from the linear system (4.22).

In the quasi-linear case $A = A(x, y, u, u_x, u_y)$, $B = \cdots$, $C = \cdots$, this method is feasible using the same formulas, since the data

$$A(P_j) = A(x(P_j), y(P_j), u(P_j), u_x(P_j), u_y(P_j)), \quad \text{etc.}$$

is directly available.

Now we want to achieve a method of characteristics, which is consistent of order two. Therefore we discretise the ODEs via trapezoidal rule, i.e., an implicit scheme. For the family of characteristic curves given by $\dot{y} = \alpha\dot{x}$, it holds

$$y(Q_1) - y(P_1) = \int_{\tau(P_1)}^{\tau(Q_1)} \dot{y}\, \mathrm{d}\tau = \int_{\tau(P_1)}^{\tau(Q_1)} \alpha(\tau)\dot{x}\, \mathrm{d}\tau = \int_{x(P_1)}^{x(Q_1)} \alpha(x)\, \mathrm{d}x$$

and for $\dot{y} = \beta\dot{x}$ analogously. A discretisation by trapezoidal rule yields

$$\begin{aligned} y(Q_1) - y(P_1) &= \tfrac{1}{2}(\alpha(P_1) + \alpha(Q_1))(x(Q_1) - x(P_1)), \\ y(Q_1) - y(P_2) &= \tfrac{1}{2}(\beta(P_2) + \beta(Q_1))(x(Q_1) - x(P_2)). \end{aligned} \tag{4.23}$$

Since $\alpha(Q_1) = \alpha(x(Q_1), y(Q_1))$ and $\beta(Q_1) = \beta(x(Q_1), y(Q_1))$, we obtain a nonlinear system (4.23) for the unknowns $x(Q_1), y(Q_1)$. Newton's method yields an approximative solution. Given the solution $x(Q_1), y(Q_1)$ of the system (4.23), the terms $\alpha(Q_1), \beta(Q_1)$ can be evaluated.

We apply the equations (4.20) to determine $u_x, y_y$ again. The discretisation of second order yields

$$\begin{aligned}
&(A(P_1)\alpha(P_1) + A(Q_1)\alpha(Q_1))(u_x(Q_1) - u_x(P_1)) \\
&+ (C(P_1) + C(Q_1))(u_y(Q_1) - u_y(P_1)) \\
&= (f(P_1) + f(Q_1))(y(Q_1) - y(P_1)), \\
&(A(P_2)\beta(P_2) + A(Q_1)\beta(Q_1))(u_x(Q_1) - u_x(P_2)) \\
&+ (C(P_2) + C(Q_1))(u_y(Q_1) - u_y(P_2)) \\
&= (f(P_2) + f(Q_1))(y(Q_1) - y(P_2)).
\end{aligned} \tag{4.24}$$

For $f = f(x, y)$, the evaluation $f(Q_1)$ can be obtained from the solution $x(Q_1), y(Q_1)$ of nonlinear system (4.23). Accordingly, $\alpha(Q_1), \beta(Q_1)$ are also known from $x(Q_1), y(Q_1)$. We obtain a linear system (4.24) for the unknowns $u_x(Q_1), u_y(Q_1)$ again. It can be shown that the coefficient matrix is regular for $P_2$ sufficiently close to $P_1$. A formula for the unknowns $u_x(Q_1), u_y(Q_1)$ can be derived by Cramer's rule. Finally, the exact solution satisfies

$$u(Q_1) = u(P_1) + \int_{\mathcal{K}_\alpha^{(1)}} \dot{u} \, \mathrm{d}\tau = u(P_1) + \int_{\mathcal{K}_\alpha^{(1)}} u_x \, \mathrm{d}x + u_y \, \mathrm{d}y$$

using the part of the characteristic curve $\mathcal{K}_\alpha^{(1)}$ from $P_1$ to $Q_1$, since it holds $\dot{u} = u_x\dot{x} + u_y\dot{y}$. Trapezoidal rule yields the approximation

$$\begin{aligned}
u(Q_1) \;=\; & u(P_1) + \tfrac{1}{2}(u_x(P_1) + u_x(Q_1))(x(Q_1) - x(P_1)) \\
& + \tfrac{1}{2}(u_y(P_1) + u_y(Q_1))(y(Q_1) - y(P_1)),
\end{aligned} \tag{4.25}$$

which is consistent of second order.

In the semi-linear case $f = f(x, y, u, u_x, u_y)$, we solve the equations (4.24) together with (4.25) for the unknowns $u(Q_1), u_x(Q_1), u_y(Q_1)$, which represents a nonlinear system in general.

In the quasi-linear case $A = A(x, y, u, u_x, u_y)$, $B = \cdots$, $C = \cdots$, we solve a nonlinear system consisting of (4.23), (4.24), (4.25) for the five unknowns $x(Q_1), y(Q_1), u(Q_1), u_x(Q_1), u_y(Q_1)$.

In the linear case (also with non-constant coefficients), all grid points can be calculated a priori without the determination of $u, u_x, u_y$. However, this is not a significant advantage. In the quasi-linear case, the grid points have to be computed successively together with $u, u_x, u_y$.

# Chapter 5

# Hyperbolic Systems of First Order

We consider systems of PDEs of first order now. Hyperbolic systems exhibit similar properties as hyperbolic PDEs of second order. For example, the speed of the transport of information is finite again.

## 5.1 Systems of two equations

A semi-linear PDE of second order

$$A(x,y)w_{xx} + 2B(x,y)w_{xy} + C(x,y)w_{yy} = f(x,y,w_x,w_y)$$

can be transformed into a system of two PDEs of first order via $u := w_x$, $v := w_y$ using the compatibility condition $w_{xy} = w_{yx}$

$$\begin{aligned} A(x,y)u_x + 2B(x,y)u_y + C(x,y)v_y &= f(x,y,u,v), \\ v_x - u_y &= 0. \end{aligned} \tag{5.1}$$

More general, we discuss a system of PDEs of first order in the form

$$\begin{aligned} a_1u_x + b_1u_y + c_1v_x + d_1v_y &= f_1(x,y,u,v) \\ a_2u_x + b_2u_y + c_2v_x + d_2v_y &= f_2(x,y,u,v) \end{aligned}$$

with coefficients depending on $x$ and $y$. Equivalently, we write the system as

$$\begin{pmatrix} a_1 & c_1 \\ a_2 & c_2 \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} b_1 & d_1 \\ b_2 & d_2 \end{pmatrix} \frac{\partial}{\partial y} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} f_1(x,y,u,v) \\ f_2(x,y,u,v) \end{pmatrix}. \tag{5.2}$$

We consider a curve $\mathcal{K} := \{(x(\tau), y(\tau) : \tau \in [\tau_0, \tau_{\text{end}}]\}$ again. A Cauchy problem is specified by initial conditions

$$u(x(\tau), y(\tau)) = u_0(\tau), \qquad v(x(\tau), y(\tau)) = v_0(\tau) \tag{5.3}$$

with predetermined functions $u_0, v_0 : [\tau_0, \tau_{\text{end}}] \to \mathbb{R}$. Thereby, we also obtain the information $\dot{u} = \dot{u}_0$, $\dot{v} = \dot{v}_0$ along the curve $\mathcal{K}$. Furthermore, it holds

$$\dot{u} := \frac{\mathrm{d}u}{\mathrm{d}\tau} = u_x\dot{x} + u_y\dot{y}, \qquad \dot{v} := \frac{\mathrm{d}v}{\mathrm{d}\tau} = v_x\dot{x} + v_y\dot{y}.$$

Together with the PDEs (5.2), we obtain the linear system

$$\begin{pmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ \dot{x} & \dot{y} & 0 & 0 \\ 0 & 0 & \dot{x} & \dot{y} \end{pmatrix} \begin{pmatrix} u_x \\ u_y \\ v_x \\ v_y \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \dot{u} \\ \dot{v} \end{pmatrix}. \tag{5.4}$$

The determinant $D$ of the involved coefficient matrix reads

$$D := \dot{y}^2(a_1c_2 - c_1a_2) - \dot{x}\dot{y}(a_1d_2 - d_1a_2 + b_1c_2 - c_1b_2) + \dot{x}^2(b_1d_2 - d_1b_2).$$

We introduce the abbreviations

$$\begin{aligned} \bar{A} &:= \det \begin{pmatrix} a_1 & c_1 \\ a_2 & c_2 \end{pmatrix}, \quad \bar{C} := \det \begin{pmatrix} b_1 & d_1 \\ b_2 & d_2 \end{pmatrix}, \\ \bar{B} &:= \frac{1}{2}\left[\det \begin{pmatrix} a_1 & d_1 \\ a_2 & d_2 \end{pmatrix} + \det \begin{pmatrix} b_1 & c_1 \\ b_2 & c_2 \end{pmatrix}\right]. \end{aligned} \tag{5.5}$$

It follows $D = \bar{A}\dot{y} - 2\bar{B}\dot{x}\dot{y} + \bar{C}\dot{x}^2$. Again characteristic curves are defined by the property $D = 0$, i.e., the linear system is not uniquely solvable.

**Definition 16 (characteristics)** *The characteristic curves (or: characteristics) of a system (5.2) of PDEs of first order are the real-valued solutions $y(x)$ of the ordinary differential equation*

$$y'(x) = \frac{\bar{B}(x,y) \pm \sqrt{\bar{B}(x,y)^2 - \bar{A}(x,y)\bar{C}(x,y)}}{\bar{A}(x,y)} \tag{5.6}$$

*with the functions from (5.5) assuming $\bar{A}(x,y) \neq 0$.*

A curve including initial conditions (5.3) of a Cauchy problem must not be tangential to a characteristic curve. We use the definition of the characteristic curves to classify the systems (5.2) according to PDEs of second order.

**Definition 17** *The system (5.2) of first order is called*

$$\begin{aligned} &\textit{elliptic} && \textit{if } \bar{A}\bar{C} - \bar{B}^2 > 0 \ , \\ &\textit{parabolic} && \textit{if } \bar{A}\bar{C} - \bar{B}^2 = 0 \ , \\ &\textit{hyperbolic} && \textit{if } \bar{A}\bar{C} - \bar{B}^2 < 0 \ . \end{aligned}$$

A hyperbolic system exhibits two different families of characteristics, a parabolic system has one family of characteristics and an elliptic system does not include characteristics at all.

For the system (5.1), it holds $a_1 = A$, $b_1 = 2B$, $d_1 = C$, $b_2 = -1$, $c_2 = 1$ and the other coefficients are zero. It follows $\bar{A} = A$, $\bar{B} = B$ and $\bar{C} = C$. Hence Definition 17 is in agreement to the classification of PDEs of second order.

We can apply the characteristic curves to construct a numerical method for solving the systems (5.2). Along a characteristic curve, it holds

$$\operatorname{rank} \begin{pmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ \dot{x} & \dot{y} & 0 & 0 \\ 0 & 0 & \dot{x} & \dot{y} \end{pmatrix} \leq 3,$$

since the matrix is singular. Nevertheless, we assume that the PDE system has a solution of a Cauchy problem, where the initial curve is not a characteristics. Consequently, the solution satisfies the linear system (5.4) also along the characteristics. It follows

$$\operatorname{rank} \begin{pmatrix} a_1 & b_1 & c_1 & d_1 & f_1 \\ a_2 & b_2 & c_2 & d_2 & f_2 \\ \dot{x} & \dot{y} & 0 & 0 & \dot{u} \\ 0 & 0 & \dot{x} & \dot{y} & \dot{v} \end{pmatrix} \leq 3.$$

If we choose four columns of this extended matrix, then the corresponding determinant results to zero. For example, it holds

$$\begin{aligned} & \det \begin{pmatrix} a_1 & b_1 & c_1 & f_1 \\ a_2 & b_2 & c_2 & f_2 \\ \dot{x} & \dot{y} & 0 & \dot{u} \\ 0 & 0 & \dot{x} & \dot{v} \end{pmatrix} \\ = \quad & -\dot{x}(a_1 b_2 \dot{u} + b_2 f_2 \dot{x} + a_2 f_1 \dot{y} - b_2 f_1 \dot{x} - a_1 f_2 \dot{y} - a_2 b_1 \dot{u}) \\ & + \dot{v}(b_1 c_2 \dot{x} + a_2 c_1 \dot{y} - c_1 b_2 \dot{x} - a_1 c_2 \dot{y}) \quad = \quad 0. \end{aligned}$$

This relation can be used in a method of characteristics to determine the solution of a Cauchy problem.

We consider the special case of a system (5.2)

$$\frac{\partial}{\partial x} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} b_1 & d_1 \\ b_2 & d_2 \end{pmatrix} \frac{\partial}{\partial y} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} f_1(x, y, u, v) \\ f_2(x, y, u, v) \end{pmatrix}, \tag{5.7}$$

i.e., $a_1 = c_2 = 1$, $a_2 = c_1 = 0$. The system (5.2) is equivalent to a PDE of the form (5.7) provided that the first matrix is regular. It follows $\bar{A} = 1$, $\bar{B} = \frac{1}{2}(b_1 + d_2)$, $\bar{C} = b_1 d_2 - b_2 d_1$. The characteristic curves of the system (5.2) are defined by

$$y'(x) = -\tfrac{1}{2}(b_1 + d_2) \pm \sqrt{\tfrac{1}{4}(b_1 + d_2)^2 - (b_1 d_2 - b_2 d_1)}.$$

The system (5.7) is hyperbolic if and only if

$$\tfrac{1}{4}(b_1 + d_2)^2 > b_1 d_2 - b_2 d_1. \tag{5.8}$$

We investigate the eigenvalues $\lambda$ of the matrix in (5.7). The characteristic polynomial reads

$$\det \begin{pmatrix} b_1 - \lambda & d_1 \\ b_2 & d_2 - \lambda \end{pmatrix} = \lambda^2 - \lambda(b_1 + d_2) + b_1 d_2 - b_2 d_1$$

and thus

$$\lambda_{1/2} = \tfrac{1}{2}(b_1 + d_2) \pm \sqrt{(b_1 + d_2)^2 - 4(b_1 d_2 - b_2 d_1)}.$$

Hence two different eigenvalues $\lambda_1$ and $\lambda_2$ exist if the condition (5.8) holds. Consequently, the matrix in (5.7) is real diagonalisable (matrix is diagonalisable and all eigenvalues are real).

This property motivates the definition of hyperbolic systems of PDEs including $n \geq 2$ unknowns.

**Definition 18 (linear hyperbolic systems)** *A linear system of PDEs*

$$\frac{\partial u}{\partial t} + A \frac{\partial u}{\partial x} = f(x,t,u) \tag{5.9}$$

*with solution* $u : (x_0, x_1) \times (t_0, t_1) \to \mathbb{R}^n$ *and a constant matrix* $A \in \mathbb{R}^{n \times n}$ *is called hyperbolic if and only if the matrix* $A$ *is real diagonalisable.*

In the case $n = 2$, a hyperbolic system (5.9) w.r.t. Definition 17 is also hyperbolic w.r.t. Definition 18. Vice versa, a hyperbolic system (5.9) w.r.t. Definition 18 is hyperbolic or parabolic w.r.t. Definition 17 for $n = 2$. A corresponding definition of elliptic and parabolic PDEs with $n > 2$ unknowns does not exist. The special case $n = 1$ is always hyperbolic.

A hyperbolic system (5.9) can be decoupled in the following sense. Let $A = SDS^{-1}$ with a regular matrix $S \in \mathbb{R}^{n \times n}$ and a diagonal matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n \times n}$. Using $v := S^{-1}u$, it follows

$$\begin{aligned}
\tfrac{\partial u}{\partial t} + A \tfrac{\partial u}{\partial x} &= f(x,t,u) \\
\tfrac{\partial u}{\partial t} + SDS^{-1} \tfrac{\partial u}{\partial x} &= f(x,t,u) \\
S^{-1}\tfrac{\partial u}{\partial t} + DS^{-1} \tfrac{\partial u}{\partial x} &= S^{-1} f(x,t,u) \\
\tfrac{\partial S^{-1}u}{\partial t} + D \tfrac{\partial S^{-1}u}{\partial x} &= S^{-1} f(x,t,u) \\
\tfrac{\partial v}{\partial t} + D \tfrac{\partial v}{\partial x} &= S^{-1} f(x,t,Sv).
\end{aligned}$$

Defining $g := S^{-1}f$, we obtain the equations

$$\frac{\partial v_j}{\partial t} + \lambda_j \frac{\partial v_j}{\partial x} = g_j(x,t,v_1,\ldots,v_n) \quad \text{for } \; j = 1,\ldots,n \; ,$$

where the left-hand side is decoupled. In case of $f \equiv 0$ (no source term), the system (5.9) can be decoupled completely into $n$ separate PDEs.

## 5.2 Conservation laws

We introduce an example of a conservation law based on the conservation of mass. Assume that a long tube is given, which is filled with gas. The (mass) density $\rho$ and the velocity $v$ of the molecules shall be the same in each cross section of the tube. Consequently, these values depend just on a single space dimension along the tube and on the time. Let the velocity $v$ be a predetermined function, whereas the density is unknown a priori.

The mass of the gas within the space domain $[x_1, x_2]$ $(x_1 < x_2)$ in the tube at time $t$ is

$$M(x_1, x_2, t) := \int_{x_1}^{x_2} \rho(x, t) \, \mathrm{d}x.$$

The flux of mass across a point $x$ at time $t$ reads

$$\tilde{f}(\rho(x, t), v(x, t)) := \rho(x, t) \cdot v(x, t).$$

In contrast to the density $\rho > 0$, the velocity $v \in \mathbb{R}$ can be negative. The conservation of mass implies that the amount of gas in $[x_1, x_2]$ can change only by the inflow or the outflow of gas at the boundaries. It follows a first formulation of a conservation law

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{x_1}^{x_2} \rho(x, t) \, \mathrm{d}x = \rho(x_1, t)v(x_1, t) - \rho(x_2, t)v(x_2, t), \tag{5.10}$$

where an integral in space and a time derivative is involved. We obtain a pure integral form of the conservation law by an integration of (5.10) in the time interval $[t_1, t_2]$ $(0 \leq t_1 < t_2)$

$$\begin{aligned} \int_{x_1}^{x_2} \rho(x, t_2) \, \mathrm{d}x \;=\; & \int_{x_1}^{x_2} \rho(x, t_1) \, \mathrm{d}x + \int_{t_1}^{t_2} \rho(x_1, t)v(x_1, t) \, \mathrm{d}t \\ & - \int_{t_1}^{t_2} \rho(x_2, t)v(x_2, t) \, \mathrm{d}t. \end{aligned} \tag{5.11}$$

Assuming $\rho, v \in C^1$, an equivalent partial differential equation can be ob-

tained. It holds

$$\begin{aligned}\rho(x,t_2)-\rho(x,t_1) &= \int_{t_1}^{t_2}\frac{\partial}{\partial t}\rho(x,t)\,\mathrm{d}t,\\ \rho(x_2,t)v(x_2,t)-\rho(x_1,t)v(x_1,t) &= \int_{x_1}^{x_2}\frac{\partial}{\partial x}(\rho(x,t)v(x,t))\,\mathrm{d}x.\end{aligned}$$

Inserting these equalities in (5.11) yields

$$\int_{t_1}^{t_2}\int_{x_1}^{x_2}\left[\frac{\partial}{\partial t}\rho(x,t)+\frac{\partial}{\partial x}(\rho(x,t)v(x,t))\right]\,\mathrm{d}x\,\mathrm{d}t=0. \tag{5.12}$$

The integrand is continuous due to the assumption $\rho, v \in C^1$. Since the equation (5.12) holds for arbitrary intervals $[x_1, x_2]$ and $[t_1, t_2]$, the fundamental theorem of variational calculus implies

$$\rho_t + (\rho v)_x = 0. \tag{5.13}$$

We have achieved a pure differential equation of the conservation law. Remark that (5.11) and (5.13) are equivalent for smooth solutions only. The integral form (5.11) may have non-smooth or even discontinuous solutions, which cannot satisfy the differential equation (5.13).

If the velocity $v$ is not predetermined but a function in dependence on the density $v = g(\rho)$, then the PDE (5.13) exhibits the more general form of a scalar nonlinear conservation law

$$\rho_t + f(\rho)_x = 0, \tag{5.14}$$

where $f$ is a given flux function. If $v$ does not depend on the (mass) density only, then other conserved quantities have to be added to obtain a system with as many unknowns as equations. For example, we arrange

$$\begin{aligned}(\rho v)_t + (\rho v^2 + p)_x &= 0 \qquad \text{conservation of momentum}\\ E_t + (v(E+p))_x &= 0 \qquad \text{conservation of energy}\end{aligned}$$

with the momentum density $\rho v$, the energy $E$ and the pressure $p$. The pressure has to be specified as a function of $\rho, \rho v, E$ according to the physical laws of gas dynamics. For example, an ideal gas satisfies

$$p = (\gamma - 1)(E - \tfrac{1}{2}\rho v^2) \tag{5.15}$$

with a constant $\gamma \in \mathbb{R}$ like $\gamma = \frac{5}{3}$.

We obtain the Euler equations of gas dynamics, which represent a system of three conservation laws. The conserved quantities are

$$u(x,t) = \begin{pmatrix} u_1(x,t) \\ u_2(x,t) \\ u_3(x,t) \end{pmatrix} = \begin{pmatrix} \rho(x,t) \\ \rho(x,t)v(x,t) \\ E(x,t) \end{pmatrix}$$

and the corresponding flux function reads

$$f(u) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ v(E+p) \end{pmatrix} = \begin{pmatrix} u_2 \\ u_2^2/u_1 + p(u_1,u_2,u_3) \\ (u_2/u_1)(u_3 + p(u_1,u_2,u_3)) \end{pmatrix}.$$

For the pressure, it holds $p(u_1,u_2,u_3) = (\gamma-1)(u_3 - \frac{1}{2}u_2^2/u_1)$ in case of (5.15).

In general, a system of conservation laws for $m$ quantities $u : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R}^m$ with corresponding flux function $f : \mathbb{R}^m \to \mathbb{R}^m$ exhibits the differential equations

$$u_t + f(u)_x = 0. \tag{5.16}$$

An integral form like (5.10) is given by

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{x_1}^{x_2} u(x,t)\, \mathrm{d}x = f(u(x_1,t)) - f(u(x_2,t)). \tag{5.17}$$

A (pure) integral form of the conservation law (5.16) like (5.11) reads

$$\begin{aligned} \int_{x_1}^{x_2} u(x,t_2)\, \mathrm{d}x \;=\; & \int_{x_1}^{x_2} u(x,t_1)\, \mathrm{d}x \\ & + \int_{t_1}^{t_2} f(u(x_1,t))\, \mathrm{d}t - \int_{t_1}^{t_2} f(u(x_2,t))\, \mathrm{d}t, \end{aligned} \tag{5.18}$$

where the integration is done in each component separately.

### Hyperbolic systems

In the following, we assume that the flux function satisfies $f \in C^1$. This assumption is given in most of the practical cases.

**Definition 19 (hyperbolic system)** *A conservation law (5.16) is called hyperbolic, if the Jacobian matrix $\frac{\partial f}{\partial u}$ is real diagonalisable for all (relevant) $u$. The conservation law (5.16) is called strictly hyperbolic, if the system is hyperbolic and all eigenvalues of $\frac{\partial f}{\partial u}$ are pairwise different.*

In the linear case $f(u) = Au$, this definition is in agreement with Definition 18. In particular, each linearisation of a nonlinear system (5.16) is a hyperbolic system, which can be decoupled. Most of the conservation laws applied in practice are hyperbolic.

In case of several space dimensions, each space coordinate requires an own flux function. Given three space dimensions $(x, y, z)$, the conserved quantities $u : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^m$ satisfy the system

$$u_t + f(u)_x + g(u)_y + h(u)_z = 0 \tag{5.19}$$

with the flux functions $f, g, h : \mathbb{R}^m \to \mathbb{R}^m$. The system (5.19) is called hyperbolic, if all linear combinations $\alpha\frac{\partial f}{\partial u} + \beta\frac{\partial g}{\partial u} + \gamma\frac{\partial h}{\partial u}$ of the Jacobian matrices are real diagonalisable for each $\alpha, \beta, \gamma \in \mathbb{R}$ and all (relevant) values $u$. Furthermore, a general conservation law with source term reads

$$u_t + f(u)_x + g(u)_y + h(u)_z = q(x, y, z, t, u)$$

with a function $q : \mathbb{R}^3 \times \mathbb{R}_0^+ \times \mathbb{R}^m \to \mathbb{R}^m$.

#### Example: Shallow water equations

We consider a single space dimension first. Let $h(x, t)$ be the water level (height) of a river and $u(x, t)$ the mean velocity. Let the bottom of the river be planar. The one-dimensional shallow water equations are given by the conservation law

$$\frac{\partial}{\partial t}\begin{pmatrix} h \\ uh \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} uh \\ u^2h + \frac{1}{2}gh^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where $g$ is the gravitation constant. This system reflects the conservation of mass and momentum written in dependence on the water level. If the bottom of the river exhibits a profile given by a smooth function $S(x)$, then it follows a conservation law with source term

$$\frac{\partial}{\partial t}\begin{pmatrix} h \\ uh \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} uh \\ u^2h + \frac{1}{2}gh^2 \end{pmatrix} = \begin{pmatrix} 0 \\ -ghS'(x) \end{pmatrix}.$$

For a planar river bottom, it holds $S(x) \equiv$ const. and thus the above conservation law is recovered.

In case of two space dimensions, the river bottom is specified by a profile $S(x, y)$ and $h(x, y, t)$ represents the water level. Let $u(x, y, t)$ and $v(x, y, t)$ be the mean velocities in the directions $x$ and $y$, respectively. We obtain a conservation law with source term

$$\frac{\partial}{\partial t}\begin{pmatrix} h \\ uh \\ vh \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} uh \\ u^2h + \frac{1}{2}gh^2 \\ uvh \end{pmatrix} + \frac{\partial}{\partial y}\begin{pmatrix} vh \\ uvh \\ v^2h + \frac{1}{2}gh^2 \end{pmatrix} = \begin{pmatrix} 0 \\ -gh\frac{\partial S}{\partial x}(x,y) \\ -gh\frac{\partial S}{\partial y}(x,y) \end{pmatrix},$$

where two flux functions are included. The system is symmetric in $x$ and $y$. For a planar river bottom $S(x, y) \equiv$ const., it follows a system of conservation laws. It can be shown that the systems are hyperbolic. Remark that no three-dimensional shallow water equations exist, since the third space dimension is given by the dependent variable $h(x, y, t)$.

## Weak solutions

Given a conservation law in the PDE form (5.16), a corresponding solution $u$ has to be smooth to satisfy the system per definition. However, non-smooth or even discontinuous solutions are often physically reasonable. Hence we will define weak solutions of conservation laws now.

We consider the initial value problem

$$\begin{aligned} u_t + f(u)_x &= 0 \\ u(x,t) &= u_0(x) \quad \text{for } x \in \mathbb{R} \end{aligned}$$

with solution $u : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R}$. If $u \in C^1$ holds, then $u$ is called a classical solution. We always assume $f \in C^1$. We define the set of test functions

$$C_0^1 := \{\phi : \mathbb{R}^2 \to \mathbb{R},\ \phi \in C^1,\ \text{supp}(\phi) \text{ is compact}\}.$$

The equation $u_t + f(u)_x = 0$ implies that it holds

$$\phi u_t + \phi f(u)_x = 0 \qquad \text{for each } \phi \in C_0^1$$

in each point $(x,t) \in \mathbb{R} \times \mathbb{R}_0^+$. It follows

$$\int_0^\infty \int_{-\infty}^{+\infty} [\phi u_t + \phi f(u)_x] \, \mathrm{d}x \, \mathrm{d}t = 0 \qquad \text{for each } \phi \in C_0^1. \tag{5.20}$$

Remark that the integrand is continuous due to the above assumptions. Integration by parts yields

$$\begin{aligned} \int_0^\infty \phi u_t \, \mathrm{d}t &= [\phi u]_{t=0}^{t\to\infty} - \int_0^\infty \phi_t u \, \mathrm{d}t = -\phi(x,0)u(x,0) - \int_0^\infty \phi_t u \, \mathrm{d}t, \\ \int_{-\infty}^{+\infty} \phi f(u)_x \, \mathrm{d}x &= [\phi f(u)]_{x\to-\infty}^{x\to+\infty} - \int_{-\infty}^{+\infty} \phi_x f(u) \, \mathrm{d}x = -\int_{-\infty}^{+\infty} \phi_x f(u) \, \mathrm{d}x, \end{aligned}$$

since the support of $\phi$ is bounded. Inserting these equations in (5.20) results in

$$\int_0^\infty \int_{-\infty}^{+\infty} [\phi_t u + \phi_x f(u)] \, \mathrm{d}x \, \mathrm{d}t = -\int_{-\infty}^{+\infty} \phi(x,0)u(x,0) \, \mathrm{d}x \tag{5.21}$$

for each $\phi \in C_0^1$. The partial derivatives have been shifted to the smooth test functions. Now we can define a broader class of solutions.

**Definition 20 (weak solution)** *Given a conservation law $u_t + f(u)_x = 0$, a locally integrable function $u : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R}$ is a weak solution if the condition (5.21) holds for all test functions $\phi \in C_0^1$.*

On the one hand, a weak solution $u$ satisfying $u \in C^1$ is also a classical solution. On the other hand, weak solutions may exist with $u \notin C^1$. Remark that the condition (5.21) can be verified for locally integrable functions. Whereas classical solutions of initial value problems are unique, many weak solutions of a conservation law often exist. The generalisation to systems of conservation laws is straightforward by components.

Moreover, the integral form (5.18) represents the conservation law, where also just integrable functions are required. The condition (5.18) can be evaluated for functions $u \in \mathcal{U}$ with

$$\begin{aligned} \mathcal{U} := \{u : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R} : u(\cdot,t) &\text{ piecewise continuous for each } t \geq 0, \\ f(u(x,\cdot)) &\text{ piecewise continuous for each } x \in \mathbb{R}\}, \end{aligned}$$

for example. The condition (5.21) requires functions from the space

$$\mathcal{V} := \{u : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R} : u \text{ locally integrable in } \mathbb{R} \times \mathbb{R}_0^+\}.$$

Assuming $u \in \mathcal{U} \cap \mathcal{V}$, it can be shown that $u$ satisfies (5.21) for all test functions $\phi \in C_0^1$ if and only if $u$ fulfills (5.18) for arbitrary boundaries $x_1 < x_2$, $0 \leq t_1 < t_2$.

### Characteristic curves

As a motivation, we consider an initial value problem of the linear advection equation

$$\begin{aligned} u_t + a u_x &= 0 \qquad (a \text{ constant}) \\ u(x,t) &= u_0(x) \quad \text{for } x \in \mathbb{R}. \end{aligned}$$

The corresponding solution is

$$u(x,t) = u_0(x - at).$$

For $u_0 \in C^1$ a classical solution results, whereas $u_0 \notin C^1$ yields a weak solution provided that $u_0$ is integrable. We define the characteristic curves of the advection equation as a family of parallel straight lines

$$x(t) := \xi + at$$

with the parameter $\xi \in \mathbb{R}$. In particular, it holds

$$\tfrac{\mathrm{d}}{\mathrm{d}t} x(t) = a.$$

We obtain

$$u(x(t),t) = u_0(x(t) - at) = u_0(\xi + at - at) = u_0(\xi),$$

i.e., the solution $u$ is constant along each characteristic curve.

Now we investigate a scalar nonlinear conservation law. Assuming $u \in C^1$, it holds the equivalence

$$u_t + f(u)_x = 0 \quad \Leftrightarrow \quad u_t + f'(u) u_x = 0,$$

where the right-hand equation represents the quasilinear form of the conservation law. Let initial conditions $u(x,0) = u_0(x)$ be given. According to the linear advection, we define characteristic curves.

**Definition 21** *Considering a scalar conservation law $u_t + f(u)_x = 0$, the corresponding characteristic curves are the solutions of the ordinary differential equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f'(u(x(t),t)) \tag{5.22}$$

*for a given classical or weak solution $u$.*

Assuming a classical solution $u \in C^1$, we conclude

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}u(x(t),t) &= \frac{\partial}{\partial t}u(x(t),t) + (\frac{\partial}{\partial x}u(x(t),t))\frac{\mathrm{d}}{\mathrm{d}t}x(t)\\ &= \frac{\partial}{\partial t}u(x(t),t) + (\frac{\partial}{\partial x}u(x(t),t))f'(u(x(t),t))\\ &= \frac{\partial}{\partial t}u(x(t),t) + \frac{\partial}{\partial x}f(u(x(t),t)) \;=\; 0.\end{aligned}$$

It follows that the solution is constant along each characteristic curve, i.e.,

$$u(x(t),t) = u(x(0),0) = u_0(x(0)) \quad \text{for all } \; t \geq 0.$$

Consequently, the ODE (5.22) can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f'(u(x(0),0)) = f'(u_0(x(0))).$$

Since the right-hand side of the ODE is constant, the characteristic curves are straight lines again. However, the straight lines are not parallel in general. The information from the initial values $u(x,0) = u_0(x)$ is transported along the characteristic curves with finite speed.

In case of linear hyperbolic systems $u_t + Au_x = 0$, it holds $A = SDS^{-1}$. Thus system can be decoupled into the the separate scalar equations

$$\frac{\partial v_j}{\partial t} + \lambda_j \frac{\partial v_j}{\partial x} = 0 \quad \text{for } \; j = 1, \ldots, m$$

with $v := S^{-1}u$. Hence $m$ families of characteristic curves exist, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \lambda_j \quad \Rightarrow \quad x(t) = \xi + \lambda_j t \quad \text{for } \; j = 1, \ldots, m$$

with parameters $\xi \in \mathbb{R}$. The transport of information proceeds along these $m$ families of characteristic curves in case of hyperbolic systems.

**Burgers' equation**

The Burgers' equation represents a benchmark problem for scalar nonlinear conservation laws, namely

$$u_t + \left(\tfrac{1}{2}u^2\right)_x = 0 \tag{5.23}$$

Thus the flux function $f(u) = \frac{1}{2}u^2$ is chosen relatively simple. The equivalent quasilinear formulation for $u \in C^1$ reads

$$u_t + uu_x = 0.$$

Hence the corresponding characteristic curves are

$$x(t) = \xi + u(x(0), 0)t \quad \text{for } \xi \in \mathbb{R}$$

We investigate a Cauchy problem $u(x, 0) = u_0(x)$. If $u_0 \in C^1$ holds, then a classical solution $u$ ($u \in C^1$) exists in each space interval $I := [a, b]$ for $0 < t < T_I$. However, the final time $T_I$ may be small.

For example, we consider $\xi_1 < \xi_2$ and the corresponding characteristic curves

$$x_1(t) = \xi_1 + u_0(\xi_1)t, \quad x_2(t) = \xi_2 + u_0(\xi_2)t.$$

Assuming $u_0(\xi_1) \neq u_0(\xi_2)$, these straight lines intersect at the time

$$T = \frac{\xi_1 - \xi_2}{u_0(\xi_2) - u_0(\xi_1)}. \tag{5.24}$$

If $u_0(\xi_1) > u_0(\xi_2)$ holds, then it follows $T > 0$. A classical solution is constant along each characteristic curve, which implies $u(x_1(T), T) = u_0(\xi_1)$ and $u(x_2(T), T) = u_0(\xi_2)$ in contrast to $u_0(\xi_1) \neq u_0(\xi_2)$ Consequently, a classical solution of (5.23) does not exist for $t \geq T$ or earlier. Nevertheless, weak solutions of the Cauchy problem exist, see Definition 20.

Since weak solutions typically exhibit discontinuities, we consider a corresponding benchmark problem: the Riemann problem. A Riemann problem consists of some conservation law together with initial conditions

$$u(x, t = 0) = \begin{cases} u_l, & x < 0 \\ u_r, & x > 0 \end{cases} \tag{5.25}$$

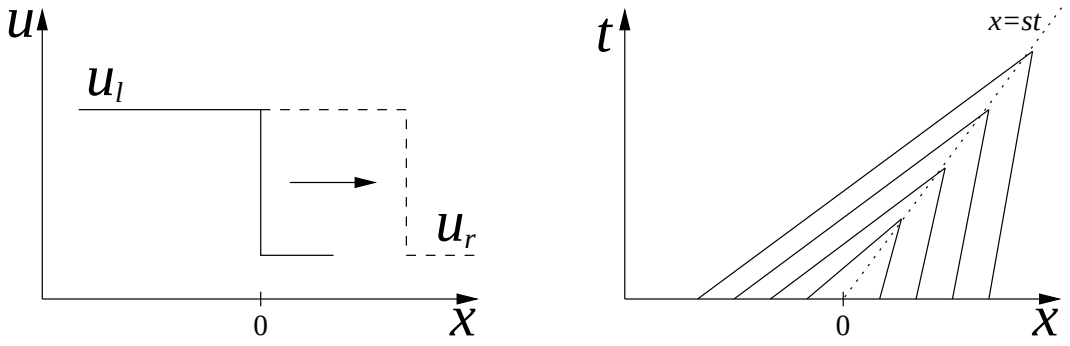



Figure 18: Shock wave for Burgers' equation with $u_l > u_r$.

for constants $u_l, u_r \in \mathbb{R}$ satisfying $u_l \neq u_r$. Hence a single discontinuity appears at $x = 0$. Depending on the choice of the constants, two cases have to be discussed.

**Case 1:** $u_l > u_r$
A unique weak solution exists given by

$$u(x,t) = \begin{cases} u_l, & x < st \\ u_r, & x > st \end{cases} \quad \text{with} \quad s = \tfrac{1}{2}(u_l + u_r). \tag{5.26}$$

Solutions of this type a called *shock waves*. The discontinuity from the initial values at $x = 0$ is transported in time with the shock speed $s$. The value of the shock speed will be derived in the next subsection. The corresponding characteristic curves are illustrated in Fig. 18. The characteristics enter the shock, which indicates a physically reasonable solution.

**Case 2:** $u_l < u_r$
The shock wave (5.26) represents a weak solution of (5.23) again. The characteristic curves are shown in Fig. 19. This weak solution is not physically reasonable, since the characteristics leave the shock now. Nevertheless, an infinite number of weak solutions exists in this case. The physically reasonable solution is a *rarefaction wave* given by

$$u(x,t) = \begin{cases} u_l, & x < u_l t, \\ \frac{x}{t}, & u_l t \leq x \leq u_r t, \\ u_r, & x > u_r t. \end{cases}$$

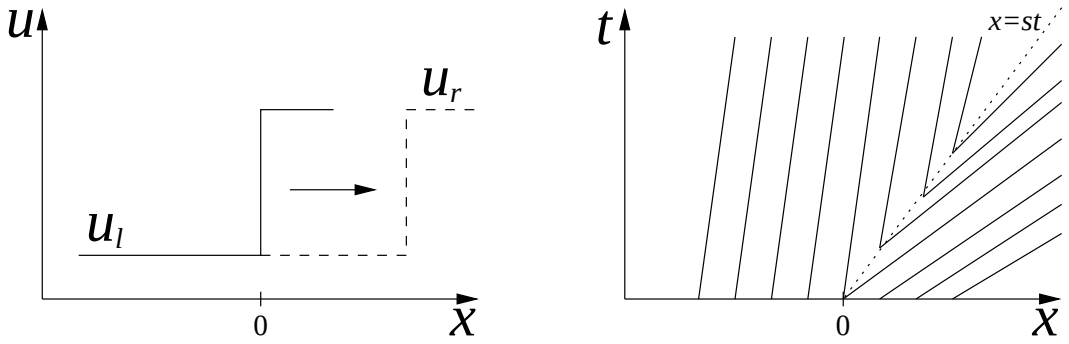



Figure 19: Shock wave for Burgers' equation with $u_l < u_r$.

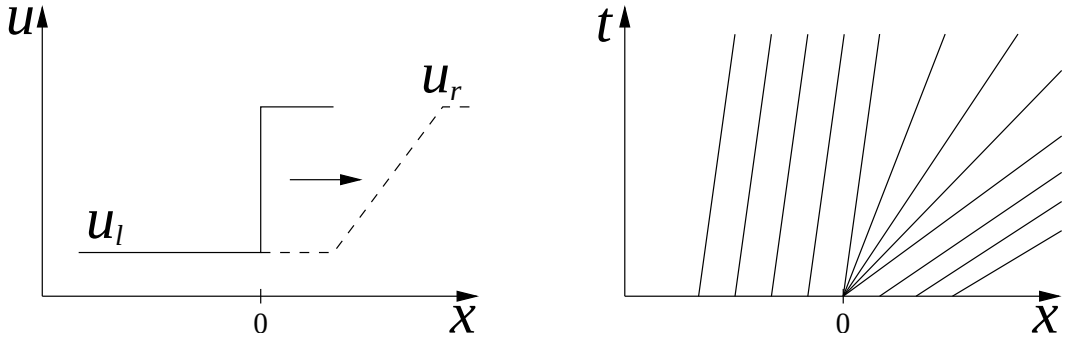



Figure 20: Rarefaction wave for Burgers' equation with $u_l < u_r$.

Moreover, this weak solution is not smooth but continuous. Fig. 20 illustrates the corresponding characteristic curves.

In case of nonlinear systems of conservation laws, a Riemann problem (5.25) typically implies shock waves as well as rarefaction waves in each component.

### Shock speed

For systems of conservation laws $u_t + f(u)_x = 0$, shock waves

$$u(x,t) = \begin{cases} u_l, & x < st \\ u_r, & x > st \end{cases} \tag{5.27}$$

with $u_l, u_r \in \mathbb{R}^m$ often represent a weak solution. The corresponding shock speed $s \in \mathbb{R}$ has to be determined appropriately. For fixed $t > 0$, we can choose $M > 0$ sufficiently large such that $u(x,t) = u_l$ holds for all $x \leq -M$

and $u(x,t) = u_r$ holds for all $x \geq M$, since the transport of information is finite in hyperbolic equations. On the one hand, the general relation (5.17) yields

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{-M}^{M} u(x,t)\, \mathrm{d}x = f(u(-M,t)) - f(u(M,t)) = f(u_l) - f(u_r).$$

On the other hand, the specific solution (5.27) implies

$$\int_{-M}^{M} u(x,t)\, \mathrm{d}x = (M + st)u_l + (M - st)u_r$$

and thus

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{-M}^{M} u(x,t)\, \mathrm{d}x = s(u_l - u_r).$$

It follows the *Rankine-Hugoniot condition*

$$f(u_l) - f(u_r) = s(u_l - u_r). \tag{5.28}$$

For systems of conservation laws, the condition (5.28) represents $m$ equations for the scalar shock speed $s$. Hence this condition will not be satisfied for arbitrary $u_l, u_r$ in general. For linear systems with $f(u) = Au$ with $A \in \mathbb{R}^{m \times m}$, the condition (5.28) is equivalent to

$$A(u_l - u_r) = s(u_l - u_r).$$

Thus $u_l - u_r$ has to be an eigenvector of $A$ and the shock speed $s$ results to the corresponding eigenvalue.

In case of the Riemann problem (5.25) for scalar conservation laws, the shock wave (5.27) always represents a weak solution provided that

$$s = \frac{f(u_l) - f(u_r)}{u_l - u_r}.$$

Strictly speaking, we have just shown that the criterion (5.28) of Rankine-Hugoniot represents a necessary condition. It can be shown that the condition is also sufficient. In case of the Burgers' equation (5.23), the shock speed reads

$$s = \frac{\frac{1}{2}u_l^2 - \frac{1}{2}u_r^2}{u_l - u_r} = \frac{\frac{1}{2}(u_l + u_r)(u_l - u_r)}{u_l - u_r} = \frac{1}{2}(u_l + u_r),$$

which has been already used in (5.26).

## 5.3 Numerical methods for linear systems

In this section, we discuss finite difference methods for linear hyperbolic systems. The linear case illustrates the typical behaviour of the numerical techniques, which are generalised to the nonlinear case in Section 5.4. The solution of initial value problems of linear hyperbolic systems can be determined exactly by decoupling the equations, i.e., without using numerical methods.

### Preliminaries

We consider a linear system of conservation laws

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0 \tag{5.29}$$

with solution $u : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R}^m$ and a constant matrix $A \in \mathbb{R}^{m\times m}$. As initial conditions, let a Cauchy problem

$$u(x, t = 0) = u_0(x) \qquad \text{for } \; x \in \mathbb{R} \tag{5.30}$$

be given with a predetermined function $u_0 : \mathbb{R} \to \mathbb{R}^m$.

To construct finite difference methods, we introduce a grid in the domain of dependence. Let $h = \Delta x$ and $k = \Delta t$ be the step sizes in space and time, respectively. We consider the grid points

$$(x_j, t_n) := (jh, nk) \quad \text{for } \; j \in \mathbb{Z}, \;\; n \in \mathbb{N}_0.$$

We also apply the intermediate points

$$x_{j+\frac{1}{2}} := x_j + \tfrac{h}{2} = \left(j + \tfrac{1}{2}\right) h.$$

Let $u_j^n := u(x_j, t_n)$ be the evaluations of a classical solution in the grid points. We want to determine approximations $U_j^n \in \mathbb{R}^m$ for $u_j^n$. In case of a weak solution, the pointwise evaluation $u_j^n$ is not well-defined in general. Alternatively, the value $U_j^n$ can be seen as an approximation of the cell average

$$\bar{u}_j^n := \frac{1}{h}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_n)\, \mathrm{d}x. \tag{5.31}$$

The integral form of the conservation law describes the evolution of the cell averages in time. The initial values $U_j^0$ are defined either pointwise by $u_0(x_j)$ or as cell averages $\bar{u}_j^0$. A comparison to the exact solution is feasible by using the function

$$U_k(x,t) := U_j^n \quad \text{for } (x,t) \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \times [t_n, t_{n+1}). \tag{5.32}$$

This piecewise constant function exhibits just the subscript $k$, since the ratio $\frac{k}{h}$ of the step sizes is assumed to be constant. Thus $k \to 0$ is equivalent to $h \to 0$.

A finite difference method determines the approximation $U_j^1$ in the first time layer by using the initial values $U_j^0$. Successively, a one-stage method applies the data $U_j^n$ to obtain the approximations $U_j^{n+1}$. In a multi-stage method with $l+1$ stages, the data $U_j^{n-l}, \ldots, U_j^{n-1}, U_j^n$ yield the approximations $U_j^{n+1}$. However, we consider only one-stage methods in the following, since multi-stage methods are not used in practice for conservation laws.

The Cauchy problem (5.29), (5.30) defines a solution in the complete domain of dependence $\mathbb{R} \times \mathbb{R}^+$. In a numerical method, we have to choose a bounded domain. Let the interval $a \le x \le b$ be given in space. Boundary conditions are required at the boundaries of this finite interval. In case of periodic initial conditions $u_0$ with period $b-a$, periodic boundary conditions

$$u(a,t) = u(b,t) \quad \text{for all } t \ge 0$$

are reasonable. However, this situation is rather seldom in practice. Alternatively, a Riemann problem (5.25) implies the conditions

$$u(a,t) = u_l, \quad u(b,t) = u_r \quad \text{for } 0 \le t \le T \tag{5.33}$$

provided that the boundaries are sufficiently far away from the initial discontinuity at $x = 0$, since the information travels with finite speed. In particular, the boundary conditions (5.33) are feasible with $u_l = u_r = 0$ in case of initial data $u_0$ with compact support.

### Construction of finite difference methods

Many finite difference methods are feasible for solving linear hyperbolic systems (5.29). Often the methods result from replacing the partial derivatives by difference formulas. For example, using the explicit Euler scheme in time and a symmetric discretisation in space yields

$$\tfrac{1}{k}\left(U_j^{n+1} - U_j^n\right) + A\tfrac{1}{2h}\left(U_{j+1}^n - U_{j-1}^n\right) = 0 \tag{5.34}$$

and thus

$$U_j^{n+1} = U_j^n - \tfrac{k}{2h}A\left(U_{j+1}^n - U_{j-1}^n\right). \tag{5.35}$$

However, the method (5.35) is unstable, which can be shown by the concept of von-Neumann. In contrast, a discretisation via the implicit Euler scheme results in a stable method, i.e.,

$$\tfrac{1}{k}\left(U_j^{n+1} - U_j^n\right) + A\tfrac{1}{2h}\left(U_{j+1}^{n+1} - U_{j-1}^{n+1}\right) = 0. \tag{5.36}$$

If $N$ grid points are given in space, it follows a linear system with $mN$ algebraic equations for the unknown approximations. Thus a significant computational effort appears in comparison to an explicit method like (5.34).

For parabolic PDEs (like the heat equation), we obtain significant restrictions on the step sizes in explicit methods by the stability. In case of implicit methods, these restrictions are not given. For hyperbolic systems, the same situation appears qualitatively. However, the conditions on the step sizes are less strong in the explicit techniques. It follows that explicit methods are more efficient than implicit methods in case of hyperbolic PDEs. Moreover, explicit methods mimic the finite speed of the transport of information in hyperbolic systems. Consequently, we consider just explicit techniques in the following.

The explicit finite difference method (5.35) can be modified into a stable method via a substitution of the central approximation $U_j^n$ by the arithmetic mean of the neighbouring approximations. It follows the *Lax-Friedrichs method*

$$U_j^{n+1} = \tfrac{1}{2}\left(U_{j-1}^n + U_{j+1}^n\right) - \tfrac{k}{2h}A\left(U_{j+1}^n - U_{j-1}^n\right). \tag{5.37}$$

A simpler choice of the difference formula for the space derivative yields

$$\tfrac{1}{k}\left(U_j^{n+1} - U_j^n\right) + A\tfrac{1}{h}\left(U_j^n - U_{j-1}^n\right) = 0$$

and thus

$$U_j^{n+1} = U_j^n - \tfrac{k}{h} A \left( U_j^n - U_{j-1}^n \right). \tag{5.38}$$

Alternatively, a similar scheme is obtained by

$$\tfrac{1}{k} \left( U_j^{n+1} - U_j^n \right) + A \tfrac{1}{h} \left( U_{j+1}^n - U_j^n \right) = 0$$

and thus

$$U_j^{n+1} = U_j^n - \tfrac{k}{h} A \left( U_{j+1}^n - U_j^n \right). \tag{5.39}$$

The methods (5.38) and (5.39) are called the *upwind methods* due to the applications in gas dynamics. However, a necessary condition for the stability of the technique (5.38) is that all eigenvalues of $A$ are non-negative. Likewise, the scheme (5.39) requires that all eigenvalues of $A$ are non-positive.

Furthermore, methods of higher order can be constructed via a Taylor expansion. Assuming a classical solution $u \in C^3$ in time, we obtain

$$u(x, t+k) = u(x,t) + k u_t(x,t) + \tfrac{1}{2} k^2 u_{tt}(x,t) + \mathcal{O}(k^3). \tag{5.40}$$

The linear conservation law (5.29) allows for the substitutions

$$u_t = -A u_x, \quad u_{tt} = -A u_{xt} = -A u_{tx} = -A(-A u_x)_x = A^2 u_{xx}. \tag{5.41}$$

It follows

$$u(x, t+k) = u(x,t) - k A u_x(x,t) + \tfrac{1}{2} k^2 A^2 u_{xx}(x,t) + \mathcal{O}(k^3).$$

We replace the derivatives in space by symmetric difference formulas of second order, which results in the *Lax-Wendroff method*

$$U_j^{n+1} = U_j^n - \tfrac{k}{2h} A (U_{j+1}^n - U_{j-1}^n) + \tfrac{k^2}{2h^2} A^2 (U_{j+1}^n - 2U_j^n + U_{j-1}^n). \tag{5.42}$$

A one-sided discretisation of the space derivatives yields the *Beam-Warming method*

$$U_j^{n+1} = U_j^n - \tfrac{k}{2h} A (3U_j^n - 4U_{j-1}^n + U_{j-2}^n) + \tfrac{k^2}{2h^2} A^2 (U_j^n - 2U_{j-1}^n + U_{j-2}^n). \tag{5.43}$$

Likewise, methods of higher order can be derived.

Table 1: Finite difference methods for linear systems $u_t + Au_x = 0$.

| name | formula |
|---|---|
| expl. Euler | $U_j^{n+1} = U_j^n - \frac{k}{2h}A(U_{j+1}^n - U_{j-1}^n)$ |
| impl. Euler | $U_j^{n+1} = U_j^n - \frac{k}{2h}A(U_{j+1}^{n+1} - U_{j-1}^{n+1})$ |
| upwind (left-hand) | $U_j^{n+1} = U_j^n - \frac{k}{h}A(U_j^n - U_{j-1}^n)$ |
| upwind (right-hand) | $U_j^{n+1} = U_j^n - \frac{k}{h}A(U_{j+1}^n - U_j^n)$ |
| Lax-Friedrichs | $U_j^{n+1} = \frac{1}{2}(U_{j-1}^n + U_{j+1}^n) - \frac{k}{2h}A(U_{j+1}^n - U_{j-1}^n)$ |
| Lax-Wendroff | $U_j^{n+1} = U_j^n - \frac{k}{2h}A(U_{j+1}^n - U_{j-1}^n) + \frac{k^2}{2h^2}A^2(U_{j+1}^n - 2U_j^n + U_{j-1}^n)$ |
| Beam-Warming | $U_j^{n+1} = U_j^n - \frac{k}{2h}A(3U_j^n - 4U_{j-1}^n + U_{j-2}^n) + \frac{k^2}{2h^2}A^2(U_j^n - 2U_{j-1}^n + U_{j-2}^n)$ |

The above techniques are constructed using the differential equations. However, we are also interested in weak solutions, which satisfy the corresponding integral formulation. It is surprising that methods based on the differential equations are often suitable also for weak solutions, which are neither smooth nor continuous.

## Consistency, Stability and Convergence

We analyse the convergence of explicit finite difference methods now. This analysis is similar to the case of ordinary differential equations.

We require a specific notation. Let the application of a one-stage method be given by a operator $\mathcal{H}_k$, i.e.,

$$U^{n+1} = \mathcal{H}_k(U^n), \qquad U_j^{n+1} = \mathcal{H}_k(U^n; j),$$

where $U^n$ is the vector including all approximations $U_j^n \in \mathbb{R}^m$. For example, the method (5.35) implies the operator

$$\mathcal{H}_k(U^n; j) = U_j^n - \tfrac{k}{2h}A\left(U_{j+1}^n - U_{j-1}^n\right). \tag{5.44}$$

The difference operator $\mathcal{H}_k$ can also be applied to a function $v : \mathbb{R} \to \mathbb{R}^m$. Thereby, the difference formula is centered around an arbitrary point $x \in \mathbb{R}$. In case of (5.35), the evaluation $\mathcal{H}_k(v)$ is defined by

$$\mathcal{H}_k(v; x) = (\mathcal{H}_k(v))(x) = v(x) - \tfrac{k}{2h} A \left( v(x+h) - v(x-h) \right).$$

In particular, the difference operator can be applied to the piecewise constant function (5.32). It holds

$$U_k(x, t+k) = \mathcal{H}_k(U_k(\cdot, t); x), \tag{5.45}$$

where the evaluation of the operator coincides for the discrete data and the constructed function. Due to this identity, we apply the same symbol in our notation.

We consider just linear methods in this chapter, i.e., the operators $\mathcal{H}_k$ are linear. Hence it holds

$$\mathcal{H}_k \left( \alpha U^n + \beta V^n \right) = \alpha \mathcal{H}_k \left( U^n \right) + \beta \mathcal{H}_k \left( V^n \right) \quad (\alpha, \beta \in \mathbb{R}).$$

The operator is also well-defined in case of an infinite number of grid points, where $U^n \in \mathbb{R}^\infty$ holds. In case of a grid with $N$ points, it follows $U^n \in \mathbb{R}^{mN}$. Thus the operator is given by a matrix of size $mN \times mN$. The method can be written as $U^{n+1} = \mathcal{H}_k U^n$ using the same symbol for the operator as well as the matrix.

We are interested in the error of the approximations with respect to the exact solution. In case of classical solutions, the *global error* is defined by

$$E_j^n := U_j^n - u_j^n.$$

In case of weak solutions, we consider the errors with respect to the cell averages, i.e.,

$$\bar{E}_j^n := U_j^n - \bar{u}_j^n.$$

We define the error function

$$E_k(x, t) := U_k(x, t) - u(x, t) \tag{5.46}$$

using (5.32). It follows the pointwise evaluations $E_j^n$ in $(x_j, t_n)$ and the corresponding cell averages $\bar{E}_j^n$. The *convergence* of the method is defined via the global error.

**Definition 22** *A method $U^{n+1} = \mathcal{H}_k(U^n)$ with the global error (5.46), where $U_k(x,t)$ is computed recursively using $\mathcal{H}_k$ and the initial data $u_0$, is called convergent with respect to a norm $\|\cdot\|$, if it holds*

$$\lim_{k\to 0} \|E_k(\cdot,t)\| = 0 \tag{5.47}$$

*for each $t \geq 0$ and all initial values $u_0$ (in some class).*

The convergence of a method depends on the choice of the norm. For simplicity, we consider a scalar function $v : \mathbb{R} \to \mathbb{R}$. In case of a classical (or at least continuous) solution, the convergence is optimal in the maximum norm

$$\|v\|_\infty = \sup\{|v(x)| : x \in \mathbb{R}\}.$$

In case of weak solutions, the integral form of the conservation law suggests the integral norm

$$\|v\|_1 = \int_{-\infty}^{+\infty} |v(x)| \, \mathrm{d}x. \tag{5.48}$$

We apply the norm $\|\cdot\|_1$ and write simply $\|\cdot\|$ in the following. The integral norm can be applied to the discrete data via

$$\|U^n\|_1 := \|U_k(\cdot,t_n)\|_1 = h \sum_{j=-\infty}^{+\infty} |U_j^n| \qquad (U_j^n \in \mathbb{R}). \tag{5.49}$$

The convergence (5.47) is defined for a vector-valued error function (5.46). The method is convergent if and only if the global error converges to zero in each component. Hence we consider the norm component-wise. A corresponding vector norm can be applied. For example, the maximum norm yields

$$\lim_{k\to 0} \|E_k(\cdot,t)\| = 0 \quad :\Leftrightarrow \quad \lim_{k\to 0} \max\{\|(E_k(\cdot,t))_p\|_1 : p = 1,\ldots,m\} = 0.$$

Since the global error cannot be estimated directly, we require further concepts to analyse the convergence. We insert the exact solution of the conservation law into the formula of a finite difference method. For example, the Lax-Friedrichs method (5.37) can be written in the form

$$\tfrac{1}{k}\left(U_j^{n+1} - \tfrac{1}{2}\left(U_{j-1}^n + U_{j+1}^n\right)\right) + \tfrac{1}{2h} A \left(U_{j+1}^n - U_{j-1}^n\right) = 0. \tag{5.50}$$

The *local error* follows from an evaluation at an exact solution, i.e.,

$$\begin{aligned} L_k(x,t) &:= \tfrac{1}{k}\left(u(x,t+k) - \tfrac{1}{2}\left(u(x-h,t)+u(x+h,t)\right)\right) \\ &\quad + \tfrac{1}{2h}A\left(u(x+h,t) - u(x-h,t)\right). \end{aligned}$$

If the solution is sufficiently smooth, a Taylor expansion yields ($u \equiv u(x,t)$)

$$\begin{aligned} L_k(x,t) &= \tfrac{1}{k}\left(\left(u + ku_t + \tfrac{1}{2}k^2u_{tt} + \mathcal{O}(k^3)\right) - \left(u + \tfrac{1}{2}h^2u_{xx} + \mathcal{O}(h^4)\right)\right) \\ &\quad + \tfrac{1}{2h}A\left(2hu_x + \mathcal{O}(h^3)\right) \\ &= u_t + Au_x + \tfrac{1}{2}\left(ku_{tt} - \tfrac{h^2}{k}u_{xx}\right) + \mathcal{O}(k^2), \end{aligned} \tag{5.51}$$

where $r := \frac{k}{h}$ is assumed to be constant. Since $u$ satisfies the differential equations (5.29), we apply the substitutions (5.41) and obtain

$$L_k(x,t) = \tfrac{1}{2}k\left(A^2 - \tfrac{h^2}{k^2}I\right)u_{xx}(x,t) + \mathcal{O}(k^2). \tag{5.52}$$

It follows the pointwise convergence

$$\lim_{k\to 0} L_k(x,t) = 0 \quad \text{for each } \; x \in \mathbb{R} \; \text{ and each } \; t \geq 0.$$

The partial derivatives of the solution can be bounded by the derivatives of the initial values $u_0$. For each component $p = 1, \ldots, m$, we obtain an estimate

$$|(L_k(x,t))_p| \leq C_p k \quad \text{for all } \; k < k_0$$

and arbitrary $x$, i.e., also in the maximum norm. The constants $C_p$ depend just on the initial data $u_0$ of the solution. If the support of the initial values is compact, then it follows a finite integral norm for each $t \geq 0$ and it holds

$$\|(L_k(\cdot,t))_p\|_1 \leq \tilde{C}_p k \quad \text{for all } \; k < \tilde{k}_0,$$

where the constants $\tilde{C}_p$ depend on the initial data.

Likewise, the consistency of a general one-stage method can be defined.

**Definition 23** *A finite difference method* $U^{n+1} = \mathcal{H}_k(U^n)$ *with the local error*

$$L_k(x,t) = \tfrac{1}{k}\left[u(x,t+k) - \mathcal{H}_k(u(\cdot,t);x)\right] \tag{5.53}$$

*is called consistent, if it holds*

$$\lim_{k\to 0} \|L_k(\cdot,t)\| = 0$$

*for each* $t \geq 0$ *and all initial values* $u_0$ *(in same class). A finite difference method is consistent of order* $q$*, if for each initial values with compact support and each* $T > 0$ *some constants* $C_L \geq 0$ *and* $k_0 > 0$ *exist such that*

$$\|L_k(\cdot,t)\| \leq C_L k^q \quad \textit{for all } \; k < k_0, \;\; t \leq T. \tag{5.54}$$

Again the consistency depends on the choice of the norm. We consider the integral norm $\|\cdot\|_1$. The application to vector-valued functions is done as for the global error.

For a consistent method, the local error can be reduced by choosing sufficiently small step sizes. However, the consistency alone is not sufficient for the convergence of the method, i.e., a reduction of the global error.

To guarantee the convergence of a method, we require the *stability* of the finite difference scheme. The local error (5.53) implies

$$u(x,t+k) = \mathcal{H}_k(u(\cdot,t);x) + kL_k(x,t). \tag{5.55}$$

Using (5.45), the linearity of the operator yields a recursion for the global error

$$E_k(x,t+k) = \mathcal{H}_k(E_k(\cdot,t);x) - kL_k(x,t). \tag{5.56}$$

The global error at time $t+k$ consists to two parts: a contribution of the global error in the previous time layer and a new contribution from the current local error. We solve the recursion and obtain at time $t_n$

$$E_k(\cdot,t_n) = \mathcal{H}_k^n(E_k(\cdot,0)) - k\sum_{i=1}^{n} \mathcal{H}_k^{n-i}(L_k(\cdot,t_{i-1})), \tag{5.57}$$

where $\mathcal{H}_k^i$ denotes the $i$-fold application of the operator $\mathcal{H}_k$.

The global error (5.57) includes an error from the initial values, which results from changing the initial function $u_0$ into discrete data. However, this error vanishes in the limit case $h \to 0$. The global error is bounded at time $t_n$ provided that the local errors in (5.57) are not accumulated by the evaluations of the operators $\mathcal{H}_k$. It follows the concept of stability.

**Definition 24** *The finite difference method $U^{n+1} = \mathcal{H}_k(U^n)$ is called stable (according to Lax-Richtmyer), if for each $T \geq 0$ it exist constants $C_S \geq 0$ and $k_0 > 0$ satisfying*

$$\|\mathcal{H}_k^n\| \leq C_S \quad \textit{for all } \; nk \leq T, \;\; k < k_0. \tag{5.58}$$

The definition of the stability involves the norm of the operators $\mathcal{H}_k^n$. For a linear operator $\mathcal{G} : \mathbb{R}^\infty \to \mathbb{R}^\infty$, the operator norm is given by

$$\|\mathcal{G}\| := \sup_{V \neq 0} \frac{\|\mathcal{G}(V)\|}{\|V\|} \qquad (V \in \mathbb{R}^\infty)\,,$$

where we apply the norm (5.49) on $\mathbb{R}^\infty$. This norm is generalised to a system of equations by considering the maximum again. In case of $N$ grid points in space, an arbitrary vector norm can be used on $\mathbb{R}^{mN}$.

Since it holds $\|\mathcal{H}_k^n\| \leq \|\mathcal{H}_k\|^n$, the condition $\|\mathcal{H}_k\| \leq 1$ is sufficient for the stability of the method. Some growth of the operator norm is allowed. For example, the condition

$$\|\mathcal{H}_k\| \leq 1 + \alpha k \quad \text{for all } \; k < k_0$$

implies

$$\|\mathcal{H}_k^n\| \leq \|\mathcal{H}_k\|^n \leq (1 + \alpha k)^n \leq \mathrm{e}^{\alpha k n} \leq \mathrm{e}^{\alpha T} \quad \text{for all } \; nk \leq T,$$

i.e., the stability property (5.58).

We reconsider the Lax-Friedrichs method (5.37) in case of the linear advection equation $u_t + a u_x = 0$, i.e.,

$$U_j^{n+1} = \tfrac{1}{2}\left(U_{j-1}^n + U_{j+1}^n\right) - \tfrac{ak}{2h}\left(U_{j+1}^n - U_{j-1}^n\right).$$

The corresponding norm (5.49) on $\mathbb{R}^\infty$ becomes

$$\begin{aligned}\|U^{n+1}\| &= h\sum_{j=-\infty}^{+\infty}\left|U_j^{n+1}\right| \\ &\le \tfrac{h}{2}\left[\sum_{j=-\infty}^{+\infty}\left|1-\tfrac{ak}{h}\right|\cdot\left|U_{j+1}^n\right| + \sum_{j=-\infty}^{+\infty}\left|1+\tfrac{ak}{h}\right|\cdot\left|U_{j-1}^n\right|\right].\end{aligned}$$

The condition

$$\left|\frac{ak}{h}\right| \le 1 \tag{5.59}$$

guarantees that the coefficients are not negative. We estimate

$$\begin{aligned}\|U^{n+1}\| &\le \tfrac{h}{2}\left[\left(1-\tfrac{ak}{h}\right)\textstyle\sum_j\left|U_{j+1}^n\right| + \left(1+\tfrac{ak}{h}\right)\sum_j\left|U_{j-1}^n\right|\right] \\ &= \tfrac{1}{2}\left[\left(1-\tfrac{ak}{h}\right)\|U^n\| + \left(1+\tfrac{ak}{h}\right)\|U^n\|\right] = \|U^n\|.\end{aligned}$$

Thus the condition (5.59) implies $\|\mathcal{H}_k\| \le 1$ and the method is stable. The derivation of the stability condition uses the same strategy as in the direct estimation for parabolic equations in Sect. 3.3.

A linear hyperbolic system (5.29) can be decoupled into $m$ linear advection equations by a linear transformation. It follows that the Lax-Friedrichs method is stable in case of

$$\left|\frac{\lambda_p k}{h}\right| \le 1 \quad \text{for all } \; p = 1,\ldots,m\;, \tag{5.60}$$

where $\lambda_p \in \mathbb{R}$ represent the eigenvalues of the matrix $A$.

The stability concept of Lax-Richtmyer agrees to the Lipschitz-continuous dependence of the numerical solution with respect to the initial values. Let $U^n, V^n$ be the approximations computed from $U_0, V_0$, respectively. It follows

$$\begin{aligned}\|U^n - V^n\| &= \|\mathcal{H}_k^n(U^0) - \mathcal{H}_k^n(V^0)\| = \|\mathcal{H}_k^n(U^0 - V^0)\| \\ &\le \|\mathcal{H}_k^n\|\cdot\|U^0 - V^0\| \le C_S\|U^0 - V^0\|.\end{aligned}$$

Alternatively, the stability concept of von-Neumann can be carried over to the case of linear hyperbolic equations.

As for ordinary differential equations, the convergence is equivalent to the stability in case of consistent methods. This result is given by the equivalence theorem of Lax.

**Theorem 15 (Lax)** *Considering the linear hyperbolic system (5.29), let the linear method $U^{n+1} = \mathcal{H}_k(U^n)$ be consistent. It follows that stability (according to Lax-Richtmyer) and convergence of the method are equivalent.*

The proof can be found in, for example, J.C. Strikwerda: Finite Difference Schemes and Partial Differential Equations. Wadsworth & Brooks/Cole, 1989.

We just show the more important part, namely that the stability implies the convergence. The equation (5.57) implies

$$\|E_k(\cdot, t_n)\| \leq \|\mathcal{H}_k^n\| \cdot \|E_k(\cdot, 0)\| + k \sum_{i=1}^{n} \|\mathcal{H}_k^{n-i}\| \cdot \|L_k(\cdot, t_{i-1})\|.$$

The stability condition (5.58) allows for the estimate

$$\|E_k(\cdot, t_n)\| \leq C_S \left( \|E_k(\cdot, 0)\| + k \sum_{i=1}^{n} \|L_k(\cdot, t_{i-1})\| \right).$$

If the method is consistent of order $q$, then the property (5.54) yields the estimate

$$\|E_k(\cdot, t_n)\| \leq C_S\big(\|E_k(\cdot, 0)\| + TC_L k^q\big) \quad \text{for all } \; k < k_0$$

and each time $nk = t_n \leq T$. If the error vanishes in the initial data, then we obtain the convergence immediately. Otherwise, we demand that the errors in the initial values exhibit the magnitude $\mathcal{O}(h^q)$. We assume that $r = \frac{k}{h}$ is constant. It follows for $t = t_n$

$$\|E_k(\cdot, t)\| \leq C_E k^q \quad \text{for all } \; t \leq T \; \text{ and } \; k < k_0.$$

In particular, the order of consistency coincides with the order of convergence.

### CFL condition

A linear hyperbolic system (5.29) can be decoupled into $m$ separate linear advection equations $v_t + \lambda_j v_x = 0$, where $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ are the eigenvalues of the matrix $A$. It follows that the solution $u$ in a point $(x^*, t^*)$ depends on the initial values (5.30) at $t = 0$ just in the space points

$$\mathcal{D}(x^*, t^*) := \{x^* - \lambda_j t^* : j = 1, \ldots, m\}. \tag{5.61}$$

The set (5.61) is called the analytical domain of dependence of the solution.

A natural requirement for a numerical method is that the approximation in a grid point $(x_j, t_n)$ is computed using initial data close to the space points (5.61) for $x^* = x_j$, $t^* = t_n$. Otherwise, we could modify the initial data such that the exact solution changes in $(x_j, t_n)$, whereas the numerical approximation remains constant.

For example, we discuss the important case of an explicit method, where the approximation $U_j^n$ depends on the three previous approximations $U_l^{n-1}$ for $l = j-1, j, j+1$. Successively, the approximation $U_j^n$ is calculated by a finite set of initial values $U_l^0$, see Fig. 21. It follows the numerical domain of dependence

$$\mathcal{D}_k(x_j, t_n) \subset \{x : |x - x_j| \leq nh\}$$

at time $t = 0$. More general, a point $(x^*, t^*)$ with $t^* = nk$ exhibits the numerical domain of dependence

$$\mathcal{D}_k(x^*, t^*) \subset \left\{x : |x - x^*| \leq \tfrac{h}{k} t^*\right\}.$$

For constant $r = \frac{k}{h}$, the limit $k \to 0$ yields

$$\mathcal{D}_0(x^*, t^*) := \left\{x : |x - x^*| \leq \tfrac{t^*}{r}\right\}, \tag{5.62}$$

since the grid points become dense in the limit. The numerical domain of dependence (5.62) has to include the analytical domain of dependence (5.61), i.e.,

$$\mathcal{D}(x^*, t^*) \subset \mathcal{D}_0(x^*, t^*). \tag{5.63}$$

The requirement (5.63) is called the CFL condition due to Courant, Friedrichs and Lewy. We have shown that the CFL condition is necessary for the

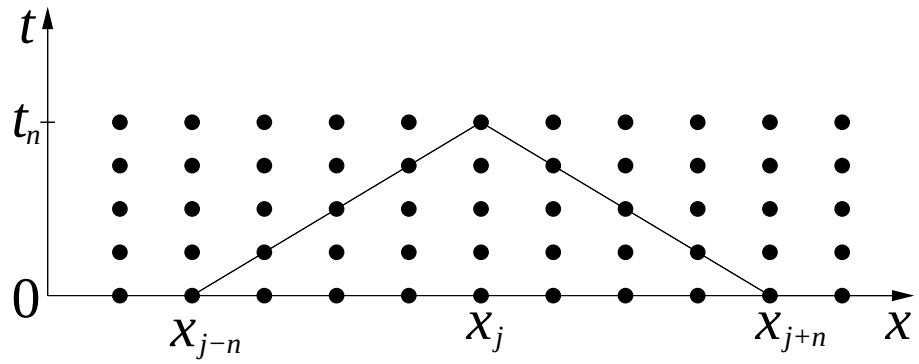



Figure 21: Numerical domain of dependence.

convergence of a numerical method. According to Theorem 15, the CFL condition is also necessary for the stability in case of a consistent method. Hence a consistent technique violating the CFL condition cannot be stable.

In our example, the CFL condition (5.63) results to the equivalent requirement

$$|(x^* - \lambda_p t^*) - x^*| \leq \tfrac{t^*}{r} \quad \Leftrightarrow \quad \left|\tfrac{\lambda_p k}{h}\right| \leq 1 \tag{5.64}$$

for all $p = 1, \ldots, m$. On the one hand, the CFL condition coincides with the stability condition (5.60) of the Lax-Friedrichs method (5.37). On the other hand, the explicit Euler method (5.35) satisfies the CFL condition (5.64) for sufficiently small step size $k$. However, the explicit Euler method is unstable for all step sizes. Hence the CFL property represents just a necessary condition for stability.

In particular, a CFL condition has to be satisfied for one-sided methods like the upwind schemes. Considering the linear advection equation $u_t + a u_x = 0$, the left-sided method (5.38) reads

$$U_j^{n+1} = U_j^n - a\tfrac{k}{h}\left(U_j^n - U_{j-1}^n\right), \tag{5.65}$$

whereas the right-sided scheme (5.39) results to

$$U_j^{n+1} = U_j^n - a\tfrac{k}{h}\left(U_{j+1}^n - U_j^n\right). \tag{5.66}$$

Both techniques are consistent of first order. A corresponding CFL condition can be constructed, which depends on the sign of the velocity $a$. In

case of $a > 0$, the CFL condition is never satisfied by the method (5.66), whereas the method (5.65) fulfills the CFL condition for sufficiently small time step size $k$. In case of $a < 0$, the properties are vice versa. Moreover, the upwind methods are stable if and only if their CFL condition is satisfied.

### Simulation of weak solutions

In the above analysis of consistency and convergence, we have assumed sufficiently smooth solutions, i.e., classical solutions. However, weak solutions appear in practice. We expect a corresponding critical behaviour for non-smooth solutions. Typically, discontinuities appear at isolated locations. To investigate the performance of the methods, we consider a Riemann problem of the scalar linear advection equation

$$\begin{aligned} u_t + a u_x = 0, \qquad & u : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R}, \\ u_0(x) = & \begin{cases} 1 & \text{for } x < 0 \\ 0 & \text{for } x > 0. \end{cases} \end{aligned} \tag{5.67}$$

with constant velocity $a \neq 0$. The unique solution is $u(x,t) = u_0(x - at)$. Due to the discontinuity, a difference formula for $u_x$ becomes unbounded in case of $h \to 0$. The local error of the method does not converge to zero. Thus Theorem 15 cannot be applied.

Alternatively, the initial data $u_0$ can be approximated by smooth functions $u_0^\varepsilon$, where the limit $\varepsilon \to 0$ recovers $u_0$. Given a method, which is consistent and stable for smooth solutions, it follows the convergence again. However, the order of convergence can be reduced significantly. In particular, a numerical solution includes obvious errors using some finite grid.

The simulation of the Riemann problem (5.67) for $a = 1$ illustrates the critical behaviour. We apply the Lax-Friedrichs method (5.37), the upwind scheme (5.65), the Lax-Wendroff method (5.42) and the Beam-Warming technique (5.43). Fig. 22 illustrates the results. The corresponding behaviour is typical:

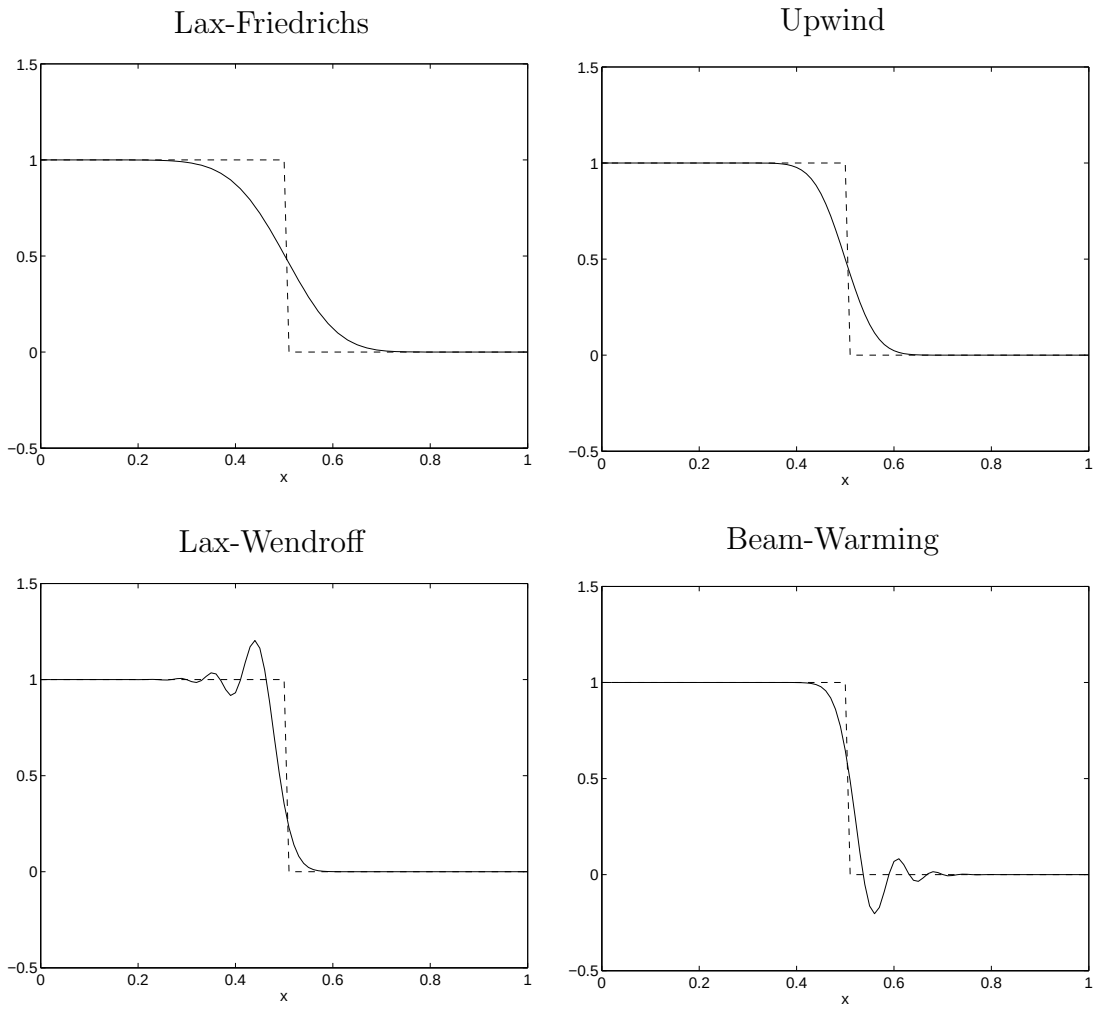



Figure 22: Numerical solutions using step sizes $h = 0.01$, $k = 0.005$ (solid line) and exact solution (dashed line) of Riemann problem for $a = 1$ at time $t = 0.5$.

- Methods of first order yield smeared solutions, i.e., an incorrect numerical diffusion appears. The shape of the discontinuity is not reproduced. Nevertheless, the decay of the numerical approximations is centered around the correct location of the discontinuity.
- Methods of second order reproduce the shape of the discontinuity significantly better. The position of the discontinuity is resolved correctly again. However, incorrect oscillations appear close to the discontinuity.

This qualitative behaviour of the finite difference methods can be explained by the concept of modified partial differential equations.

Furthermore, the global error can be estimated with respect to the integral norm (5.48) in case of the Riemann problem (5.67). Considering $h \to 0$ and $\frac{k}{h}$ constant, it follows

$$\|u(\cdot,t) - U_k(\cdot,t)\|_1 \approx C_t h^q \quad \text{for each } \ t \geq 0$$

with $q = \frac{1}{2}$ in case of the Lax-Friedrichs method (5.37) and $q = \frac{2}{3}$ in case of the Lax-Wendroff scheme (5.42). Hence the classical order of consistency is reduced significantly.

## 5.4 Conservative methods for nonlinear systems

Now we construct numerical methods for general nonlinear systems of conservation laws

$$u_t + f(u)_x = 0 \tag{5.68}$$

with an unknown solution $u : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R}^m$ and a smooth flux function $f : \mathbb{R}^m \to \mathbb{R}^m$. We assume that the system (5.68) is hyperbolic. A grid $x_j := jh$ and $t_n := nk$ is applied again with step sizes $h$ and $k$ in space and time, respectively. In particular, we want to determine approximations of weak solutions of the system (5.68). Accordingly, corresponding cells are considered in the domain of dependence, see Figure 23. The integral form (5.18) of the conservation law (5.68) yields the equations

$$\begin{aligned}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t_{n+1})\,\mathrm{d}x = &\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t_n)\,\mathrm{d}x \\ &+ \int_{t_n}^{t_{n+1}} f(u(x_{j-\frac{1}{2}},t))\,\mathrm{d}t - \int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}},t))\,\mathrm{d}t\end{aligned} \tag{5.69}$$

for each $j \in \mathbb{Z}$ and $n \in \mathbb{N}_0$. Dividing (5.69) by $h$, we obtain an equation for the evolution of the cell averages (5.31)

$$\bar{u}_j^{n+1} = \bar{u}_j^n + \frac{1}{h}\int_{t_n}^{t_{n+1}} f(u(x_{j-\frac{1}{2}},t))\,\mathrm{d}t - \frac{1}{h}\int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}},t))\,\mathrm{d}t. \tag{5.70}$$

Numerical methods are constructed using the equation (5.70) now.

**Definition 25 (conservative method)** *A finite difference method is called conservative, if it can be written in the form*

$$\begin{aligned}U_j^{n+1} \;=\; U_j^n - \tfrac{k}{h}\big[&F(U_{j-p}^n, U_{j-p+1}^n, \ldots, U_{j+q}^n) \\ &- F(U_{j-p-1}^n, U_{j-p}^n, \ldots, U_{j+q-1}^n)\big]\end{aligned} \tag{5.71}$$

*with a fixed function $F : \mathbb{R}^{p+q+1} \to \mathbb{R}^m$ and some integers $p, q \geq 0$.*

The most important case is $p = 0$ and $q = 1$ in (5.71), where the method reads

$$U_j^{n+1} = U_j^n - \tfrac{k}{h}\left[F(U_j^n, U_{j+1}^n) - F(U_{j-1}^n, U_j^n)\right]. \tag{5.72}$$

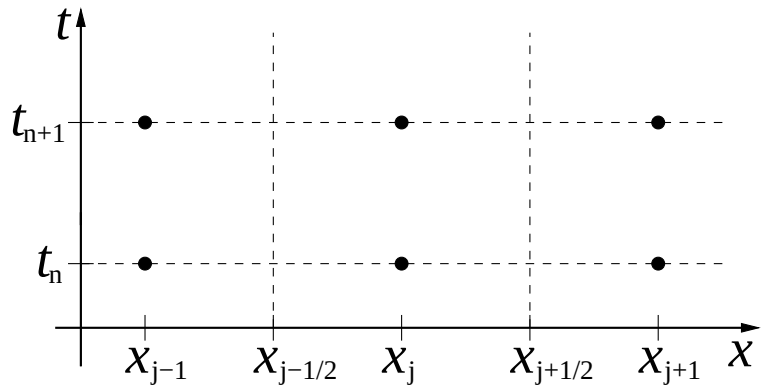



Figure 23: Cells in grid for finite difference method.

For the formula (5.71), we apply the short notation

$$U_j^{n+1} = U_j^n - \tfrac{k}{h}\left[F(U^n; j) - F(U^n; j-1)\right], \tag{5.73}$$

where $U^n$ represents the complete data at time $t_n$. A comparison of (5.70) and (5.73) shows that we want to achieve an approximation

$$F(U^n; j) \approx \frac{1}{k}\int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}}, t))\,\mathrm{d}t. \tag{5.74}$$

Thus the function $F$ is called the numerical flux function of the method.

To achieve a reasonable approximation in (5.74), a natural requirement is that a constant flux function is approximated exactly, i.e.,

$$F(u, u, \ldots, u) = f(u) \qquad \text{for each (relevant) } u \in \mathbb{R}^m. \tag{5.75}$$

Yet the condition (5.75) is not sufficient to obtain a convergent method. We demand a slightly stronger property.

**Definition 26 (consistency of conservative method)**
*A conservative method (5.71) is called consistent, if the local Lipschitz condition*

$$\left\|F(U_{j-p}^n, U_{j-p+1}^n, \ldots, U_{j+q}^n) - f(u)\right\| \le C \max_{-p\le i\le q} \|U_{j+i} - u\| \tag{5.76}$$

*holds for each relevant $u \in \mathbb{R}^m$ with a constant $C$ (which may depend on $u$) in an arbitrary vector norm.*

Sufficient for the Lipschitz condition (5.76) is the elementary property (5.75) and $F \in C^1$. Often $f \in C^1$ implies also $F \in C^1$. Obviously, the condition (5.76) implies (5.75).

Remark that we do not define an order of consistency in case of weak solutions. A consistency of higher order typically requires sufficiently smooth solutions. Smooth solutions are classical solutions, where the usual concept of consistency applies. In contrast, weak solutions are not smooth.

### Discrete conservation

Consistent conservative methods according to Definition 25 have an advantageous property in solving conservation laws, namely the principle of discrete conservation. Let initial values $u(x,0) = u_0(x)$ be given with $u_0(x) = u_{-\infty}$ for $x \le \alpha$ and $u_0(x) = u_{+\infty}$ for $x \ge \beta$. In particular, these assumptions are fulfilled for initial values with compact support. We choose $a < b$ and $T > 0$ such that

$$u(a,t) = u_{-\infty}, \quad u(b,t) = u_{+\infty} \qquad \text{for all } \ 0 \le t \le T$$

holds. On the one hand, the integral form (5.18) yields

$$\int_a^b u(x,t_N)\,\mathrm{d}x = \int_a^b u(x,0)\,\mathrm{d}x - t_N\,[f(u_{+\infty}) - f(u_{-\infty})] \tag{5.77}$$

for $t_N = Nk \le T$. On the other hand, we sum up the formulas of the conservative method (5.73)

$$\begin{aligned} h\sum_{j=J}^{L} U_j^{n+1} &= h\sum_{j=J}^{L} U_j^n - k\sum_{j=J}^{L}[F(U^n;j) - F(U^n;j-1)] \\ &= h\sum_{j=J}^{L} U_j^n - k[F(U^n;L) - F(U^n;J-1)] \\ &= h\sum_{j=J}^{L} U_j^n - k[f(u_{+\infty}) - f(u_{-\infty})] \end{aligned}$$

using some $J, L$ satisfying $Jh \ll \alpha$ and $Lh \gg \beta$. We have applied the consistency (5.75) of the conservative method in the last equality. Recursively,

we obtain

$$h\sum_{j=J}^{L} U_j^N = h\sum_{j=J}^{L} U_j^0 - t_N[f(u_{+\infty}) - f(u_{-\infty})].$$

We assume that the initial values are the exact cell averages, i.e., $U_j^0 = \bar{u}_j^0$. It follows

$$h\sum_{j=J}^{L} U_j^0 = \int_{x_{J-\frac{1}{2}}}^{x_{L+\frac{1}{2}}} u(x,0)\,\mathrm{d}x.$$

and thus

$$h\sum_{j=J}^{L} U_j^N = \int_{x_{J-\frac{1}{2}}}^{x_{L+\frac{1}{2}}} u(x,0)\,\mathrm{d}x - t_N[f(u_{+\infty}) - f(u_{-\infty})]. \tag{5.78}$$

A comparison of (5.77) and (5.78) yields the crucial equality

$$h\sum_{j=J}^{L} U_j^N = \int_{x_{J-\frac{1}{2}}}^{x_{L+\frac{1}{2}}} u(x,t_N)\,\mathrm{d}x. \tag{5.79}$$

The finite difference method defines an approximating function (5.32) satisfying

$$\int_{x_{J-\frac{1}{2}}}^{x_{L+\frac{1}{2}}} U_k(x,t_N)\,\mathrm{d}x = \int_{x_{J-\frac{1}{2}}}^{x_{L+\frac{1}{2}}} u(x,t_N)\,\mathrm{d}x$$

due to (5.79). Hence the integral of the approximation $U_k$ coincides with the integral of the exact solution in case of a consistent conservative method. No errors occur in the relation (5.79). The integral form (5.18) represents a conservation law. In view of (5.79), the conservative method exhibits a discrete conservation of the same quantities.

### Construction of conservative methods

Now we want to obtain conservative methods of the form (5.72). Thereby, we construct approximations according to (5.74), i.e., the time integral at the intermediate space point is approximated by the data of the two neighbouring grid points.

A simple choice of the numerical flux function is given by

$$F_{\mathrm{l}}(U_j, U_{j+1}) := f(U_j) \quad \text{or} \quad F_{\mathrm{r}}(U_j, U_{j+1}) := f(U_{j+1}).$$

The consistency (5.75) follows straightforward in both cases. A corresponding CFL condition determines the choice of either $F_{\mathrm{l}}$ or $F_{\mathrm{r}}$ in case of scalar conservation laws ($m = 1$).

Alternatively, we approximate the integral (5.74) by the arithmetic mean of the data in the two neighbouring grid points, i.e.,

$$F(U_j, U_{j+1}) := \tfrac{1}{2}\left(f(U_j) + f(U_{j+1})\right).$$

Again the consistency (5.75) is obvious. It follows the finite difference scheme

$$\begin{aligned} U_j^{n+1} &= U_j^n - \tfrac{k}{h}\left[\tfrac{1}{2}f(U_j) + \tfrac{1}{2}f(U_{j+1}) - \tfrac{1}{2}f(U_{j-1}) - \tfrac{1}{2}f(U_j)\right] \\ &= U_j^n - \tfrac{k}{2h}\left[f(U_{j+1}^n) - f(U_{j-1}^n)\right]. \end{aligned}$$

This method is just the explicit Euler method (5.35) in the nonlinear case, which represents an unstable scheme. Again the method can be stabilised by replacing the central approximation $U_j^n$ by the arithmetic mean of the neighbouring approximations. It follows the Lax-Friedrichs method, cf. (5.37),

$$U_j^{n+1} = \tfrac{1}{2}\left(U_{j-1}^n + U_{j+1}^n\right) - \tfrac{k}{2h}\left[f(U_{j+1}^n) - f(U_{j-1}^n)\right].$$

The method can be written in the form (5.72) with the numerical flux function

$$F(U_j, U_{j+1}) = \tfrac{h}{2k}\left(U_j - U_{j+1}\right) + \tfrac{1}{2}\left(f(U_j) + f(U_{j+1})\right). \tag{5.80}$$

Hence the method is conservative and the consistency (5.75) is satisfied. The first term in (5.80) does not imply a reasonable approximation for the integral (5.74). However, this term becomes tiny for $U_j \approx U_{j+1}$. Thus the first term just causes the desired stabilisation of the method. Furthermore, the method (5.80) is convergent of order one in case of classical solutions.

Nonlinear generalisations of the Lax-Wendroff method (5.42) are also feasible. Considering sufficiently smooth solutions, the Taylor expansion (5.40) exists. The conservation law (5.68) yields

$$u_t = -f(u)_x, \quad u_{tt} = -f(u)_{xt} = -f(u)_{tx} = -(A(u)u_t)_x = (A(u)f(u)_x)_x,$$

where $A(u) \in \mathbb{R}^{m\times m}$ represents the Jacobian matrix of the flux function $f$ assuming $f \in C^1$. We replace the time derivatives in the Taylor expansion (5.40) and obtain

$$u(x,t+k) = u(x,t) - kf(u)_x + \tfrac{1}{2}k^2(A(u)f(u)_x)_x + \mathcal{O}(h^3).$$

Now we substitute the space derivatives by centered difference formulas of second order. An inner as well as outer approximation is applied to the term $(A(u)f(u)_x)_x$. It follows the scheme

$$\begin{aligned} U_j^{n+1} &= U_j^n - \tfrac{k}{2h}\left[f(U_{j+1}^n) - f(U_{j-1}^n)\right] \\ &\quad + \tfrac{k^2}{2h^2}\left[A_{j+\frac{1}{2}}(f(U_{j+1}^n) - f(U_j^n)) - A_{j-\frac{1}{2}}(f(U_j^n) - f(U_{j-1}^n))\right] \end{aligned}$$

with $A_{j\pm\frac{1}{2}} := A(u(x_{j\pm\frac{1}{2}}))$. Since the intermediate values $u(x_{j\pm\frac{1}{2}})$ are unknown, we replace them using a linear interpolation with the two neighbouring values, which also represents an approximation of second order. It follows

$$\begin{aligned} U_j^{n+1} &= U_j^n - \tfrac{k}{2h}\left[f(U_{j+1}^n) - f(U_{j-1}^n)\right] \\ &\quad + \tfrac{k^2}{2h^2}\big[A(\tfrac{1}{2}(U_j^n + U_{j+1}^n))(f(U_{j+1}^n) - f(U_j^n)) \\ &\qquad - A(\tfrac{1}{2}(U_{j-1}^n + U_j^n))(f(U_j^n) - f(U_{j-1}^n))\big]. \end{aligned} \tag{5.81}$$

The corresponding numerical flux function reads

$$F(U_j, U_{j+1}) = \tfrac{1}{2}\left(f(U_j) + f(U_{j+1})\right) - \tfrac{k}{2h}A(\tfrac{1}{2}(U_j + U_{j+1}))(f(U_{j+1}) - f(U_j)).$$

The consistency (5.75) follows from the first term, whereas the second term cancels out. Thus the first term is crucial for approximating weak solutions. The second term can be seen as a correction, which causes a consistent approximation of second order in case of sufficiently smooth solutions.

A drawback of the finite difference method (5.81) is that the Jacobian matrices of the flux function have to be evaluated, which increases the computational effort. Similar finite difference methods can be constructed, which avoid evaluations of the Jacobians. The Richtmyer two-step Lax-Wendroff

method reads

$$\begin{aligned} U_{j+\frac{1}{2}}^{n+\frac{1}{2}} &= \tfrac{1}{2}\left(U_j^n + U_{j+1}^n\right) - \tfrac{k}{2h}\left[f(U_{j+1}^n) - f(U_j^n)\right] \\ U_j^{n+1} &= U_j^n - \tfrac{k}{h}\left[f\left(U_{j+\frac{1}{2}}^{n+\frac{1}{2}}\right) - f\left(U_{j-\frac{1}{2}}^{n+\frac{1}{2}}\right)\right]. \end{aligned} \tag{5.82}$$

MacCormack's method is given by

$$\begin{aligned} U_j^* &= U_j^n - \tfrac{k}{h}\left[f(U_{j+1}^n) - f(U_j^n)\right] \\ U_j^{n+1} &= \tfrac{1}{2}\left(U_j^n + U_j^*\right) - \tfrac{k}{2h}\left[f\left(U_j^*\right) - f\left(U_{j-1}^*\right)\right]. \end{aligned} \tag{5.83}$$

All three techniques (5.81), (5.82), (5.83) reduce to the original Lax-Wendroff method (5.42) in case of linear conservation laws ($f(u) = Au$). The three methods are convergent of order two in case of sufficiently smooth functions. Moreover, each method can be written in the form (5.71) with a numerical flux function. Hence the methods are conservative. It can be shown that the schemes are consistent according to Definition 26. In conclusion, the three methods are also appropriate for the determination of weak solutions.

The method (5.81) has been constructed based on a Taylor expansion assuming a sufficiently smooth solution. It is surprising that such methods are also appropriate in case of weak solutions. Moreover, they have the advantageous property to achieve reasonable approximations in non-smooth parts of the solution, whereas they switch automatically to approximations of second order in smooth parts of the solution.

Considering discontinuous solutions like in a Riemann problem, the nonlinear generalisations of the Lax-Friedrichs method and the Lax-Wendroff scheme exhibit the same behaviour as in the linear case, cf. Section 5.3. In particular, the discrete conservation (5.79) can be observed in Figure 22.