Preface

These lecture notes are intended as a written companion to the course “Numerical Methods IV: Partial Differential Equations” taught in summer 2019 at the University of Vienna. The course is part of an interdisciplinary masters programme called Computational Science and intends to acquaint students with a basic knowledge of numerical methods for partial differential equations.

In creating these notes I have used the following references. They are highly recommended for further reading.

- *Partial Differential Equations* by Lawrence C. Evans, American Mathematical Society 2010

In addition, I have used the lecture notes

- *Einführung in die Numerik der Differentialgleichungen* by Winfried Auzinger (held in summer term 2007 at TU Wien),
- *Numerical Methods IV: Partial Differential Equations* by Leonidas Mindrinos and Otmar Scherzer (held in summer terms 2017 and 2018 at the University of Vienna).

Clemens Kirisits
## Contents

Preface iii

1 Partial Differential Equations 1
   1.1 Introduction ................................................. 1
   1.2 Explicit Solutions ........................................... 4
      1.2.1 Boundary and Initial Conditions ...................... 6
      1.2.2 The Linear Transport Equation ......................... 7
      1.2.3 The Wave Equation ...................................... 8
      1.2.4 Laplace’s Equation ..................................... 10
      1.2.5 Heat Equation .......................................... 11
   1.3 Well-posed Problems ......................................... 13
   1.4 Numerical Analysis of PDEs .................................. 15
   1.5 Classification of PDEs ....................................... 15
      1.5.1 The Two-variable Case .................................. 16
      1.5.2 The General Case ...................................... 17

2 The Finite Difference Method 21
   2.1 Finite Difference Approximations ......................... 21
      2.1.1 First Order Derivatives ............................... 21
      2.1.2 Higher Order Derivatives ............................. 23
   2.2 Elliptic Problems ........................................... 25
      2.2.1 Deriving the Difference Equations .................... 25
      2.2.2 Consistency, Stability and Convergence ............... 28
      2.2.3 The Neumann Problem .................................. 31
      2.2.4 More General Problems ................................ 33
   2.3 Parabolic Problems .......................................... 35
      2.3.1 An Explicit Scheme .................................... 35
      2.3.2 Von Neumann Stability Analysis ....................... 36
      2.3.3 Convergence ........................................... 37
      2.3.4 The $\theta$-scheme ................................... 38
      2.3.5 The Lax Equivalence Theorem ......................... 41
   2.4 Hyperbolic Problems ........................................ 42
      2.4.1 The Lax-Friedrichs Scheme ........................... 42
      2.4.2 The Upwind Scheme .................................... 44
      2.4.3 The Courant-Friedrichs-Lewy Condition ............... 45
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.4</td>
<td>Higher Order Methods</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>The Finite Element Method</td>
<td>49</td>
</tr>
<tr>
<td>3.1</td>
<td>A One-Dimensional Example</td>
<td>49</td>
</tr>
<tr>
<td>3.2</td>
<td>Principles of the Finite Element Method</td>
<td>52</td>
</tr>
<tr>
<td>3.2.1</td>
<td>The Weak Formulation</td>
<td>53</td>
</tr>
<tr>
<td>3.2.2</td>
<td>The Galerkin Method</td>
<td>54</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Piecewise Polynomials</td>
<td>57</td>
</tr>
<tr>
<td>3.3</td>
<td>A Two-Dimensional Example</td>
<td>58</td>
</tr>
<tr>
<td>3.4</td>
<td>Convergence</td>
<td>63</td>
</tr>
</tbody>
</table>

Appendix

65
Chapter 1
Partial Differential Equations

1.1 Introduction

Our understanding of the fundamental processes of the natural world is based to a large extent on partial differential equations. Examples are the vibrations of solids, the flow of fluids, the diffusion of chemicals, the spread of heat, the structure of molecules, the interactions of photons and electrons, and the radiation of electromagnetic waves. Partial differential equations also play a central role in modern mathematics, especially in geometry and analysis. The availability of powerful computers is gradually shifting the emphasis in partial differential equations away from the analytical computation of solutions and toward both their numerical analysis and the qualitative theory.\(^1\)

A partial differential equation (PDE) is an equation which relates certain partial derivatives of an unknown function to each other and possibly to the function itself. A defining property is that the unknown must be a function of two or more variables.

For example, a first-order PDE in two dimensions takes the following general form

\[
F(x, y, u, u_x, u_y) = 0.
\]

Here, \(x\) and \(y\) are the independent variables, \(u = u(x, y)\) is the unknown dependent variable, and \(F\) is some given real-valued function specifying the particular form of the equation. Subscripts denote partial derivatives, for instance,

\[
u_x(x, y) = \frac{\partial u}{\partial x}(x, y) = \lim_{h \to 0} \frac{1}{h} (u(x + h, y) - u(x, y)).
\]

The order of a PDE is the order of the highest appearing derivative. Thus, a second-order PDE in two dimensions can be written as

\[
F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0. \tag{1.1.1}
\]

\(^1\)Quoted from the preface of [9].
Recall that the mixed derivatives agree for a twice continuously differentiable function: $u_{xy} = u_{yx}$. Solving a PDE means finding all solutions. A solution to (1.1.1), for instance, is a function $u$ that satisfies the equation for all $x, y$ ranging over some given region of $\mathbb{R}^2$. As a general rule higher order equations are usually more difficult to solve than lower order ones.

**Linearity**

In the field of PDE a fundamental distinction is drawn between linear and nonlinear equations. A PDE is called *linear*, if $F$ depends linearly on $u$ and all its derivatives. To be more precise let us write a general PDE in operator form

$$L(u) = f.$$  

(1.1.2)

Here, $f$ is some given function (that does not depend on $u$) and $L$ denotes a *differential operator*. An operator takes a function $u$ and produces a new function $L(u)$. The word “differential” implies that the new function $L(u)$ involves certain derivatives of $u$. Now, the PDE (1.1.2) is linear, if $L$ is a linear operator, that is,

$$L(u + v) = L(u) + L(v)$$

$$L(cu) = cL(u)$$

for all functions $u$ and $v$ and all numbers $c$. In this case the brackets enclosing the argument are often omitted and one simply writes $Lu$. Recall that differentiation itself is always linear operation. That is, the simplest differential operator $Lu = u_x$ always satisfies the two requirements above.

If the right hand side in (1.1.2) vanishes identically (i.e. equals zero for all $x$), in symbols $f \equiv 0$, then the resulting equation

$$L(u) = 0$$

is called *homogeneous*. An important implication of linearity is the following. If $u$ and $v$ solve the same homogeneous linear equation, then so does their sum $u + v$. This is sometimes called the *superposition principle*. More generally, consider $m$ homogeneous solutions $u_1, \ldots, u_m$, that is, each $u_i$ solves the homogeneous linear PDE

$$Lu_i = 0.$$  

Then any linear combination of the $u_i$ is also a homogeneous solution

$$L(c_1 u_1 + \cdots + c_m u_m) = 0, \quad c_i \in \mathbb{R}.$$  

Another consequence of linearity is this: If $u$ is a homogeneous solution and $v$ an inhomogeneous one, then $u + v$ is another inhomogeneous solution. Therefore, for linear PDEs it is easy to construct new solutions from previously found ones.
Introduction

Many physical phenomena can be modelled well by linear equations. Generally speaking, however, improving the accuracy of PDE-based models often means taking more nonlinearities into account. From a mathematical viewpoint it must be said that linear PDEs are usually easier to solve than nonlinear ones, both analytically and numerically. Not surprisingly, both mathematical theory and numerical methods are better developed for the former. In particular, most numerical methods will transform a linear PDE $Lu = f$ in a very natural way into a linear system of equations $Ax = b$, for which there is a plethora of efficient algorithms. Therefore, in this lecture we will mainly consider numerical methods for linear PDEs.

Examples of Important PDEs

Below we look at a few important examples of PDEs and introduce some notation along the way. They are generally stated in their simplest form neglecting physical constants, coefficients, etc.

1. One of the simplest PDEs is the linear transport equation

$$u_t + \nabla u \cdot b = 0.$$  \hspace{1cm} (1.1.3)

Here $\nabla u = (u_{x_1}, \ldots, u_{x_d})^T$ denotes the spatial gradient of $u$, $b \in \mathbb{R}^d$ is a given vector and $a \cdot b = \sum_i a_i b_i$ is the inner product on $\mathbb{R}^d$. Equation (1.1.3) is linear, homogeneous and of first order. It describes the transport of an initially given quantity with velocity $b$. See Section 1.2.2 below.

2. A famous example of a homogeneous linear second-order PDE is Laplace’s equation

$$\Delta u = 0.$$  \hspace{1cm} (1.1.4)

The differential operator $\Delta u = \sum_{i=1}^d u_{x_i x_i}$ is the $d$-dimensional Laplace operator. Solutions of (1.1.4) are called harmonic functions. Laplace’s equation describes diffusion processes such as heat conduction in equilibrium.

3. The inhomogeneous version of (1.1.4) is called Poisson’s equation

$$\Delta u = f.$$  \hspace{1cm} (1.1.5)

Physically, the inhomogeneity $f$ represents a source or sink, which can create or deplete, respectively, some of the quantity which is represented by $u$.

4. The diffusion equation, also called heat equation, is another well-known second order linear PDE. It reads

$$u_t - \Delta u = f$$

and models the evolution in time of heat, chemical concentration etc. In equilibrium state ($u_t = 0$) it reduces to Poisson’s equation.
5. The *wave equation*, which in its simplest form reads

$$u_{tt} - \Delta u = 0,$$

(1.1.5)

can describe a wide range of phenomena including the vibration of solids (such as strings, membranes, elastic bodies) but also fluid waves or sound waves. The operator $\frac{\partial^2}{\partial t^2} - \Delta$ is often called *wave* or *d’Alembert operator*. See also Section 1.2.3.

6. We conclude this section with an example of a nonlinear second-order PDE. The *minimal surface equation*

$$\text{div} \frac{\nabla u}{\sqrt{1 + |\nabla u|^2}} = 0$$

characterizes surfaces $(x, y, z) = (x, y, u(x, y))$ which have minimal total surface area subject to certain constraints, such as a soap film resulting from dipping a wire frame into soap solution. Recall that the *divergence* of a vector function $v = (v^1, \ldots, v^d)$ is given by

$$\text{div} v = v^1_{x_1} + \cdots + v^d_{x_d},$$

while by $|v|$ we denote its Euclidean length, that is,

$$|v| = \sqrt{(v^1)^2 + \cdots + (v^d)^2}.$$  

### 1.2 Explicit Solutions

Arguably the most straightforward way to study a given PDE is to try to find solutions right away. Let us do just that with a few simple examples.

**Example 1.1.** First, consider the *ordinary* differential equation (ODE) $u'' = 0$. Its solutions are $u(x) = ax + b$, where $a$ and $b$ are arbitrary constants. Next, consider a very similar PDE. Let us try find all functions $u(x, y)$ satisfying

$$u_{xx} = 0.$$  

(1.2.1)

Clearly, $u(x, y) = ax + b$ is again a solution. But this answer is far from including all solutions! Since now $u$ may also depend on $y$ the general solution is actually $u(x, y) = xf(y) + g(y)$, where $f$ and $g$ are *arbitrary functions*. Equation (1.2.1) is essentially an ODE, but since the unknown is a function of more than one variable, the solution contains arbitrary functions instead of arbitrary constants.

**Example 1.2.** For another simple example consider the PDE

$$u_{xy} = 0.$$  

(1.2.2)

Let us start by integrating with respect to $y$. We obtain $u_x(x, y) = f(x)$ where $f$ is an arbitrary function depending only on $x$. After integrating with respect to $x$ we get the general solution $u(x, y) = F(x) + G(y)$, where $F' = f$. 


Example 1.3. The one-dimensional wave equation reads

\[ u_{tt} - u_{xx} = 0. \]  \hfill (1.2.3)

This PDE has a very convenient property: just as the binomial \( a^2 - b^2 = (a + b)(a - b) \) can be written as a product of two first-order terms, so can the wave operator be “factored” into two first-order differential operators

\[ \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} = \left( \frac{\partial}{\partial t} + \frac{\partial}{\partial x} \right) \left( \frac{\partial}{\partial t} - \frac{\partial}{\partial x} \right). \]  \hfill (1.2.4)

Inspired by this let us make an educated guess: If we could transform \((t, x)\) into a new pair of coordinates \((s, y)\) such that

\[ \frac{\partial}{\partial s} = \frac{\partial}{\partial t} + \frac{\partial}{\partial x} \quad \text{and} \quad \frac{\partial}{\partial y} = \frac{\partial}{\partial t} - \frac{\partial}{\partial x}, \]

up to multiplication by a constant, then the wave equation should essentially reduce to equation (1.2.2). It turns out that the coordinate transformation that achieves just this is very straightforward

\[ s = t + x, \quad y = t - x. \]

Let us verify by using the chain rule. Regarding \(s\) and \(y\) as functions of \((t, x)\) we have for every function \(u = u(t, x)\) that

\[ u_t = u_s s_t + u_y y_t = u_s + u_y, \]
\[ u_x = u_s s_x + u_y y_x = u_s - u_y. \]

Adding and subtracting these two equations, respectively, shows that

\[ \frac{\partial}{\partial t} + \frac{\partial}{\partial x} = 2 \frac{\partial}{\partial s} \quad \text{and} \quad \frac{\partial}{\partial t} - \frac{\partial}{\partial x} = 2 \frac{\partial}{\partial y}. \]

Together with the factorization (1.2.4) it follows that (1.2.3) is equivalent to \(u_{sy} = 0\). We conclude from Example 1.2 that the solution of (1.2.3) is given by

\[ u(x, t) = F(t + x) + G(t - x) \]  \hfill (1.2.5)

for two arbitrary functions \(F, G\). Thus, \(u\) is the sum of a left-moving and a right-moving wave.

We have seen that solutions to PDEs can, and often do, contain arbitrary functions. In order to have a uniquely determined solution we need to impose additional restrictions. Such restrictions are derived from the underlying physics and they come in two main types: initial conditions and boundary conditions.
1.2.1 Boundary and Initial Conditions

PDEs are usually formulated in some physical region $\Omega \subset \mathbb{R}^d$. A more precise way of stating problem (1.1.2) is therefore: Find a function $u : \Omega \to \mathbb{R}$ such that

$$L(u)(x) = f(x) \quad \text{for all } x \in \Omega.$$ 

Or, in a time-dependent problem, we look for functions $u(x,t)$ satisfying

$$L(u)(x,t) = f(x,t) \quad \text{for all } x \in \Omega \text{ and } t \in (0,T).$$ 

Here $T$ is either a positive number or equal to $\infty$.

Boundary conditions prescribe the behaviour of admissible solutions $u$ as $x$ approaches the boundary $\partial \Omega$ of $\Omega$. Recall that the boundary of a three-dimensional domain $\Omega$ is a surface, while the boundary of a two-dimensional domain is a curve. In one space dimension $\Omega$ is an interval $(a,b)$ and its boundary reduces to the two endpoints $\partial \Omega = \{a,b\}$. The two most common types of boundary conditions are discussed below.

A *Dirichlet boundary condition* specifies the values of $u$ itself on the boundary of $\Omega$, that is,

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

where $\phi$ is a given function. For time-dependent PDEs boundary conditions are usually imposed at all times

$$u(x,t) = \phi(x,t) \quad \text{for all } x \in \partial \Omega \text{ and } t \in [0,T).$$

Neumann boundary conditions on the other hand specify the normal derivative $\partial u/\partial n$ of $u$:

$$\frac{\partial u}{\partial n}(x) = \phi(x) \quad \text{for all } x \in \partial \Omega,$$

with an analogous modification to time-dependent problems. The normal derivative of $u$ can be computed as $\nabla u \cdot n$, where $n$ is the *outward pointing unit normal*. In other words the vector $n = n(x) \in \mathbb{R}^d$ is perpendicular to $\partial \Omega$, it points outward of $\Omega$ and it satisfies $|n| = 1$. In one-dimensional problems, $\Omega = (a,b)$, the normal derivative is just the regular partial derivative with respect to $x \in \mathbb{R}$. If $\phi$ vanishes for all $x \in \partial \Omega$ then the boundary condition is called *homogeneous*.

Initial conditions determine the unknown at $t = 0$

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

For certain problems, such as the wave equation (1.1.5), it is common to also prescribe the initial value of the time derivative

$$u_t(x,0) = \psi(x) \quad \text{for all } x \in \Omega.$$ 

Initial value problems are sometimes called *Cauchy problems*.
Example 1.4. Consider the heat equation

\[
\begin{aligned}
&u_t - \alpha \Delta u = 0 \quad \text{in } \Omega \times (0, +\infty), \\
&u = \phi \quad \text{at } t = 0, \\
&u = \psi \quad \text{on } \partial \Omega,
\end{aligned}
\]

whose solution \(u(x, t)\) describes the evolution of temperature in an object \(\Omega\) over time. Here, \(\alpha > 0\) is the thermal diffusivity, which describes the rate of heat transfer in the material under consideration. The initial condition \(\phi(x)\) determines the initial temperature distribution in \(\Omega\), while the Dirichlet boundary condition \(\psi(t)\) can model a situation where the object is immersed in a comparatively large reservoir of specified temperature \(\psi(t)\). A homogeneous Neumann boundary condition

\[
\frac{\partial u}{\partial n} = 0 \quad \text{on } \partial \Omega,
\]

on the other hand would mean that no heat can flow across the object’s boundary. In other words \(\Omega\) is perfectly insulated.

In the following sections we try to find explicit solution for a few initial and boundary value problems.

1.2.2 The Linear Transport Equation

Consider the initial value problem for the linear transport equation (1.1.3) in one space dimension

\[
\begin{aligned}
&u_t + bu_x = 0 \quad \text{in } \mathbb{R} \times (0, \infty), \\
&u = \phi \quad \text{for } t = 0,
\end{aligned}
\]

where \(b \in \mathbb{R}\) and the function \(\phi : \mathbb{R} \to \mathbb{R}\) are given. Observe that the transport equation can be written as \((u_x, u_t) \cdot (b, 1) = 0\), which means that the directional derivative of \(u\) along \((b, 1)\) vanishes for all \(x \in \mathbb{R}\) and \(t > 0\). In other words, \(u\) must be constant on the lines

\[
\begin{pmatrix} x \\ t \end{pmatrix} + s \begin{pmatrix} b \\ 1 \end{pmatrix}, \quad s \in \mathbb{R}.
\]

(1.2.6)

Therefore, setting \(s = -t\) and invoking the initial condition we obtain

\[
u(x, t) = u(x + bs, t + s) = u(x - bt, 0) = \phi(x - bt).
\]

We have found an explicit formula for the unique solution of the initial value problem. The function \(u\) at \(t > 0\) is simply a shifted, or transported, version of the initial condition \(\phi\), where \(b\) determines the direction and the speed of this shifting.
Next, consider the inhomogeneous version of the previous problem

\[
\begin{cases}
    u_t + bu_x = f & \text{in } \mathbb{R} \times (0, \infty), \\
    u = \phi & \text{for } t = 0,
\end{cases}
\]

where \( f = f(x, t) \) is a given function. Now, the directional derivative of \( u \) along \((b, 1)\) no longer vanishes, but it is equal to \( f \). We can express this relation as

\[
\frac{d}{ds} u(x + bs, t + s) = f(x + bs, t + s). \tag{1.2.7}
\]

Integrating both sides from \( s = -t \) to \( s = 0 \) gives

\[
u(x + bs, t + s) \bigg|_{s=0}^{s=0} = \int_{-t}^{0} f(x + bs, t + s) \, ds
\]

and therefore

\[
u(x, t) = \phi(x - bt) + \int_{-t}^{0} f(x + bs, t + s) \, ds.
\]

Finally, essentially the same approach works when we include a zero order term in the transport equation

\[
u_t + bu_x + cu = f. \tag{1.2.8}
\]

Only this time the ODE that results from restricting the PDE to the characteristic lines (1.2.6) cannot be integrated as directly as (1.2.7). Note that (1.2.8) actually constitutes a general first-order linear PDE with constant coefficients.\(^2\)

Thus we have found a way to solve a sizeable class of PDEs.

The approach we have employed above is a special case of the so-called method of characteristics. It generally consists in reducing a PDE into an appropriate family of ODEs and also works for certain nonlinear first-order equations. See Chapter 3.2 in [3] for a detailed treatment.

### 1.2.3 The Wave Equation

Consider the initial value problem for the one-dimensional wave equation

\[
\begin{cases}
    u_{tt} - u_{xx} = 0 & \text{in } \mathbb{R} \times (0, \infty), \\
    u = \phi & \text{for } t = 0, \\
    u_t = \psi & \text{for } t = 0.
\end{cases}
\]

Instead of changing variables this time we solve the wave equation by reducing it to a linear transport equation. Again, the reason why this works is the “factorization” (1.2.4).

---

\(^2\)Having more than two independent variables, \( x \in \mathbb{R}^d \) for example, does not complicate the method of characteristics in any way.
Explicit Solutions

Introducing the new unknown $v = u_t - u_x$ we can rewrite the wave equation

$$v_t + v_x = 0,$$

which is the homogeneous transport equation with $b = 1$. Therefore

$$v(x, t) = v(x - t, 0) = u_t(x, t) - u_x(x, t).$$

This is another linear transport equation, but this time in the original unknown $u$ and with inhomogeneity $v$ and $b = -1$. We obtain

$$u(x, t) = \phi(x + t) + \int_{x-t}^{x+t} v(p, 0) \, dp,$$

where in the last step we substituted $p = x - t - 2s$. We can get rid of the auxiliary unknown $v$ by recalling its definition and using the pair of initial conditions

$$v(x, 0) = u_t(x, 0) - u_x(x, 0) = \psi(x) - \phi'(x),$$

which finally leads to

$$u(x, t) = \phi(x + t) + \frac{1}{2} \int_{x-t}^{x+t} (\psi(p) - \phi'(p)) \, dp$$

$$= \frac{1}{2} \left( \phi(x + t) + \phi(x - t) + \int_{x-t}^{x+t} \psi(p) \, dp \right). \quad (1.2.9)$$

This expression for $u$ is also called d’Alembert’s formula. Notice that the solution is indeed of the form (1.2.5).

In three space dimensions

$$\begin{cases}
    u_{tt} - \Delta u = 0 & \text{in } \mathbb{R}^3 \times (0, \infty), \\
    u = \phi & \text{for } t = 0, \\
    u_t = \psi & \text{for } t = 0,
\end{cases}$$

the solution is given by Kirchhoff’s formula

$$u(x, t) = \frac{1}{4\pi t} \int_{\partial B(x,t)} \psi \, dS + \frac{\partial}{\partial t} \left( \frac{1}{4\pi t} \int_{\partial B(x,t)} \phi \, dS \right),$$

where $B(x, t)$ denotes the ball with center $x$ and radius $t$. 
1.2.4 Laplace’s Equation

First, we will show that Laplace’s equation is rotationally invariant. Let us try to verify this in the \((x, y)\)-plane. Consider a new pair of variables \((\xi, \eta)\) resulting from \((x, y)\) by rotation through an angle \(\alpha\)

\[
\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.
\]

Applying the chain rule as in Example 1.3 we obtain

\[
\begin{align*}
  u_x &= u_\xi \xi_x + u_\eta \eta_x = u_\xi \cos \alpha - u_\eta \sin \alpha, \\
  u_y &= u_\xi \xi_y + u_\eta \eta_y = u_\xi \sin \alpha + u_\eta \cos \alpha, \\
  u_{xx} &= u_{\xi\xi} (\cos \alpha)^2 - 2u_{\xi\eta} \cos \alpha \sin \alpha + u_{\eta\eta} (\sin \alpha)^2, \\
  u_{yy} &= u_{\xi\xi} (\sin \alpha)^2 + 2u_{\xi\eta} \cos \alpha \sin \alpha + u_{\eta\eta} (\cos \alpha)^2.
\end{align*}
\]

Recalling that \((\sin \alpha)^2 + (\cos \alpha)^2 = 1\), we have

\[
\Delta u = u_{xx} + u_{yy} = u_{\xi\xi} + u_{\eta\eta}.
\]

This rotational invariance suggests looking for solutions having the same property. In polar coordinates \((r, \theta)\) that means we want to find solutions which are independent of the angle \(\theta\) and only depend on the distance from the origin \(r\). Such functions are called radial solutions. We make the ansatz \(u(x, y) = v(r)\). Recall that \(r = \sqrt{x^2 + y^2}\) and therefore \(\partial r/\partial x = x/r\). Using the chain rule we get

\[
\begin{align*}
  u_x &= v'(r) \frac{\partial r}{\partial x} = v'(r) \frac{x}{r}, \\
  u_{xx} &= v''(r) \frac{x^2}{r^2} + v'(r) \left( \frac{1}{r} - \frac{x^2}{r^3} \right),
\end{align*}
\]

and similarly for \(y\). Therefore \(\Delta u = v'' + v'/r\) and we can rewrite Laplace’s equation as the following ODE

\[
0 = \frac{d}{dr} \left( rv'(r) \right),
\]

which is solved by

\[
v(r) = c_1 \log r + c_2 \tag{1.2.10}
\]

for arbitrary constants \(c_1\) and \(c_2\). Note that this solution has a singularity at the origin unless \(c_1 = 0\).

Now consider the Dirichlet problem for Laplace’s equation

\[
\begin{cases}
  \Delta u = 0 \quad \text{in } \Omega, \\
  u = \phi \quad \text{on } \partial \Omega,
\end{cases}
\]
where we let \( \Omega \) be the annulus
\[
\Omega = \{(x, y) \in \mathbb{R}^2 : a^2 < x^2 + y^2 < b^2\}
\]
with \( 0 < a < b \). That is, \( \Omega \) is a ring-shaped domain bounded by two concentric circles with radii \( a \) and \( b \), respectively. In order comply with the radiality of \( u \), the Dirichlet boundary condition \( \phi : \partial \Omega \to \mathbb{R} \) must take the following form
\[
\phi(r) = \begin{cases} 
A, & r = a, \\
B, & r = b,
\end{cases}
\]
for two numbers \( A, B \). We can use these two conditions to determine the constants \( c_1 \) and \( c_2 \) in (1.2.10).

The \( d \)-dimensional boundary value problem
\[
\begin{cases} 
\Delta u = 0 & \text{in } B(0,1), \\
u(x) = \phi & \text{on } \partial B(0,1),
\end{cases}
\]
where \( B(0,1) = \{x \in \mathbb{R}^d : |x| < 1\} \) is the \( d \)-dimensional unit ball, can be shown to be solved by Poisson's formula
\[
u(x) = \frac{1 - |x|^2}{d\alpha(d)} \int_{\partial B(0,1)} \frac{\phi(y)}{|x-y|} dS(y),
\]
where \( \alpha(d) \) denotes the volume of \( B(0,1) \).

### 1.2.5 Heat Equation

We demonstrate Fourier’s method for solving the heat equation in one space dimension with homogeneous boundary conditions
\[
\begin{cases} 
u_t = \alpha u_{xx} & \text{in } (0, L) \times (0, \infty), \\
u = 0 & \text{for } x = 0 \text{ and } x = L, \\
u = \phi & \text{at } t = 0.
\end{cases}
\]

(1.2.11) The first step is to separate variables meaning that we look for a solution having the following form
\[u(x, t) = X(x)T(t)\]

We try to determine the functions \( X, T \) by first plugging their product into the heat equation yielding \( XT'' = \alpha X''T \). Dividing by \( \alpha XT \) gives the following identity
\[
\frac{T'}{\alpha T} = \frac{X''}{X}.
\]
The fraction on the left is independent of \( x \), while the one on the right is independent of \( t \). It follows that both fractions must be equal to a constant, which we denote by \(-\lambda\) for reasons that will become clear below. Now the heat equation boils down to a pair of independent ODEs

\[
\begin{align*}
T' &= -\lambda \alpha T, \\
X'' &= -\lambda X.
\end{align*}
\]

First, assume \( \lambda = 0 \). Then \( X(x) = Bx + C \), and the boundary conditions imply that \( B = C = 0 \). Therefore, the trivial \( u \equiv 0 \) is the only solution. Next assume \( \lambda < 0 \). Now we get \( X(x) = Be^{\sqrt{-\lambda}x} + Ce^{-\sqrt{-\lambda}x} \). Again the boundary conditions imply that \( u \equiv 0 \). Therefore, in order to have a nontrivial solution, \( \lambda > 0 \) must hold, which gives

\[
\begin{align*}
T(t) &= A \exp (-\lambda \alpha t), \\
X(x) &= B \cos(x \sqrt{\lambda}) + C \sin(x \sqrt{\lambda}).
\end{align*}
\]

The boundary condition at \( x = 0 \) implies that \( B = 0 \), while at \( x = L \) we obtain \( 0 = C \sin(L \sqrt{\lambda}) \). Since we are not interested in \( C = 0 \), which would lead to \( u \equiv 0 \) once again, we must have \( L \sqrt{\lambda} = k \pi \) for any integer \( k \). Thus, with the definition

\[ \lambda_k = \left( \frac{k \pi}{L} \right)^2, \quad k \in \mathbb{N}, \]

we obtain the following collection of solutions to pair of ODEs above

\[
\begin{align*}
T_k(t) &= \exp (-\lambda_k \alpha t) \\
X_k(x) &= \sin \left( x \sqrt{\lambda_k} \right)
\end{align*}
\]

Since the PDE in (1.2.11) is linear and homogeneous, every linear combination

\[ u(x, t) = \sum_k A_k T_k(t) X_k(x), \quad (1.2.12) \]

where \( A_k \) are constants, is again a solution of (1.2.11) provided that the initial condition is satisfied:

\[ \phi(x) = \sum_k A_k \sin \left( \frac{k \pi x}{L} \right). \quad (1.2.13) \]

This might seem like a very restrictive assumption on \( \phi \), but it turns out that practically any function \( \phi : (0, L) \to \mathbb{R} \) can be expressed as a Fourier sine series of the form (1.2.13) as long as the sum is infinite and the \( A_k \) are the Fourier coefficients of \( \phi \)

\[ A_k = \frac{2}{L} \int_0^L \phi(x) \sin \left( x \sqrt{\lambda_k} \right) \, dx. \]
Moreover, it can be shown that in this case the corresponding infinite linear combination in (1.2.12) remains a solution to (1.2.11).

In $d$ space dimensions, $x \in \mathbb{R}^d$, the separation ansatz $u(x,t) = T(t)v(x)$ for the heat equation leads to the following eigenvalue problem for $v$

$$\Delta v = -\lambda v,$$

combined with the given boundary conditions of the $d$-dimensional heat equation. If there are eigenvalues $\lambda_k$ and eigenfunctions $v_k$, then again any linear combination

$$u(x,t) = \sum_k A_k \exp(-\lambda_k \alpha t)v_k(x),$$

will solve the heat equation. As before the coefficients $A_k$ are determined by the initial condition.

### 1.3 Well-posed Problems

Employing tools like coordinate transformations, the method of characteristics, separation of variables and Fourier expansions, we have been able to find explicit expressions for the solutions to a few PDEs. Unfortunately, it is a fundamental truth that for most PDEs this is not possible. On the upside, however, it often is possible to deduce important properties of solutions of PDEs without actually solving them. One of the most fundamental properties is well-posedness.

A mathematical problem, in our case a PDE with initial and/or boundary conditions, is said to be well-posed, if it satisfies the following three conditions.

**Existence** There is at least one solution.

**Uniqueness** There is at most one solution.

**Stability** The solution depends in a stable way on the data of the problem.

That is, a slight perturbation of the data should lead to a small change in the solution.

By data we mean the collection of all given quantities that specify the problem, that is, initial and boundary condition as well as inhomogeneities. Consider, for instance, the inhomogeneous heat equation

$$\begin{cases}
  u_t - \Delta u = f & \text{in } \Omega \times (0, \infty), \\
  u = \psi & \text{on } \partial \Omega \times [0, \infty), \\
  u = \phi & \text{at } t = 0.
\end{cases}$$

Here, the data are the three functions $f(x,t)$, $\psi(x,t)$ and $\phi(x)$. Stability means that, if one or more of these functions is slightly changed, then the resulting change in $u$ should be small too. Such a statement requires a notion of “distance” for functions, which typically comes in the form of a function norm.
The heat equation, as well as all the other initial and boundary value problems considered in Sections 1.2.2 to 1.2.5, can be shown to be well-posed under the right assumptions and with the right “distances”. In contrast, backward diffusion, i.e. the problem of determining \( u(x, t) \) from \( u(x, T) \) where \( t < T \), is a typical example of an ill-posed problem.

**Example 1.5.** Consider the problem of solving a linear system of equations \( Ax = b \). In this case the data is just the vector \( b \). If the system is overdetermined (more equations than unknowns), then there is no solution \( x \). If, however, \( b \) lies in the range of \( A \), then there is at least one solution. If \( A \) is invertible, then there is exactly one vector \( x \) such that \( Ax = b \). Concerning stability recall that a small perturbation in \( b \) leads to a small perturbation in \( x \), if \( A \) is a well-conditioned matrix. That is, solving \( Ax = b \) is a well-posed problem in the above sense, if the condition number \( \|A\|\|A^{-1}\| \) is small.

**Example 1.6.** There are fundamental PDEs for which even the basic question of existence has not been resolved. As a famous example consider the Navier-Stokes equations, which describe the motion of fluids and form one of the pillars of fluid mechanics. For an incompressible fluid filling all of \( \mathbb{R}^3 \) the Navier-Stokes equations consist of the following nonlinear system of PDEs

\[
\begin{cases}
  u_t + J_u u = \nu \Delta u - \nabla p + f & \text{in } \mathbb{R}^3 \times (0, \infty), \\
  \text{div} \: u = 0 & \text{in } \mathbb{R}^3 \times (0, \infty), \\
  u = \phi & \text{at } t = 0.
\end{cases}
\]

The unknowns are the velocity field \( u : \mathbb{R}^3 \times (0, \infty) \to \mathbb{R}^3 \) and the pressure of the fluid \( p : \mathbb{R}^3 \times (0, \infty) \to \mathbb{R} \). The matrix \( J_u = (u_i)_i^j \) is the Jacobian of \( u \), \( \nu > 0 \) is the viscosity of the fluid, \( f : \mathbb{R}^3 \times (0, \infty) \to \mathbb{R}^3 \) is an external force such as gravity, and \( \phi : \mathbb{R}^3 \to \mathbb{R}^3 \) is a divergence-free initial velocity.

It is an open problem whether this initial value problem admits smooth global solutions \( u, p \) that are physically reasonable (i.e. of finite energy). In fact, in May 2000 the Clay Mathematics Institute included this problem in its list of seven Millenium Prize Problems,\(^3\) for which there is a total prize fund of 7,000,000 USD. The precise official statement of the Navier-Stokes problem was given by Charles Fefferman.\(^4\)

A common requirement for many mathematical models, besides being physically realistic, is well-posedness in the above sense. Concerning stability, keep in mind that in many applications the data of a PDE are obtained from real-world measurements, which can never be absolutely error-free. Moreover, well-posedness is also relevant from a numerical analyst’s point of view. The reason is that a well-posed PDE can usually be expected to lend itself to a stable numerical algorithm. In Chapters 2 and 3 we will often tacitly assume that the underlying PDE is well-posed.

---

\(^3\)http://www.ams.org/notices/200606/fea-jaffe.pdf
\(^4\)http://www.claymath.org/sites/default/files/navierstokes.pdf
1.4 Numerical Analysis of PDEs

The third main approach of studying PDEs, besides discovering analytical solutions and developing rigorous theory, is the design of numerical methods. Let us try to clarify what we generally expect from a numerical method.

Consider a general PDE in the following abstract form

\[ P(u, g) = 0, \]  

(1.4.1)

where \( g \) are the data of the problem, \( u \) is the solution and \( P \) specifies the relation between the two. A numerical method is a sequence of approximate problems

\[ P_N(u_N, g_N) = 0, \quad N \in \mathbb{N}, \]  

(1.4.2)

where \( g_N \) is an approximation of the original data, \( P_N \) specifies the form of the approximate problem, and \( u_N \) is the resulting solution. The parameter \( N \) controls the quality of approximation, in other words, small values should correspond to crude approximations, while increasing \( N \) means refining the approximation. Note that, in contrast to PDEs, approximate problems are typically finite-dimensional, meaning that the space of candidate solutions can in some way be identified with \( \mathbb{R}^n \). Consequently, transforming the original problem (1.4.1) into (1.4.2) involves a discretization process of some sort.

Ultimately, we want the numerical method to be convergent, meaning that

\[ \| u - u_N \| \to 0 \quad \text{as} \quad N \to \infty, \]

for an appropriately chosen norm \( \| \cdot \| \). With a convergent method we can make the error \( u - u_N \) arbitrarily small by simply increasing \( N \). It turns out that direct verification of convergence is often difficult. However, a recurring theme in the numerical analysis of PDEs is that

\[ \text{consistency} + \text{stability} \Rightarrow \text{convergence}. \]

A numerical method \( (P_N)_{N \in \mathbb{N}} \) is consistent, if

\[ P_N(u, g) \to 0 \quad \text{as} \quad N \to \infty, \]

while stability means the usual robustness of \( P_N \) with respect to small changes in the data \( g_N \). Roughly speaking consistency is just the basic requirement that the discretized problems \( P_N \) actually approximate \( P \) in the limit, while stability essentially asserts that each \( P_N \) is a well-posed problem.

1.5 Classification of PDEs

There are three main types of second-order linear PDEs: elliptic, parabolic and hyperbolic ones, the three prototypical representatives being Laplace’s, the heat and the wave equation. The classification is meaningful, because solutions to PDEs of one class will behave similarly in many ways, while being distinctly different from solutions to PDE of the other two classes.
1.5.1 The Two-variable Case

Consider first a general second-order linear differential operator in two independent variables

$$Lu = a_{11} u_{xx} + 2a_{12} u_{xy} + a_{22} u_{yy} + a_{1} u_x + a_{2} u_y + a_{0} u, \quad (1.5.1)$$

where we assume that at least one of the highest order coefficients $a_{11}$, $a_{12}$ or $a_{22}$ is nonzero. The classification of $L$, and consequently of any equation $Lu = f$, is based on these three coefficients. More specifically, define the discriminant of $L$ as

$$D = a_{11} a_{22} - a_{12}^2.$$ 

Then $L$ is called

- **elliptic**, if $D > 0$,
- **parabolic**, if $D = 0$,
- **hyperbolic**, if $D < 0$.

In order to make sense of this definition, consider a general linear coordinate transformation

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = B \begin{pmatrix} x \\ y \end{pmatrix},$$

for an invertible $2 \times 2$ matrix $B$. Then, using the chain rule as we have done before, we can transform $L$ and write the result in an analogous way

$$\tilde{a}_{11} u_{\xi\xi} + 2\tilde{a}_{12} u_{\xi\eta} + \tilde{a}_{22} u_{\eta\eta} + \tilde{a}_{1} u_{\xi} + \tilde{a}_{2} u_{\eta} + \tilde{a}_{0} u.$$ 

Denote the new discriminant by $\tilde{D} = \tilde{a}_{11} \tilde{a}_{22} - \tilde{a}_{12}^2$. Now we have the following result.

**Theorem 1.1.** The sign of the discriminant is invariant under linear coordinate transformations, that is, $\text{sgn} \, \tilde{D} = \text{sgn} \, D$. In addition, depending on $\text{sgn} \, D$, the operator (1.5.1) can be reduced in the following ways.

- **If** $D > 0$, **then there is a** $B$ **such that** $Lu = u_{\xi\xi} + u_{\eta\eta} + l. \ o. \ t.$
- **If** $D = 0$, **then there is a** $B$ **such that** $Lu = u_{\xi\xi} + l. \ o. \ t.$
- **If** $D < 0$, **then there is a** $B$ **such that** $Lu = u_{\xi\xi} - u_{\eta\eta} + l. \ o. \ t.$

We postpone the proof until Section 1.5.2.

Theorem 1.1 essentially says that, up to lower order terms, an elliptic operator can be reduced to the Laplace operator. Note that a hyperbolic operator can not only be transformed into the form given by Theorem 1.1, but also—according to Example 1.3—into a single mixed term $u_{\xi\eta} + l. \ o. \ t.$
Example 1.7. Let $D = 0$. A change of variables that transforms $L$ in the way postulated by Theorem 1.1 is given by

$$ B = \begin{pmatrix} -2a_{21} & 2a_{11} \\ 1 & 0 \end{pmatrix}. $$

It satisfies $a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy} = a_{22}u_{\eta\eta}.$

Remark 1.1. The three classes of PDEs are named after conic sections because of the following analogy. Let us turn the differential equation $Lu = f$ into an algebraic one by replacing $u_{xx}$ with $x^2$, $u_{xy}$ with $xy$ and so on. We obtain

$$ a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + a_1x + a_2y + a_0 = f. \quad (1.5.2) $$

If the operator (1.5.1) is elliptic, then the set of all points $(x, y) \in \mathbb{R}^2$ satisfying this equation turns out to be an ellipse. Similarly, in the parabolic case equation (1.5.2) describes a parabola, while in the hyperbolic case we get a hyperbola. Note that, however, the conic sections obtained in this way include degenerate cases. Therefore, in order to have “typical” conic sections also the lower order coefficients must be set to appropriate values. For example, setting all coefficients to zero except $a_{11} = -a_2 = 1$ gives the well-known parabola $y = x^2$ (and $D = 0$).

1.5.2 The General Case

Now consider a second-order linear operator in $d$ independent variables $x = (x_1, \ldots, x_d)$

$$ Lu = \sum_{i,j=1}^{d} a_{ij}u_{x_ix_j} + \sum_{i=1}^{d} a_{i}u_{x_i} + a_0u. \quad (1.5.3) $$

Note that we can assume $a_{ij} = a_{ji}$, because the mixed derivatives of $u$ agree. We want to find out how the matrix $A = (a_{ij})$ of second-order coefficients of $L$ changes under a linear coordinate transformation $\xi = (\xi_1, \ldots, \xi_d) = Bx$, where $B = (b_{ij}) \in \mathbb{R}^{d \times d}$ is a regular matrix. The chain rule yields

$$ \frac{\partial}{\partial x_j} = \sum_{k=1}^{d} b_{kj} \frac{\partial}{\partial \xi_k}, \quad \frac{\partial^2}{\partial x_i \partial x_j} = \left( \sum_{\ell=1}^{d} b_{i\ell} \frac{\partial}{\partial \xi_\ell} \right) \left( \sum_{k=1}^{d} b_{kj} \frac{\partial}{\partial \xi_k} \right) = \sum_{k, \ell=1}^{d} b_{kj}b_{i\ell} \frac{\partial^2}{\partial \xi_k \xi_\ell}. \quad (1.5.4) $$
Now the highest-order terms of $L$ transform according to
\[
\sum_{i,j=1}^{d} a_{ij} u_{x_i x_j} = \sum_{i,j=1}^{d} a_{ij} \left( \sum_{k,\ell=1}^{d} b_{kj} b_{\ell i} \frac{\partial^2}{\partial \xi_k \xi_\ell} \right) u
\]
\[
= \sum_{k,\ell=1}^{d} \left( \sum_{i,j=1}^{d} b_{\ell i} a_{ij} b_{kj} \right) \tilde{a}_{k\ell} u_{\xi_k \xi_\ell}.
\]

In matrix notation the relationship between the old and new coefficients becomes much clearer
\[
\tilde{A} = BAB^T,
\]
where $B^T$ denotes the transpose of $B$.

Now, the spectral theorem guarantees, for every symmetric matrix $A$, the existence of an orthogonal matrix $B$ such that $BAB^T$ is a diagonal matrix whose entries are the eigenvalues of $A$. This means that every operator of the form (1.5.3) can be linearly transformed into an operator without mixed terms. The spectral theorem helps us prove Theorem 1.1.

**Proof of Theorem 1.1.** First observe that in two dimensions
\[
A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}
\]
and therefore $D = \det A$. Now we exploit the properties of determinants to compute $\tilde{D}$
\[
\tilde{D} = \det \tilde{A} = \det(BAB^T) = (\det B)^2 \det A = (\det B)^2 D.
\]
And since $(\det B)^2 > 0$, the signs of $D$ and $\tilde{D}$ must agree. \qed

We now classify the operator $L$ according to its eigenvalues:

**Definition 1.1.** The operator $L$ as defined in (1.5.3) is **elliptic**, if all eigenvalues of its second-order coefficient matrix $A$ are positive or if all are negative. It is **parabolic**, if exactly one eigenvalue is zero and all the others have the same sign. It is **hyperbolic**, if one eigenvalue is positive and all others are negative, or vice versa.

**Remark 1.2.**

- Note that $L$ being elliptic means that $A$ is a **definite** matrix (either positive or negative), while for a parabolic operator $A$ is **semidefinite**. If $L$ is hyperbolic, then $A$ is **indefinite**.

- Definition 1.1 is consistent with the one given in Section 1.5.1, because $D = \det(A) = \lambda_1 \lambda_2$, where $\lambda_1$ and $\lambda_2$ are the eigenvalues of $A$. 
• In general the coefficients of an operator are not constant but depend on the same variables as the function \( u \). In this case a differential operator can be of one type in one region and of another type in another region. For example, the equation \( xu_{xx} + u_{yy} = f \) is elliptic for \( x > 0 \), parabolic for \( x = 0 \) and hyperbolic for \( x < 0 \).

• In applications, elliptic PDEs usually turn up in stationary (time-independent) problems, while parabolic and hyperbolic PDEs model phenomena where one of the independent variables is time. Therefore, the following definitions of parabolic and hyperbolic operators are sometimes used. The operator \( \partial / \partial t + L \) is parabolic if, for each fixed time \( t \), \( L \) is an elliptic operator in the space variables. Identifying, without loss of generality,\(^5\) ellipticity with positive definiteness of \( A \), we say \( \partial^2 / \partial t^2 - L \) is hyperbolic, if \( L \) is elliptic in the space variables for each \( t \).

• For the sake of completeness we mention that the linear transport equation \( u_t + b \cdot \nabla u = 0 \) is sometimes referred to as a first-order hyperbolic equation. Refraining from a thorough explanation we make do with recalling Section 1.2.3, where we have reduced the 1D wave equation to a system of transport equations.

---

\(^5\)If \( A \) is negative definite, multiply the equation \( Lu = f \) with \(-1\) to obtain an equivalent PDE with positive definite \( A \).
The basic idea of finite differences is simple. First, replace the domain of the PDE with a set of grid points. Next, approximate all derivatives by difference quotients. This turns a linear PDE into linear system of (algebraic) equations. The solution of this linear system should approximate the values of $u$ at the grid points. If, by refining the grid, we are able to make this approximation arbitrarily good, our finite difference method is convergent. Different finite difference methods for one and the same problem vary mainly in the way differential operators are discretized. Finite difference schemes are typically easy to derive, but can be difficult to analyse.

### 2.1 Finite Difference Approximations

Let $u$ be a function of a single variable. Suppose we want to approximate its derivative at a point $x$, given the values of $u$ at $x$ itself and at another point $y$. The definition of $u'(x)$ as a limit of difference quotients suggests to simply use

$$u(y) - u(x) \over y - x.$$  \hspace{1cm} (2.1.1)

Visually, this is the slope of the secant line that passes through the points $(x, u(x))$ and $(y, u(y))$. How can we assess the quality of such an approximation? First, we could observe that if $u$ were a linear function $u(x) = ax + b$, then the difference quotient above would actually produce the exact result $u'(x) = a$. Can we say anything about more general functions?

### 2.1.1 First Order Derivatives

Recall Taylor’s theorem, which tells us that, if $u$ is sufficiently well-behaved, then it can be approximated by Taylor polynomials. For instance, letting $y = x + h$ and assuming $u \in C^4$ we have

$$u(x + h) = u(x) + hu'(x) + \frac{h^2}{2} u''(x) + \frac{h^3}{6} u'''(x) + O(h^4).$$  \hspace{1cm} (2.1.2)
Here, we have used the big $O$ notation $O(h^4)$ as a placeholder for an expression about whose specific form we do not care, as long as, for $h \to 0$, it tends to 0 about as fast as the monomial $h^4$ or faster. Rearranging the equation above gives

$$\frac{u(x + h) - u(x)}{h} - u'(x) = \frac{h}{2} u''(x) + \frac{h^2}{6} u'''(x) + O(h^3)$$

(2.1.3)

$$= O(h).$$

For sufficiently small $h$, the dominating term on the right hand side of (2.1.3) is $u''(x)h/2$, which is linear in $h$. That is, if we reduce the distance $h$ between $x$ and $y$, then the difference between the derivative of $u$ and the difference quotient decays linearly in $h$. Thus we have found a qualitative statement about how well $u'(x)$ is approximated by a simple difference quotient for very general class of functions $u$.

We introduce some terminology. Let $h > 0$. Then

$$D^+_h u(x) = \frac{u(x + h) - u(x)}{h}$$

is called a forward difference, while

$$D^-_h u(x) = \frac{u(x) - u(x - h)}{h}$$

is a backward difference. Both are one-sided approximations to $u'$. Furthermore, motivated by the estimate (2.1.3) in the previous paragraph, we call them first order accurate.

Very often, two-sided, or centred, approximations work better than one-sided ones. From the forward and backward differences we can obtain a centred one by simply averaging them. This gives the centred difference

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

which can be visualized as the slope of the line passing through the points $(x - h, u(x - h))$ and $(x + h, u(x + h))$. In order to analyse the order of accuracy of this approximation we subtract from (2.1.2) the Taylor expansion

$$u(x - h) = u(x) - hu'(x) + \frac{h^2}{2} u''(x) - \frac{h^3}{6} u'''(x) + O(h^4)$$

(2.1.4)

and rearrange terms to find that

$$D^0_h u(x) - u'(x) = \frac{h^2}{6} u'''(x) + O(h^3) = O(h^2).$$

(2.1.5)

Therefore, the centred difference is second order accurate. Since the error involves only derivatives of order 3 or higher, the centred difference returns the exact value $u'(x)$ for every quadratic function $u(x) = ax^2 + bx + c$. 
Finite Difference Approximations

Table 2.1: Errors in approximating \( \sin'(1) \) with forward, backward and centred differences.

<table>
<thead>
<tr>
<th>( h )</th>
<th>( D_h^+ )</th>
<th>( D_h^- )</th>
<th>( D_h^0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1E-1</td>
<td>-4.29E-2</td>
<td>4.11E-2</td>
<td>-9.00E-4</td>
</tr>
<tr>
<td>5E-2</td>
<td>-2.13E-2</td>
<td>2.08E-2</td>
<td>-2.25E-4</td>
</tr>
<tr>
<td>1E-2</td>
<td>-4.22E-3</td>
<td>4.20E-3</td>
<td>-9.01E-6</td>
</tr>
<tr>
<td>5E-3</td>
<td>-2.11E-3</td>
<td>2.10E-3</td>
<td>-2.25E-6</td>
</tr>
<tr>
<td>1E-3</td>
<td>-4.21E-4</td>
<td>4.21E-4</td>
<td>-9.01E-8</td>
</tr>
</tbody>
</table>

**Remark 2.1.** Orders of accuracy can be verified experimentally. Pick a well-behaved function, for instance, \( u(x) = \sin x \). Suppose we want to approximate its derivative at \( x = 1 \). Pick a finite difference approximation \( D_h \) and compute the errors \( D_h u(1) - u'(1) \) for decreasing values of \( h \). If \( D_h \) is \( k \)-th order accurate, then whenever \( h \) is decreased by a factor of 2, say, the error should reduce by a factor of \( 2^k \). See Table 2.1. Figures are taken from Table 1.1 in [6].

### 2.1.2 Higher Order Derivatives

A common finite difference approximation for \( u''(x) \) is the centred second difference

\[
D_h^2 u(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) + \frac{h^2}{12} u'''(x) + O(h^4).
\]  

(2.1.6)

The second identity, which shows that \( D_h^2 \) is second order accurate, can be obtained by replacing \( u(x+h) \) and \( u(x-h) \) by their Taylor expansions up to fourth order, compare (2.1.2) and (2.1.4).

The centred second difference can also be obtained by computing the forward difference of a backward difference or vice versa. For instance,

\[
D_h^+ D_h^- u(x) = \frac{1}{h} \left( (D_h^- u(x+h) - D_h^- u(x)) \right) = \frac{1}{h^2} \left( \frac{u(x+h) - u(x) - u(x) + u(x-h)}{h} \right) = D_h^2 u(x).
\]

Observe that by composing two first order accurate approximations for \( u' \) we have obtained a second order accurate one for \( u'' \)! In general this cannot be expected. Here it works, however, since the two asymmetric one-sided differences add up to a symmetric two-sided second difference. Alternatively, \( D_h^2 \) can be obtained by composing two centred differences each of stepsize \( h/2 \)

\[
D_h^2 u(x) = D_{h/2}^0 D_{h/2}^0 u(x).
\]
As a general rule approximations to higher order derivatives can be found by composing lower order differences. For instance,

\[ D_h^2 u(x) = \frac{1}{2h^3} \left( u(x + 2h) - 2u(x + h) + 2u(x - h) - u(x - 2h) \right) \]

is a second order accurate approximation of \( u'''(x) \). Another option to find general finite difference approximations is the method of \textit{undetermined coefficients}. We explain this method by means of an example.

\textbf{Example 2.1.} Suppose we are given values \( u(x), u(x + h), u(x + 2h) \) and we want to find a one-sided approximation to \( u'(x) \) — let us call it \( D_h u(x) \) — which is better than first order accurate. This seems like a reasonable goal, since with one additional point we must be able to do at least as good as \( D_+ u(x) \). First, observe that all difference operators we have encountered so far turned out to be linear combinations of values of \( u \). Therefore we make the ansatz

\[ D_h u(x) = au(x) + bu(x + h) + cu(x + 2h) = u'(x) + O(h^2). \quad (2.1.7) \]

Now we replace \( u(x + h) \) by its expansion (2.1.2) and \( u(x + 2h) \) by

\[ u(x + 2h) = u(x) + 2hu'(x) + \frac{(2h)^2}{2}u''(x) + \frac{(2h)^3}{6}u'''(x) + O(h^4) \]

and collect terms giving

\[ D_h u(x) = (a + b + c)u(x) + (b + 2c)hu'(x) + (b + 4c)\frac{h^2}{2}u''(x) \]

\[ + (b + 8c)\frac{h^3}{6}u'''(x) + O(h^4). \quad (2.1.8) \]

Now, we want as many terms as possible to be zero, except for \( (b + 2c)hu'(x) \), which should equal \( u'(x) \). (Recall the second equality in (2.1.7)) On the other hand, we only have three coefficients, which cannot solve more than the following three equations

\[ a + b + c = 0 \]
\[ b + 2c = \frac{1}{h} \]
\[ b + 4c = 0. \]

The solution to this system is given by

\[ a = -\frac{3}{2h}, \quad b = \frac{2}{h}, \quad c = -\frac{1}{2h}. \]

Therefore

\[ D_h u(x) = \frac{1}{2h} \left( -3u(x) + 4u(x + h) - u(x + 2h) \right) \]
is a second order accurate one-sided approximation for \( u'(x) \), since according to (2.1.8) we have
\[
= u'(x) - \frac{h^2}{3} u'''(x) + O(h^3).
\]

In general the method of undetermined coefficients can be used to find finite difference approximations for the \( k \)-th derivative \( u^{(k)}(x) \), given \( n > k \) function values \( u(x_i) \) at arbitrary points \( x_1 < x_2 < \cdots < x_n \). The expected order of accuracy of the obtained difference operator is at least \( n - k \). Note, however, that the resulting linear system is a Vandermonde system, which can become extremely ill-conditioned for increasing \( n \), so that some care is needed for determining the coefficients in a stable way.

2.2 Elliptic Problems

We first consider the finite difference method for a model problem and introduce several key concepts along the way. Afterwards we discuss how to extend this method to more general problems.

Our model problem is Poisson’s equation on the unit square \( \Omega = (0,1) \times (0,1) \subset \mathbb{R}^2 \) with Dirichlet boundary conditions
\[
\begin{aligned}
\Delta u &= f \quad \text{in } \Omega, \\
u &= \phi \quad \text{on } \partial \Omega.
\end{aligned}
\tag{2.2.1}
\]

### 2.2.1 Deriving the Difference Equations

First we replace \( \Omega \) by a discrete domain. Let \( N \in \mathbb{N} \) and denote by \( \Omega_h \) the Cartesian grid with step size \( h = 1/N \) (in both \( x \) and \( y \) directions). That is,
\[
\Omega_h = \{(x_i, y_j) = (ih, jh) : i, j = 0, \ldots, N\}.
\]

Our aim is to find approximations to the values \( u^*(x_i, y_j) \), where \( u^* \) is the exact solution of (2.2.1).

We approximate the Laplace operator by using centred second differences
\[
\Delta u(x_i, y_j) = u_{xx}(x_i, y_j) + u_{yy}(x_i, y_j)
\]
\[
\approx \frac{1}{h^2} (u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + \frac{1}{h^2} (u_{i,j+1} - 2u_{i,j} + u_{i,j-1})
\]
\[
= \frac{1}{h^2} (u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j})
\]
\[
=: \Delta_h u_{i,j}.
\]

The difference operator \( \Delta_h \) can be visualized by the so-called five-point stencil
\[
\frac{1}{h^2} \begin{bmatrix}
1 & -4 & 1 \\
-1 & 4 & -1 \\
1 & -4 & 1
\end{bmatrix}.
\tag{2.2.2}
\]
Setting \( f_{i,j} = f(x_i, y_j) \) we obtain the following system of linear equations
\[
\Delta_h u_{i,j} = f_{i,j}, \quad 1 \leq i, j \leq N - 1.
\] (2.2.3)

Near the boundary, that is, if at least one of the indices equals 1 or \( N - 1 \), we replace the boundary values of \( u \) by the corresponding values of \( \phi \). This way we arrive at \((N - 1)^2\) equations for the \((N - 1)^2\) unknowns \( u_{i,j}, 1 \leq i, j \leq N - 1 \).

We would like to write system (2.2.3) in matrix-vector form. In order to do so we first consider the equations near the boundary and move the values of \( \phi \) to the right hand side. Next we have to rearrange the unknowns as well as the right hand sides into vectors \( u_h, f_h \in \mathbb{R}^{M^2} \), where \( M = N - 1 \). This can be done in many ways, one possibility is to order horizontally, that is, we define
\[
\begin{bmatrix}
  u_{1,j} \\
  \vdots \\
  u_{N-1,j}
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
  u[1] \\
  \vdots \\
  u[N-1]
\end{bmatrix}.
\]

Now we can write system (2.2.3) as
\[
A_h u_h = f_h,
\] (2.2.4)

where the system matrix \( A_h \in \mathbb{R}^{M^2 \times M^2} \) takes the following symmetric, block tridiagonal form
\[
A_h = \frac{1}{h^2}
\begin{bmatrix}
  T & I \\
  I & T & I \\
  \vdots & \ddots & \ddots & \ddots \\
  I & T & I \\
  I & T
\end{bmatrix},
\] (2.2.5)

where \( I \) is the \( M \)-dimensional identity matrix and the blocks \( T \) are given by
\[
T = \begin{bmatrix}
  -4 & 1 & & \\
  1 & -4 & 1 & \\
  & \ddots & \ddots & \ddots \\
  & & 1 & -4 & 1 \\
  & & & 1 & -4
\end{bmatrix} \in \mathbb{R}^{M \times M}.
\]

**Example 2.2.** Consider Laplace’s equation on the unit square, that is, problem (2.2.1) where \( f \equiv 0 \). As boundary condition we pick \(^1\)
\[
\phi(x, y) = \begin{cases}
  324x^2(1 - x), & y = 0 \\
  0, & \text{otherwise}.
\end{cases}
\]

Letting \( N = 3 \) we obtain the \( 4 \times 4 \) grid \( \Omega_{1/3} = \{ \frac{1}{3}(i, j) : i, j = 0, \ldots, 3 \} \), which has four interior points and twelve boundary points. On the latter the solution

\(^1\)This example is taken from Chapter 8.4 in [9].
is already determined by the values of $\phi$. We can visualize the discrete problem with the following matrix

\[
\begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & u_{1,2} & u_{2,2} & 0 \\
0 & u_{1,1} & u_{2,1} & 0 \\
0 & 24 & 48 & 0
\end{pmatrix}
\quad (2.2.6)
\]

where we have to fill in the four missing interior values $u_{1,1}, u_{2,1}, u_{1,2}, u_{2,2}$ by solving the equations (2.2.3). In our example these equations read

\[
u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0, \quad i,j = 1,2.
\]

As a side remark, observe that this equation implies that $u_{i,j}$ is the average of its four neighbours! This is a discrete version of the mean value property of harmonic functions. See for instance Chapters 6 and 7 in [9].

Directly plugging in the contributions of $\phi$ we obtain the following matrix-vector form of the four linear equations above

\[
\begin{pmatrix}
-4 & 1 & 1 \\
1 & -4 & 1 \\
1 & -4 & 1 \\
1 & 1 & -4
\end{pmatrix}
\begin{pmatrix}
u_{1,1} \\
u_{2,1} \\
u_{1,2} \\
u_{2,2}
\end{pmatrix}
= \begin{pmatrix}
-24 \\
-48 \\
0 \\
0
\end{pmatrix}.
\]

Compare to (2.2.4), (2.2.5). Its solution $u_h = [11, 16, 4, 5]^T$ we can fill into (2.2.6):

\[
\begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 4 & 5 & 0 \\
0 & 11 & 16 & 0 \\
0 & 24 & 48 & 0
\end{pmatrix}
\]

**Remark 2.2.** With increasing $N$, the matrix $A_h$ quickly becomes very large and sparse, and even more so for three-dimensional problems. For solving such systems iterative methods are often preferred to direct solvers. When using Matlab it is important to explicitly assemble $A_h$ as a sparse matrix. The following lines of code, taken from [6], are a sensible way to do so.

Algorithm 2.1: Assembly of $A_h$ in Matlab

```matlab
1 I = eye(M);
2 e = ones(M,1);
3 T = spdiags([e -4*e e],[-1 0 1],M,M);
4 S = spdiags([e e],[ -1 1],M,M);
5 A = (kron(I,T) + kron(S,I))/h^2;
```

Its main ingredients are the commands `spdiags` which creates a sparse band matrix by arranging the input vectors along (sub-)diagonals, and `kron` which
implements the Kronecker product. The Kronecker product \( A \odot B \) of a \( k \times \ell \) matrix \( A \) and an \( m \times n \) matrix \( B \) is the \( km \times \ell n \) matrix defined by

\[
A \odot B = \begin{bmatrix}
a_{11}B & \cdots & a_{1n}B \\
\vdots & \ddots & \vdots \\
a_{n1}B & \cdots & a_{nn}B
\end{bmatrix}.
\]

### 2.2.2 Consistency, Stability and Convergence

How well does the solution \( u_h \) of (2.2.4) approximate the exact solution \( u^* \) of Poisson’s equation?

**Definition 2.1.** We first look at the local truncation error

\[
\tau_{i,j} = \Delta_h u_{i,j}^* - f_{i,j}.
\]

If we arrange the values \( \tau_{i,j} \) into a long vector \( \tau_h \in \mathbb{R}^{M^2} \) (in the same way as \( u_{i,j} \) and \( f_{i,j} \)) we can write this definition as

\[
\tau_h = A_h u^*_h |_{\Omega_h} - f_h,
\]

where \( u^*_h |_{\Omega_h} \) denotes the exact solution of the PDE evaluated at the grid points and also rearranged into a vector.

The local truncation error basically tells us how well the solution of the PDE solves the discrete equation, or how well the discrete equation approximates the exact one. Note that we have \( f_{i,j} = \Delta u_{i,j}^* \) and therefore

\[
\tau_{i,j} = \Delta_h u_{i,j}^* - \Delta u_{i,j}^*.
\]

Thus, the local truncation error measures the approximation quality of the finite difference operator \( \Delta_h \).

**Definition 2.2.** If the truncation errors satisfies

\[
\tau_{i,j} \to 0 \quad \text{as} \quad h \to 0,
\]

for all \( i, j = 1, \ldots, M \), then the finite difference method is consistent. If, in addition, it can be shown that

\[
\tau_{i,j} = O(h^k)
\]

for some \( k > 0 \), then its order of consistency is \( k \).

Since the centred second difference, which we used to approximate the Laplacian, is second order accurate, recall (2.1.6), it follows that \( \tau_{i,j} = O(h^2) \). This tells us that our finite difference method is consistent and that its order of consistency is two.
Definition 2.3. The finite difference method is stable, if the system matrix $A_h$ is invertible and $\|A_h^{-1}\|$ admits an upper bound which is independent of $h$, that is,
\[ \|A_h^{-1}\| \leq C \]
for some constant $C > 0$.

It turns out that the finite difference scheme derived above is stable, because the matrix in (2.2.5) satisfies the following stability estimates.

Theorem 2.1. The matrix $A_h$ as given in (2.2.5) is regular. Its inverse satisfies
\[
\|A_h^{-1}\|_p \leq \begin{cases} 
\frac{1}{2\pi^2} + O(h^2), & p = 2 \\
\frac{1}{8}, & p = \infty.
\end{cases}
\]

Proof. We only prove the bound for the spectral norm, but first we have to show that $A_h$ is actually invertible. Consider the functions
\[ v^{k,\ell}(x, y) = \sin(k\pi x) \sin(\ell\pi y), \quad k, \ell = 1, \ldots, M. \]
These functions, evaluated at the grid points $(x_i, y_j)$ and rearranged into vectors, are eigenvectors of the matrix $A_h$. More precisely,
\[ A_h v^{k,\ell} \big|_{\Omega_h} = \lambda^{k,\ell} v^{k,\ell} \big|_{\Omega_h} \]
with eigenvalues
\[ \lambda^{k,\ell} = \frac{4}{h^2} \left[ \sin^2 \left( \frac{k\pi h}{2} \right) + \sin^2 \left( \frac{\ell\pi h}{2} \right) \right] > 0. \]
Since all eigenvalues are nonzero, the matrix is invertible. (In fact, it is an SPD matrix.)

Concerning the upper bound for $\|A_h^{-1}\|_2$, note that $A_h$ is symmetric and therefore the spectral norm of $A_h^{-1}$ is given by the reciprocal of the smallest eigenvalue of $A_h$. We compute
\[ \|A_h^{-1}\|_2 = \frac{1}{\lambda_{\text{min}}} = \frac{1}{\lambda^{1,1}} = \frac{h^2}{8\sin^2(\pi h/2)}. \]
Using the fact that the second order Taylor expansion of $\sin^2 x$ around zero is given by $x^2 + O(x^4)$ we further obtain
\[ = \frac{1}{2\pi^2 + O(h^2)} = \frac{1}{2\pi^2} + O(h^2), \]
where the last identity is another Taylor approximation. \qed
Next we turn to the quantity we ultimately would like to control: the global error.

**Definition 2.4.** The global error of $u_h$ is defined as the difference between $u_h$ and the exact solution $u^*$ evaluated at the grid points.

$$e_h = u_h - u^*|_{\Omega_h}.$$  

**Definition 2.5.** A finite difference method is called convergent, if $e_h \to 0$, as $h \to 0$, and convergent of order $k$, if $e_h = O(h^k)$.

The global error is related to the local truncation error via

$$A_h e_h = -\tau_h.$$  

Equation (2.2.7) says that the global error solves the same difference equations as $u_h$ but with $-\tau_h$ instead of $f_h$ on the right hand side.

We can now establish the main result of this section, that is, convergence of the finite difference method derived above, by means of a typical consistency and stability argument, recall Section 1.4.

**Theorem 2.2.** The finite difference scheme (2.2.3) for problem (2.2.1) is convergent of order two with respect to both $\| \cdot \|_2$ and $\| \cdot \|_\infty$.

**Proof.** Let $\| \cdot \|$ stand for one of the two norms.

$$\|e_h\| = \|A_h^{-1}\tau_h\| \leq \|A_h^{-1}\|\|\tau_h\| \leq C\|\tau_h\| = O(h^2).$$

**Remark 2.3.** The order of convergence of the finite difference method (2.2.3) equals the order of consistency, which again equals the order of accuracy of the centred second difference used to approximate the second derivatives $u_{xx}$ and $u_{yy}$. This suggests that using a higher order accurate finite difference approximation should directly translate into a higher order of convergence, as long as we can guarantee stability in the sense of Definition 2.3.

We can, for instance, use five points to obtain a fourth order accurate approximation of the second derivatives. This would lead to a stencil of the form

$$\frac{1}{h^2} \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix}$$

for the Laplacian, which, however, leads to problems close to the boundary and doubles the bandwidth of the resulting system matrix. A better idea is to use the following compact nine-point stencil

$$\frac{1}{h^2} \begin{bmatrix} 1 & 4 & 1 \\ 1 & -20 & 1 \\ 1 & 4 & 1 \end{bmatrix}.$$
In order to actually achieve fourth order using this stencil one has to replace the right hand side $f_{i,j} = f(x_i, x_j)$ of (2.2.3) by a weighted average of $f_{i,j}$ and its four closest neighbours

$$\frac{1}{12} (f_{i+1,j} + f_{i-1,j} + f_{i,j+1} + f_{i,j-1} + 8f_{i,j}).$$

This finite difference method for Poisson’s equation is called the *mehrstellen method*. It is fourth order consistent, the resulting system matrix $A_h$ can be shown to again satisfy $\|A_h^{-1}\|_\infty \leq 1/8$, and therefore the order of convergence is also four.

**Remark 2.4.** Recall Remark 2.1, where we have discussed a way of experimentally verifying the accuracy of a finite difference approximation. The order of convergence of a finite difference method can be verified experimentally in a very similar way. Consider an instance of Poisson’s equation for which you know the exact solution $u^*$. You can find such an instance by picking a well-behaved function $u^*$ such as $u^*(x,y) = \sin(xy)$. Then calculate $\Delta u^*$ to obtain $f$ and evaluate $u^*$ at the boundary of the domain to obtain $\phi$.

Suppose you have derived and implemented a $k$-th order convergent method. Run your algorithm for various values of $h$ and compute the global error

$$\|e_h\|_\infty = \max_{i,j} |u_{i,j} - u^*_{i,j}|$$

for each. You should observe the following behaviour: Whenever $h$ is halved, the error $\|e_h\|_\infty$ decreases by a factor of $2^k$.

### 2.2.3 The Neumann Problem

Let us replace the Dirichlet with a Neumann boundary condition and consider

$$\begin{cases}
\Delta u = f & \text{in } \Omega, \\
\frac{\partial u}{\partial n} = \phi & \text{on } \partial \Omega.
\end{cases} \quad (2.2.8)$$

Recall that $\partial u/\partial n = \nabla u \cdot n$, where $n$ is the outward pointing unit normal. Therefore, with $\Omega$ being the unit square $(0,1) \times (0,1)$ as before, the normal derivative on the boundary $\partial \Omega$ is given by

$$\frac{\partial u}{\partial n} = \begin{cases}
-u_x, & x = 0, \ 0 < y < 1, \\
u_x, & x = 1, \ 0 < y < 1, \\
-u_y, & y = 0, \ 0 < x < 1, \\
u_y, & y = 1, \ 0 < x < 1.
\end{cases}$$

It is usually advisable to discretize derivatives in the boundary or initial conditions with the same order of accuracy as the differential operator $L$. That means we should choose a second order accurate approximation for $u_x$ and $u_y$, such as
the centred difference. In order to be able to use the latter we can introduce *ghost variables*

\[ u_{-1,j}, u_{N+1,j}, \quad j = 0, \ldots, N, \]
\[ u_{i,1}, u_{i,N+1}, \quad i = 0, \ldots, N, \]

which we visualize as lying outside the original grid \( \Omega_h \). The resulting discretized boundary condition reads

\[
\begin{aligned}
&\frac{u_{-1,j} - u_{1,j}}{2h} = \phi_{0,j}, \quad j = 0, \ldots, N, \\
&\frac{u_{N+1,j} - u_{N-1,j}}{2h} = \phi_{N,j}, \quad j = 0, \ldots, N, \\
&\frac{u_{i,1} - u_{i,1}}{2h} = \phi_{i,0}, \quad i = 0, \ldots, N, \\
&\frac{u_{i,N+1} - u_{i,N-1}}{2h} = \phi_{i,N}, \quad i = 0, \ldots, N.
\end{aligned}
\]  

These \( 4(N + 1) \) equations complement the discretized PDE

\[
\Delta_h u_{i,j} = f_{i,j}, \quad 0 \leq i, j \leq N,
\]  

where, in contrast to the Dirichlet setting, the indices now run from 0 to \( N \) (and therefore the ghost variables also appear in (2.2.10).) In total we end up with \((N + 1)^2 + 4(N + 1)\) equations and unknowns. Note that there are exactly as many ghosts as there are equations in (2.2.9).

Concerning the matrix-vector form of this problem we have two options. We can either directly assemble a system matrix including the ghost variables, or we can use the boundary condition (2.2.9) to eliminate the ghosts in (2.2.10) and then set up a system only for the original unknowns \( u_{i,j}, i, j = 0, \ldots, N \). In either case, however, the linear system has two issues.

1. The system matrix resulting from (2.2.9) and (2.2.10) is singular. Therefore the linear system does not have a solution in general.
2. If it does have a solution, then it actually has infinitely many.

The lack of both existence and uniqueness is connected to analogous issues with the original problem.

**Theorem 2.3.**

1. A necessary condition for the existence of a solution to (2.2.8) is

\[
\int_{\Omega} f = \int_{\partial \Omega} \phi.
\]  

2. If \( u \) is a solution then \( u + c, c \in \mathbb{R} \), is another one.
Proof. For the first part we use the divergence theorem (see the Appendix). Since \( f = \Delta u = \text{div} \nabla u \) we have
\[
\int_{\Omega} f = \int_{\Omega} \text{div} \nabla u = \int_{\partial\Omega} \nabla u \cdot n = \int_{\partial\Omega} \phi.
\]

The second part of the theorem follows immediately from the fact that problem (2.2.8) only contains derivatives of \( u \).

Remark 2.5. The second part of Theorem 2.3 implies that, in order to have a unique solution, we have to prescribe the value of \( u(x_0, y_0) \) at one arbitrary point \((x_0, y_0) \in \Omega\).

Moreover, note that the lack of uniqueness is exclusive to the Neumann problem. The solution to the Dirichlet problem (2.2.1), if it exists, is always unique.

The solvability condition for the discrete Neumann problem (2.2.9), (2.2.10) takes a form similar to (2.2.11). It reads
\[
h^2 \sum_{i,j=0}^{N} d_id_j f_{i,j} = 2h \sum_{i,j} d_id_j \phi_{i,j}, \tag{2.2.12}
\]
where the second sum runs over all indices where \( \phi_{i,j} = \phi(x_i, y_j) \) is defined, that is, over all boundary points. The numbers \( d_k \) are given by \( d_0 = d_N = 1/2 \) and \( d_k = 1 \) otherwise.

Suppose that (2.2.12) is satisfied and that we have rewritten (2.2.9), (2.2.10) as \( A_h u_h = f_h \). As mentioned above this has system infinitely many solutions. We can compute a particular one by essentially following Remark 2.5. Pick an arbitrary index pair \((i_0, j_0)\) and set \( u_{i_0,j_0} = 0 \). Next, remove the corresponding entry from the vector of unknowns \( u_h \) and from the right hand side. Finally, also remove the row and column corresponding to \((i_0, j_0)\) from the matrix \( A_h \). Then this reduced system has a unique solution and one can show the following result.

Theorem 2.4. The finite difference method outlined above for problem (2.2.8) is convergent of order two.

Remark 2.6. Instead of using centred differences we could discretize the boundary condition with the one-sided approximation that we derived in Example 2.1. This would avoid the need for ghost variables but not the lack of uniqueness and existence. More generally, while affecting the specific form of the linear system, different consistent discretizations of the boundary condition do not avoid singularity of the resulting linear system. See Chapter 4 in [4] for more details as well as a proof for Theorem 2.4.

2.2.4 More General Problems

General domains. Extending the method discussed above to rectangular domains or to three-dimensional cuboid ones is straightforward. For more general
domains it might be necessary to use nonequidistant grids which can decrease the order of consistency. Another possibility is the finite element method (Chapter 3), which is more flexible with respect to the geometry of $\Omega$.

**Nonconstant diffusion coefficient.** Elliptic operators are sometimes given in *divergence form*

$$Lu = \text{div}(c\nabla u).$$

If the diffusion parameter $c$ is not a constant but depends on the same variables as $u$, then it is advisable *not* to differentiate out and discretize the resulting expression (which would be $\nabla c \cdot \nabla u + c \Delta u$). Instead a higher order of consistency can often be achieved by discretizing the operator directly. An example of such a direct discretization is

$$\frac{\partial}{\partial x} \left( c(x,y)u_x(x,y) \right) \bigg|_{x_i,y_j} + \frac{\partial}{\partial y} \left( c(x,y)u_y(x,y) \right) \bigg|_{x_i,y_j} \approx \frac{1}{h^2} \left( c_{i+\frac{1}{2},j}(u_{i+1,j} - u_{i,j}) - c_{i-\frac{1}{2},j}(u_{i,j} - u_{i-1,j}) \right) + \frac{1}{h^2} \left( c_{i,j+\frac{1}{2}}(u_{i,j+1} - u_{i,j}) - c_{i,j-\frac{1}{2}}(u_{i,j} - u_{i,j-1}) \right),$$

where $c_{i+\frac{1}{2},j} = c(x_i + \frac{h}{2}, y_j)$ etc. This approximation can be made sense of by first observing that a forward difference at $x$ can always be interpreted as a centred difference at $x + h/2$ of half the stepsize $D_{h/2}^0 g(x) = D_{h/2}^0 g(x + h/2)$. Therefore, denoting by $D_{h}^{0,x}$ the centred difference in $x$ direction, we can write

$$\frac{1}{h} c_{i+\frac{1}{2},j}(u_{i+1,j} - u_{i,j}) = c_{i+\frac{1}{2},j} D_{h}^{0,x} u_{i+\frac{1}{2},j} = (cD_{h}^{0,x}u)_{i+\frac{1}{2},j},$$

and analogously for the other three terms in (2.2.13). The resulting expressions can again be interpreted as centred differences, for instance,

$$\frac{1}{h^2} \left( c_{i+\frac{1}{2},j}(u_{i+1,j} - u_{i,j}) - c_{i-\frac{1}{2},j}(u_{i,j} - u_{i-1,j}) \right) = \frac{1}{h} \left( (cD_{h}^{0,x}u)_{i+\frac{1}{2},j} - (cD_{h}^{0,x}u)_{i-\frac{1}{2},j} \right) = D_{h/2}^{0,x} \left( cD_{h/2}^{0,x}u \right)_{i,j},$$

and the last line is clearly an approximation to $\frac{\partial}{\partial x} (cu_x)$ at $(x_i, y_j)$. 

\[34\]

*The Finite Difference Method*
2.3 Parabolic Problems

Our model problem in this section will be the one-dimensional heat equation with homogeneous Dirichlet boundary condition

\[ \begin{align*}
   u_t &= u_{xx}, \quad 0 < x < 1, \ t > 0 \\
   u &= 0, \quad x = 0, x = 1 \\
   u &= \phi, \quad t = 0.
\end{align*} \] 

(2.3.1)

2.3.1 An Explicit Scheme

As before, we first cover the domain \([0, 1] \times [0, \infty)\) by a discrete set of points

\((x_j, t_n) = (hj, kn), \quad j = 0, \ldots, J, \ n = 0, 1, \ldots,\)

where \(h = 1/J\) and \(k > 0\) are the stepsizes in space and time, respectively. Using centred second differences in space and forward differences in time yields

\[ \frac{u_j^{n+1} - u_j^n}{k} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}, \] 

(2.3.2)

where \(u_j^n = u(x_j, t_n)\). This is an explicit difference scheme, often called Euler method, since the solution at time \(t_{n+1}\) can be expressed explicitly in terms of the solution at time \(t_n\).

**Example 2.3.** Let us try to solve the explicit scheme for \(k = h^2\), in which case it simplifies to

\[ u_j^{n+1} = u_{j+1}^n - u_j^n + u_{j-1}^n. \] 

(2.3.3)

We set \(J = 8\) and pick the initial condition \(\phi_4 = 1\) and \(\phi_j = 0\) for all other \(j\). Starting from \(u_j^0 = \phi_j\), we can use (2.3.3) to compute \(u_j^n\) for all \(j\) and then proceed to \(n = 2\) and so on. We obtain

\[
\begin{array}{cccccccc}
    n = 0 & : & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
    n = 1 & : & 0 & 0 & 0 & 1 & -1 & 1 & 0 & 0 \\
    n = 2 & : & 0 & 0 & 1 & -2 & 3 & -2 & 1 & 0 \\
    n = 3 & : & 0 & 1 & -3 & 6 & -7 & 6 & -3 & 1 \\
\end{array}
\]

Is this a reasonable approximation to the exact solution of the heat equation? Recall Section 1.2.5, where we have shown that the solution can be written as an infinite sum

\[ u(x, t) = \sum_m A_m e^{-m^2 \pi^2 t} \sin(m \pi x). \]

Since every summand goes to zero as \(t \to \infty\), so does their sum. That means the exact solution will vanish eventually. It can also be shown that \(|u(x, t)| \leq \]

\footnote{This example is taken from Chapter 8 in [9]}
max $|\phi|$ holds for all $t$ and $x$. Both of these statements are clearly violated by the discrete solution we calculated above, which seems to actually explode rather quickly. Thus, the explicit scheme (2.3.3) is not convergent. It is also unstable in the sense that slightly changing $\phi$ leads to a very different solution eventually, i.e. for growing $n$. (Check for yourself by setting, for instance, $\phi_4 = 1 + \epsilon$.)

2.3.2 Von Neumann Stability Analysis

What went wrong in the previous example? In this section we pursue von Neumann’s approach to stability analysis in order to find an answer. It turns out that the crucial quantity is the ratio $k/h^2$.

**Theorem 2.5.** A necessary condition for stability of the explicit scheme (2.3.2) is

$$\frac{k}{h^2} \leq \frac{1}{2}.$$ 

**Proof.** Setting $s = k/h^2$ equation (2.3.2) becomes

$$u_j^{n+1} = s(u_{j+1}^n + u_{j-1}^n) + (1 - 2s)u_j^n. \quad (2.3.4)$$

We can separate variables in this difference equation, as we did in Section 1.2.5 for the heat equation, by making the ansatz $u_j^n = T_nX_j$. We obtain

$$\frac{T_{n+1}}{T_n} = 1 - 2s + s \frac{X_{j+1} - X_{j-1}}{X_j},$$

where the left hand side is independent of $j$, while the right hand side is independent of $n$. Since they are equal, both must be constant. Let us call this constant $\mu$. The equation $T_{n+1}/T_n = \mu$ is solved by

$$T_n = \mu^n T_0.$$ 

For the second equation,

$$1 - 2s + s \frac{X_{j+1} - X_{j-1}}{X_j} = \mu, \quad (2.3.5)$$

we can, inspired by Section 1.2.5, guess that a general solution $X_j$ might be a sum of a sine and a cosine. The cosine can be eliminated by using the boundary condition on the left, which leaves us with $X_j = \sin(j\theta)$. At the right boundary we must have $\sin(J\theta) = 0$ and therefore $J\theta = m\pi$, for some $m \in \mathbb{N}$, or equivalently $\theta = h\pi m$. Thus

$$X_j = \sin(jmh\pi).$$

We can compute $\mu$ by plugging this solution back into (2.3.5) and simplifying afterwards, which gives

$$1 - 2s(1 - \cos(mh\pi)) = \mu = \mu(m). \quad (2.3.6)$$
Thus a general solution to (2.3.4) is a linear combination of the form
\[ u_j^n = T_0 \sum_{m} A_m \mu(m)^n \sin(jmh\pi), \]  
(2.3.7)
where the numbers \( A_m \) would have to be determined using the initial condition.

According to (2.3.7), the growth in time of a solution is determined by the terms \( \mu(m)^n \). If \( |\mu(m)| > 1 \), for some \( m \), the solution can grow out of bounds eventually, in which case the finite difference scheme cannot be stable. Therefore \( |\mu(m)| \leq 1 \) for all \( m \in \mathbb{N} \) is necessary for stability.

Returning to (2.3.6) we observe that the cosine takes values between \(-1\) and \(1\). Thus \( 1 - \frac{4s}{k^2} \leq \mu(m)^n \leq 1 \). The lower bound \( 1 - \frac{4s}{k^2} \) should not be less than \(-1\) in order to have \( |\mu| \leq 1 \). This means \( s \leq 1/2 \).

\[ \begin{aligned}
\tau_j^n &= \frac{\hat{u}_{j+1}^n - \hat{u}_j^n}{k} - \frac{\hat{u}_{j+1}^n - 2\hat{u}_j^n + \hat{u}_{j-1}^n}{h^2} \\
\tau_j^n &= O(k) + O(h^2) \rightarrow 0, \quad \text{as} \quad h, k \rightarrow 0.
\end{aligned} \]  
(2.3.8)

Therefore the explicit scheme (2.3.2) is first order consistent. Moreover, from the second equality above it follows that we can obtain an upper bound for the maximal truncation error from upper bounds for \( u_{xxx} \) and \( u_{tt} \), if they exist. That is, if we assume \( u_{xxx}(x,t) \leq C \) and \( u_{tt}(x,t) \leq D \) for all \( x \) and \( t \), then
\[ \bar{\tau} := \max_{j,n} |\tau_j^n| \leq \frac{k}{2} \left( C + \frac{s}{6}D \right). \]  
(2.3.9)

2.3.3 Convergence

Next we would like to answer the following question: Denote the exact solution of the heat equation (2.3.1) by \( \hat{u} \) and let \( u_j^n \) be the solution of the explicit difference scheme with \( k/h^2 \leq 1/2 \). Does the global error \( e_j^n = u_j^n - \hat{u}(x_j, t^n) = u_j^n - \hat{u}_j^n \) tend to zero as we decrease the stepsizes \( h \) and \( k \)?

First, we consider the truncation error. As in Section 2.2.2 it is given by the residual of the difference scheme after plugging in the PDE solution. Thus
\[ \tau_j^n = \frac{\hat{u}_{j+1}^n - \hat{u}_j^n}{k} - \frac{\hat{u}_{j+1}^n - 2\hat{u}_j^n + \hat{u}_{j-1}^n}{h^2} \]
Recalling the Taylor expansions of the forward difference (2.1.3) and the centred second difference (2.1.6) we can write
\[ \begin{aligned}
\tau_j^n &= \hat{u}_t(x_j, t_n) + \frac{k}{2} \hat{u}_{tt}(x_j, t_n) + O(k^2) \\
&\quad - \hat{u}_{xx}(x_j, t_n) - \frac{h^2}{12} \hat{u}_{xxxx}(x_j, t_n) - O(h^4) \\
&= \frac{k}{2} \hat{u}_t(x_j, t_n) + O(k^2) - \frac{h^2}{12} \hat{u}_{xxxx}(x_j, t_n) - O(h^4) \\
&= O(k) + O(h^2) \\
&\rightarrow 0, \quad \text{as} \quad h, k \rightarrow 0.
\end{aligned} \]  
(2.3.8)

Therefore the explicit scheme (2.3.2) is first order consistent. Moreover, from the second equality above it follows that we can obtain an upper bound for the maximal truncation error from upper bounds for \( u_{xxx} \) and \( u_{tt} \), if they exist. That is, if we assume \( u_{xxx}(x,t) \leq C \) and \( u_{tt}(x,t) \leq D \) for all \( x \) and \( t \), then
\[ \bar{\tau} := \max_{j,n} |\tau_j^n| \leq \frac{k}{2} \left( C + \frac{s}{6}D \right). \]  
(2.3.9)
Theorem 2.6. Consider the heat equation (2.3.1) on a bounded time interval \([0,T]\). Assume \(k/h^2 \leq \frac{1}{2}\) and that (2.3.9) is satisfied. Then the explicit scheme (2.3.2) is convergent.

Proof. The global and truncation errors are related in a way similar to (2.2.7). We have

\[ e^{n+1}_j = s(e^n_{j+1} + e^n_{j-1}) + (1 - 2s)e^n_j - k\tau^n_j, \]

where \(s = k/h^2\). This equation follows from plugging \(\hat{u}\) into (2.3.4), which leaves a residual of \(k\tau^n_j\), and subtracting it from (2.3.4). Take absolute values on both sides and apply the triangle inequality to obtain

\[ |e^{n+1}_j| \leq s |e^n_{j+1}| + s |e^n_{j-1}| + (1 - 2s) |e^n_j| + k |\tau^n_j|. \]

Notice that the coefficients on the right are all positive and therefore absolute values can be omitted. Setting \(E^n = \max_j |e^n_j|\) and taking into account that \(s + s + 1 - 2s = 1\) gives

\[ E^{n+1} \leq E^n + k\bar{\tau}. \]

We apply this inequality recursively to obtain

\[ E^n \leq E^{n-1} + k\bar{\tau} \leq E^{n-2} + 2k\bar{\tau} \leq \cdots \leq E^0 + nk\bar{\tau} \]

Note that \(E^0 = 0\). Moreover, because of the boundedness of the time interval \([0,T]\), the index \(n\) runs from 0 to \(N\), where \(kN = T\). Therefore

\[ E^n \leq nk\bar{\tau} \leq Nk\bar{\tau} = T\bar{\tau} \leq T\frac{k}{2}\left(C + \frac{s}{6}D\right)^{k \to 0} 0. \]

\[ \square \]

2.3.4 The \(\theta\)-scheme

The explicit scheme (2.3.2) is called \textit{conditionally stable}, because it only is stable, if the two stepsizes are related in the right way. Unfortunately, the condition \(k/h^2 \leq 1/2\) can be quite severe. Suppose, for instance, that \(h = 10^{-2}\). Then \(k\) can be at most \(5 \cdot 10^{-5}\). This means that at least 20,000 time steps are required to compute the solution at time \(t = 1\). In this section we will discuss schemes that are \textit{unconditionally stable}, that is, stable without any restriction on the sizes of \(h\) or \(k\).

The main idea behind the \(\theta\)-scheme is to replace the right hand side of the explicit scheme by a weighted average — a convex combination actually — of the centred second differences at times \(t_n\) and \(t_{n+1}\). To be more precise, denote by \(D^+_{k,t}\) the forward difference operator with stepsize \(k\) in \(t\) direction, and by \(D^2_{h,x}\) the centred second difference with stepsize \(h\) in \(x\) direction.
Definition 2.6. For each $\theta \in [0, 1]$, the $\theta$-scheme for the one-dimensional heat equation is given by

$$D_t^\theta u^n_j = (1 - \theta)D_x^{2x}u^n_j + \theta D_x^{2x}u^n_{j+1}.$$  \hspace{1cm} (2.3.10)

If $\theta = 0$, we recover the explicit scheme. If $\theta > 0$, the scheme is called implicit, because we cannot explicitly express the solution at time $t_{n+1}$ in terms of the solution at time $t_n$. Instead, we have to solve a linear system of the form

$$Au^{n+1}_* = u^n,$$

where $u_* = (u_1^*, \ldots, u_{J-1}^*)^\top$.

Theorem 2.7. The $\theta$-scheme (2.3.10) is unconditionally stable for $1/2 \leq \theta \leq 1$.

Proof. Since this proof is similar to that of Theorem 2.5, we only give the main steps. Separation of variables gives a solution whose behaviour over time is determined by the growth factors $\mu(m)_n$, where

$$\mu(m) = \frac{1 - 2(1 - \theta)s(1 - \cos(mh\pi))}{1 + 2\theta s(1 - \cos(mh\pi))}.$$  \hspace{1cm} (2.3.11)

The stability requirement $|\mu(m)| \leq 1$ boils down to

$$s(1 - 2\theta)(1 - \cos(mh\pi)) \leq 1.$$

If $\theta \geq 1/2$, then this last inequality holds for sure, whatever the value of $s > 0$. Therefore, the $\theta$-scheme with $\theta \in [1/2, 1]$ is unconditionally stable.

The Crank-Nicolson Scheme

The choice $\theta = 1/2$ leads to the so-called Crank-Nicolson scheme. Note that in this case one takes the arithmetic mean of the centred differences at times $n$ and $n + 1$ on the right hand side:

$$\frac{u^{n+1}_j - u^n_j}{k} = \frac{1}{2} \left( \frac{u^n_{j+1} - 2u^n_j + u^n_{j-1}}{h^2} + \frac{u^{n+1}_{j+1} - 2u^{n+1}_j + u^{n+1}_{j-1}}{h^2} \right).$$

Setting $r = k/(2h^2)$ and moving all terms involving $n + 1$ to the left and all others to the right gives

$$-ru^{n+1}_{j+1} + (1 + 2r)u^{n+1}_j - ru^{n+1}_{j-1} = ru^n_{j+1} + (1 - 2r)u^n_j + ru^n_{j-1}.$$  \hspace{1cm} (2.3.11)

Recall that in our model problem we have homogeneous Dirichlet boundary conditions. Therefore $j$ runs from 1 to $J - 1$ and the terms $u^n_0$, $u^{n+1}_0$, $u^{n+1}_J$ vanish. Defining the $(J - 1) \times (J - 1)$ matrix

$$A_h = \frac{1}{h^2} \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix},$$  \hspace{1cm} (2.3.12)
we can rewrite (2.3.11) as
\[
\left( I - \frac{k}{2} A_h \right) u^{n+1} = \left( I + \frac{k}{2} A_h \right) u^n.
\]

(2.3.13)

Each time step requires solving a tridiagonal system. Recall that the number of flops required to solve a tridiagonal system is only linear in the number of unknowns. The Crank-Nicolson scheme is therefore not significantly slower per time step than the explicit scheme (2.3.2), where no linear system must be solved. Moreover, we can choose much larger time steps for the Crank-Nicolson scheme. Of course, the Crank-Nicolson scheme — or any \( \theta \)-scheme — can be applied to problems more general than (2.3.1). Below we briefly consider two generalizations.

**Inhomogeneous boundary condition.** In case of inhomogeneous Dirichlet boundary conditions

\[
\begin{align*}
u(0,t) &= \psi_0(t) \\
u(1,t) &= \psi_1(t)
\end{align*}
\]

the linear system (2.3.13) becomes
\[
\left( I - \frac{k}{2} A_h \right) u^{n+1} = \left( I + \frac{k}{2} A_h \right) u^n + r\psi^n,
\]

(2.3.14)

where \( \psi^n \) is the following vector
\[
\psi^n = \begin{bmatrix} \psi_0(t_n) + \psi_0(t_{n+1}) \\ 0 \\ \vdots \\ 0 \\ \psi_1(t_n) + \psi_1(t_{n+1}) \end{bmatrix}.
\]

**2D heat equation.** As another generalization, consider the two-dimensional heat equation on the unit square. Here, we first replace the centred second difference in the \( \theta \)-scheme (2.3.10) by, for instance, the five-point Laplacian (2.2.2):

\[
D^+_{k,t} u_{i,j}^n = (1 - \theta) \Delta_h u_{i,j}^n + \theta \Delta_h u_{i,j}^{n+1}.
\]

Note that the unknown \( u \) now has two subscript indices corresponding to the two spatial dimensions. Setting \( \theta = 2 \) and translating into matrix-vector form leads to essentially the same system as (2.3.14). Only \( A_h \) must be replaced by the matrix (2.2.5) arising from Poisson’s equation and the vector \( \psi^n \) must be adapted to the new boundary.

In presence of a nonconstant diffusion coefficient it is advisable to discretize as in (2.2.13).
2.3.5 The Lax Equivalence Theorem

We conclude Section 2.3 with a key result in the numerical analysis of differential equations. The Lax Equivalence Theorem — sometimes also called Lax-Richtmyer Theorem — is a mathematically precise version of the recurring theme that consistency and stability entail convergence, recall Section 1.4.

All difference schemes considered in Section 2.3 can be written in the following general form

\[ u^{n+1} = Bu^n + f^n, \tag{2.3.15} \]

for a square matrix \( B \) and a vector \( f^n \). Note that all the quantities (\( u, B \) and \( f \)) in (2.3.15) depend on the step sizes \( h \) and \( k \). For simplicity, however, we now assume that there is a fixed dependence of \( h \) on \( k \) as \( k \to 0 \).

The truncation error of method (2.3.15) is given by

\[ \tau^n = \hat{u}^{n+1} - B\hat{u}^n - f^n, \]

where \( \hat{u} \) corresponds to the exact solution of the PDE. It is consistent if \( \tau^n_j \to 0 \), and convergent if \( \|u - \hat{u}\| \to 0 \) as \( k \to 0 \). Regarding stability note that we have not yet given a precise definition for time-dependent problems.

**Definition 2.7.** We call (2.3.15) Lax-Richtmyer stable, if for every time \( T > 0 \) there is a constant \( C_T \) such that

\[ \|B^n\| \leq C_T \]

holds for all \( k > 0 \) and \( n \in \mathbb{N} \) for which \( nk \leq T \).

**Theorem 2.8 (Lax Equivalence Theorem).** For a consistent scheme of the form (2.3.15) Lax-Richtmyer stability is equivalent to convergence.

**Proof.** We do not go into details here and instead refer to [7, 8]. The main idea of the first part (i.e., a consistent and stable scheme is convergent) is analogous to the proof of Theorem 2.6. First, one derives a difference relation for the global error, thus relating it to the truncation error. Then, one exploits stability to obtain an upper bound which goes to zero as \( k \to 0 \). \( \square \)

**Remark 2.7.** The Lax Equivalence Theorem is a useful result, because it allows us to establish a property of a numerical method which is essential but hard to check (convergence) by means of to other properties which are easier to check (consistency and stability).

Moreover, note that the theorem not only guarantees that “consistency + stability ⇒ convergence,” but also that “consistency + convergence ⇒ stability.” If the second implication were not true, then there could be convergent but unstable methods. This possibility, however, is eliminated by the Lax Equivalence Theorem.
2.4 Hyperbolic Problems

The one-dimensional wave equation

\[ u_{tt} = c^2 u_{xx}, \quad (2.4.1) \]

with propagation speed \( c > 0 \), can be reduced to a system of two homogeneous linear transport equations. Setting \( v = u_t \) we can write the wave equation as

\[
\begin{align*}
    u_t + cv_x &= 0, \\
    v_t + cu_x &= 0.
\end{align*}
\]

(Differentiate the first equation with respect to \( t \) and subtract from it \( c \) times the \( x \)-derivative of the second equation.) Now, with \( p = (u+v)/2 \) and \( q = (u-v)/2 \), this system transforms to

\[
\begin{align*}
    p_t + cp_x &= 0, \\
    q_t - cq_x &= 0, \quad (2.4.2)
\end{align*}
\]

which consists of two homogeneous transport equations. This shows that we can solve (2.4.2) and then compute \( u = p + q \) to obtain a solution of (2.4.1).

Therefore, in this section we will restrict our attention to devising finite difference schemes for the problem

\[
\begin{align*}
    u_t + au_x &= 0, \quad x \in \mathbb{R}, \ t > 0 \\
    u &= \phi, \quad t = 0,
\end{align*}
\]

where \( a \in \mathbb{R} \) and \( \phi = \phi(x) \) is a given function. Recall Section 1.2.2, where we have shown that this problem has the unique solution \( u(x,t) = \phi(x-\alpha t) \). While, strictly speaking, numerical treatment of this problem is not really necessary, it will guide us to a better understanding of more general hyperbolic equations.

2.4.1 The Lax-Friedrichs Scheme

In the considerations below we cover the unbounded domain \( \mathbb{R} \times [0, \infty) \) of problem (2.4.3) with an infinite grid

\[ \Omega_{h,k} = \{(x_j, t_n) = (jh, kn) : j \in \mathbb{Z}, n \in \mathbb{N}\}. \]

As a first attempt we could try to mimic the explicit scheme for the heat equation: forward differences in time, centred differences in space, which gives

\[ u_j^{n+1} = u_j^n - \frac{ak}{2h} (u_{j+1}^n - u_{j-1}^n). \quad (2.4.4) \]

This scheme, however, turns out to be a bad choice due to stability issues: A separation ansatz will give a solution of the form \( u_j^n = \lambda^n e^{ijkh} \), where \( i \) denotes
the imaginary unit. It follows that \( |u^n_j| = |\lambda|^n \), but it can be shown that \( |\lambda| > 1 \) in general.

The difference scheme above can be improved upon by replacing \( u^n_j \) on the right hand side with the average of \( u^n_{j+1} \) and \( u^n_{j-1} \). This gives the so-called Lax-Friedrichs scheme

\[
u^{n+1}_j = \frac{1}{2} \left( u^n_{j+1} + u^n_{j-1} \right) - \frac{ak}{2h} \left( u^n_{j+1} - u^n_{j-1} \right).
\]

(2.4.5)

Setting \( \gamma = \frac{ak}{h} \) we can visualize the Lax-Friedrichs scheme with the following stencil

\[
\begin{bmatrix}
1 + \gamma & \ast & 0 \\
\frac{1-\gamma}{2} & 1 & \frac{1-\gamma}{2}
\end{bmatrix},
\]

where \( \ast \) stands for \( u^{n+1}_j \), while the second row contains the coefficients of \( u^n_{j-1} \), \( u^n_j \) and \( u^n_{j+1} \).

**Theorem 2.9.** The Lax-Friedrichs scheme (2.4.5) is conditionally stable:

\[
\text{If } \left| \frac{ak}{h} \right| \leq 1, \text{ then } \|u^{n+1}\|_1 \leq \|u^n\|_1.
\]

**Proof.** The Lax-Friedrichs scheme can be written as

\[
u^{n+1}_j = \frac{1}{2} \left( 1 - \frac{ak}{h} \right) u^n_{j+1} + \frac{1}{2} \left( 1 + \frac{ak}{h} \right) u^n_{j-1}.
\]

Taking absolute values on both sides and using the triangle inequality we get

\[
|u^{n+1}_j| \leq \frac{1}{2} \left( 1 - \frac{ak}{h} \right) |u^n_{j+1}| + \frac{1}{2} \left( 1 + \frac{ak}{h} \right) |u^n_{j-1}|.
\]

Note that because of the restriction on the stepsizes, the two numbers \( 1 \pm ak/h \) are nonnegative. Summing over \( j \) gives

\[
\|u^{n+1}\|_1 \leq \frac{1}{2} \left( 1 - \frac{ak}{h} \right) \|u^n\|_1 + \frac{1}{2} \left( 1 + \frac{ak}{h} \right) \|u^n\|_1 = \|u^n\|_1.
\]

\[\square\]

**Remark 2.8.** Why does the seemingly arbitrary replacement of \( u^n_j \) with the average \( \frac{1}{2}(u^n_{j+1} + u^n_{j-1}) \) make a difference? First, the proof above simply fails without this modification. This, admittedly, is not a very insightful explanation. Therefore, let us rewrite

\[
\frac{1}{2}(u^n_{j+1} + u^n_{j-1}) = u^n_j + \frac{1}{2}(u^n_{j+1} - 2u^n_j + u^n_{j-1}) = u^n_j + \frac{h^2}{2}D^2h^{-2}u^n_j.
\]

Plugging the right hand side back into the Lax-Friedrichs scheme yields

\[
D^+_ku^n_j = \frac{h^2}{2k}D^2h^{-2}u^n_j - aD^0h^{-1}u^n_j.
\]
Setting \( \varepsilon = \frac{h^2}{2\pi} \) we can interpret this equation as a discretization of the PDE

\[
u_t + au_x = \varepsilon u_{xx},
\]

which is the transport equation with an additional diffusive term. This so-called artifical diffusion has a stabilizing effect on the numerical solution, but can also cause a loss of accuracy. Artificial diffusion actually is a common feature of many schemes for the transport equation (including all the schemes presented below), but in the case of Lax-Friedrichs it is particularly apparent.

### 2.4.2 The Upwind Scheme

When using one-sided differences to approximate \( u_x \), stability can only be guaranteed if the direction of the difference operator is chosen in accordance with the sign of \( a \). This leads to the so-called upwind scheme

\[
u_j^{n+1} = u_j^n - \begin{cases} \frac{ak}{h} (u_j^n - u_{j-1}^n), & a > 0, \\ \frac{ak}{h} (u_{j+1}^n - u_j^n), & a < 0. \end{cases} \tag{2.4.6}
\]

The stencils for these two cases take the form

\[
\begin{bmatrix} \gamma & * \\ 1-\gamma & 0 \end{bmatrix}, \quad a > 0, \\
\begin{bmatrix} 0 & * \\ 1+\gamma & -\gamma \end{bmatrix}, \quad a < 0.
\]

**Theorem 2.10.** The upwind scheme (2.4.6) is conditionally stable:

If \( \left| \frac{ak}{h} \right| \leq 1 \), then \( \|u_{n+1}\|_1 \leq \|u^n\|_1 \).

**Proof.** Let \( a > 0 \). Then the upwind scheme can be written as

\[
u_j^{n+1} = \left(1 - \frac{ak}{h}\right) u_j^n + \frac{ak}{h} u_{j-1}^n.
\]

Noting that the two coefficients \(1 - ak/h\) and \(ak/h\) are nonnegative, we can proceed as in the proof of Theorem 2.9.

\[
\|u_{n+1}\|_1 \leq \left(1 - \frac{ak}{h}\right) \|u^n\|_1 + \frac{ak}{h} \|u^n\|_1 = \|u^n\|_1
\]

The case \( a < 0 \) works analogously. \( \Box \)

**Remark 2.9.**

1. If the “wrong” one-sided difference is chosen, the estimate above fails and the scheme cannot, in general, be convergent. The reason for this phenomenon is explored in Section 2.4.3.
2. The critical quantity for stability of both the Lax-Friedrichs and the up-wind scheme is the Courant number \( \gamma = \frac{ak}{h} \). The condition \( |\gamma| \leq 1 \) can be read as an upper bound on the time step \( k \leq h/|a| \). Notice that this bound is much less restrictive than the one required for stability of the explicit scheme for the heat equation (Theorem 2.5). This brings us to one of the fundamental differences between parabolic and hyperbolic problems: While the efficient solution of parabolic problems requires implicit schemes, hyperbolic problems are typically solved with explicit methods.

The correct choice of one-sided difference as well as the condition \( |ak/h| \leq 1 \) are two manifestations of one and the same principle that must be obeyed when solving time-dependent problems. This principle is discussed below.

2.4.3 The Courant-Friedrichs-Lewy Condition

Let \( u \) be the solution of a time-dependent PDE and consider its value \( u(\bar{x}, \bar{t}) \) at some fixed point \( (\bar{x}, \bar{t}) \in \Omega \times (0, \infty) \). We can ask the following question: Which values \( \phi(x) \) of the initial condition actually influence \( u(\bar{x}, \bar{t}) \)? This leads us to the domain of dependence \( D(\bar{x}, \bar{t}) \), which is the set of all those points in \( \Omega \) where a change in \( \phi \) results in a change of \( u(\bar{x}, \bar{t}) \).

**Example 2.4.** Consider the linear transport problem (2.4.3). The solution is constant along the characteristic lines, recall (1.2.6). Therefore, \( u(\bar{x}, \bar{t}) = \phi(\bar{x} - a\bar{t}) \) is entirely determined by the initial condition at the one point \( \bar{x} - a\bar{t} \in \mathbb{R} \). Thus

\[
D(\bar{x}, \bar{t}) = \{\bar{x} - a\bar{t}\}.
\]

This is a very special situation. In general we can expect \( u(\bar{x}, \bar{t}) \) to depend on the values of \( \phi \) on a larger set of points, such as an interval.

**Example 2.5.** For the 1D wave equation \( u_{tt} = c^2 u_{xx} \) with propagation speed \( c \) the domain of dependence of \( (\bar{x}, \bar{t}) \) is the interval

\[
D(\bar{x}, \bar{t}) = [\bar{x} - c\bar{t}, \bar{x} + c\bar{t}].
\]

Recall d’Alembert’s formula in (1.2.9), but note that \( c = 1 \) there! From each starting point \( (x, 0) \) the initial “information” \( \phi(x) \), \( \psi(x) \) propagates with velocity \( c \) both to the left and the right. This situation can be nicely visualized by drawing the characteristic triangle in the \( x-t \)-plane with vertices \( (\bar{x} - c\bar{t}, 0) \), \( (\bar{x} + c\bar{t}, 0) \) and \( (\bar{x}, \bar{t}) \).

**Example 2.6.** The heat equation \( u_t = u_{xx} \) is another extreme case, as it has infinite propagation speed: \( u(\bar{x}, \bar{t}) \) is affected by the values of \( \phi \) on the entire domain \( \Omega \), even if \( \Omega = \mathbb{R} \) and no matter how close \( \bar{t} \) is to 0. Therefore

\[
D(\bar{x}, \bar{t}) = \Omega.
\]

For finite difference schemes one defines the numerical domain of dependence \( D_{h,k}(x_j, t_n) \) in an analogous fashion, that is, as the set of all grid points \( (x_j, t_0) \) for which \( \phi(x_j) \) contributes to the value of the solution \( u_j^n \).
Example 2.7. Consider an explicit scheme. From its definition we can immediately infer which points \((x_j, t_{n-1})\) contribute to \(u_j^n\). We proceed in the same fashion from \(t_{n-1}\) to \(t_{n-2}\) and so on, until we ultimately reach \(t_0\) and thus have found the numerical domain of dependence. The grid points that we have thus visited typically form a triangular array. Note that the shape of the array does not change as \(h, k \to 0\), as long as the ratio \(k/h\) is kept fixed.

We can now state the Courant-Friedrichs-Lewy condition.

Theorem 2.11 (CFL Condition). A finite difference scheme can only be convergent if its numerical domain of dependence contains the domain of dependence of the approximated PDE, that is,

\[
D_{h,k}(x_j, t_n) \supset D(x_j, t_n),
\]

at least asymptotically for \(h, k \to 0\).

Proof. Suppose that the condition is violated. Then there is a point \(\xi \in D(x_j, t_n)\) with \(\xi \notin D_{h,k}(x_j, t_n)\). If we now change the initial condition in a small region around \(\xi\), then the exact solution of the PDE changes, but the numerical solution is unaffected. Therefore the scheme cannot be convergent for arbitrary initial conditions. \(\square\)

Remark 2.10.

1. Marking out the dependencies for the Lax-Friedrichs scheme as mentioned in Example 2.7 leads to an array of points with the shape of an isosceles triangle. Let the peak of this triangle be \((x_j, t_n)\). The CFL condition dictates that the characteristic line that runs through the point \((x_j, t_n)\) must also run through the base of the triangular array. This is guaranteed if the characteristic’s slope is steeper than the triangle’s sides. Mathematically, this means that \(1/|a| \geq k/h\) must hold. But this is just the stability requirement of Theorem 2.9!

2. Now consider the upwind scheme. The same procedure as above leads to a right-angled triangle, which basically marks out one half of the triangle obtained from the Lax-Friedrichs scheme. In order to satisfy the CFL condition, we not only have to have \(1/|a| \geq k/h\), but also the characteristic line must lean to the correct side if it is to pass through the triangle’s base. This additional restriction is taken care of by the dependence on \(\text{sgn}(a)\) of the upwind scheme.

3. Keep in mind that the CFL condition is only necessary for convergence, but not sufficient. This means: Every convergent scheme must satisfy the CFL condition, but there might be schemes which satisfy it but are still not convergent. An example is the scheme (2.4.4). The usefulness of the CFL condition lies in the fact that it is very easy to check and therefore allows one to reject poor schemes with little effort.
2.4.4 Higher Order Methods

The Lax-Friedrichs scheme as well as the upwind scheme are first order consistent. This can be shown as in (2.3.8). We conclude Section 2.4 with a brief discussion of second order methods.

One way to obtain a second order consistent method is to use a centred difference in time
\[ u_j^{n+1} = u_j^{n-1} - \frac{ak}{h} (u_{j+1}^{n} - u_{j-1}^{n}) . \]
This is the \textit{leapfrog scheme}. Its stencil reads
\[
\begin{bmatrix}
\gamma & * & -\gamma \\
1 & - & 1
\end{bmatrix}.
\]

In contrast to all other methods considered in Sections 2.3 and 2.4 it is a \textit{three-level scheme}, since it involves the solution at three time levels \( n-1, n \) and \( n+1 \). Note that, when implementing a three-level scheme, one has to store the solution at the last two time steps, thus increasing memory requirements.

Two-level higher order schemes can be obtained with the following approach. Let \( \hat{u} \) be the solution to \( u_t + au_x = 0 \). Using Taylor series expansion in time we can write
\[ \hat{u}(x, t + k) = \hat{u}(x, t) + k \hat{u}_t(x, t) + \frac{k^2}{2} \hat{u}_{tt}(x, t) + O(k^3). \]
Replacing \( \hat{u}_t \) with \( -a \hat{u}_x \) and \( \hat{u}_{tt} \) with \( a^2 \hat{u}_{xx} \) we obtain
\[ \hat{u}(x, t + k) = \hat{u}(x, t) - ak \hat{u}_x(x, t) + \frac{a^2 k^2}{2} \hat{u}_{xx}(x, t) + O(k^3). \quad (2.4.7) \]
Using centred differences to approximate the first and second \( x \)-derivatives while dropping the higher order terms gives the the \textit{Lax-Wendroff scheme}
\[ u_j^{n+1} = u_j^n - ak D_{h}^{0,x} u_j^n + \frac{a^2 k^2}{2} D_{h}^{2,x} u_j^n \]
with stencil
\[
\begin{bmatrix}
\frac{\gamma(\gamma+1)}{2} & * & \frac{\gamma(\gamma-1)}{2} \\
1 & - & 1
\end{bmatrix}.
\]

The Lax-Wendroff scheme is second order consistent, because the truncation error takes the form
\[ \tau_j^n = D_{h}^{+t} \hat{u}_j^n + a D_{h}^{0,x} \hat{u}_j^n - \frac{a^2 k^2}{2} D_{h}^{2,x} \hat{u}_j^n, \]
and the fact that the centred differences are second order accurate combined with equation (2.4.7) shows that \( \tau_j^n = O(k^2) + O(h^2) \).
The Finite Difference Method

Table 2.2: Comparison of difference schemes for the linear transport equation.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Order</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lax-Friedrichs</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Upwind</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Leapfrog</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Lax-Wendroff</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

The derivation above also shows how to obtain two-level finite differences schemes for the transport equation of even higher order: First, do a Taylor expansion in time. Then, replace $t$-derivatives with $x$-derivatives up to the desired order. Finally, approximate all $x$-derivatives by suitably high order accurate differences.
Chapter 3
The Finite Element Method

The finite element method was originally invented for problems in elasticity and structural mechanics, but, due to its high flexibility, has since been applied to virtually every field of science and engineering that can be modelled by partial differential equations. One of its main advantages over the finite difference method is the ability to handle non-rectilinear domains with ease. On the downside it should be mentioned that both theoretical framework and implementation of the finite element is somewhat less straightforward compared to finite difference methods. At its core the finite element method consists in piecewise polynomial approximation. A slightly more elaborate summary could read like this:

1. Partition the domain Ω of the PDE into elements, that is, into simple polygonal or polyhedral subsets.
2. On this partition construct piecewise polynomial basis functions $\phi_1, \ldots, \phi_N$ with small support.\(^1\)
3. Represent the approximate solution in this basis $u_h = \sum_i q_i \phi_i$.
4. Use the weak formulation of the PDE to solve for the coefficients $q_1, \ldots, q_N$.

3.1 A One-Dimensional Example

In order to illustrate the main ideas of the finite element method, we first consider an ordinary differential equation. Let $f : (0, 1) \to \mathbb{R}$ be a given function.\(^2\)

Problem 1 (Strong formulation). Find a function $u : [0, 1] \to \mathbb{R}$ such that

$$\begin{cases} -u''(x) = f(x), & x \in (0, 1), \\ u(x) = 0, & x \in \{0, 1\}. \end{cases}$$

\(^1\)The support of a function is the set $\{x : f(x) \neq 0\}$.
\(^2\)The term “strong formulation” will become clear later.
The Finite Element Method

The first step of the finite element method is to translate this differential equation into its weak formulation. To this end we multiply the ODE with a function \( v \) satisfying \( v(0) = v(1) = 0 \) and integrate both sides over the domain

\[-\int_{0}^{1} u''(x)v(x) \, dx = \int_{0}^{1} f(x)v(x) \, dx.\]

Next we integrate by parts on the left hand side

\[-\int_{0}^{1} u''(x)v(x) \, dx = -u'(x)v(x)\bigg|_{0}^{1} + \int_{0}^{1} u'(x)v'(x) \, dx\]

\[= \int_{0}^{1} u'(x)v'(x) \, dx,\]

where we have used the fact that \( v \) vanishes on the endpoints of the interval. Therefore, if \( u \) solves the boundary value problem above, then it also satisfies

\[\int_{0}^{1} u'(x)v'(x) \, dx = \int_{0}^{1} f(x)v(x) \, dx\]

for all \( v \) which vanish at the endpoints and for which this integral equation makes sense at all, i.e. the integrals are finite. Let us denote the set of all those functions by \( V \). Then the weak formulation of Problem 1 reads as follows.

**Problem 2** (Weak formulation). Find a \( u \in V \) such that

\[\int_{0}^{1} u'(x)v'(x) \, dx = \int_{0}^{1} f(x)v(x) \, dx\]

for all \( v \in V \).

Next, we discretize using the Galerkin method. We choose a finite-dimensional subspace \( V_h \) of \( V \) and solve the weak formulation on \( V_h \) instead of \( V \).

**Problem 3** (Galerkin equations). Find a \( u_h \in V_h \) such that

\[\int_{0}^{1} u_h'(x)v_h'(x) \, dx = \int_{0}^{1} f(x)v_h(x) \, dx \quad (3.1.1)\]

for all \( v_h \in V_h \).

Suppose the dimension of \( V_h \) is \( N \). Then there must be a basis \( \{\phi_1, \ldots, \phi_N\} \) of \( V_h \) meaning that every function in \( V_h \) can be expressed as a linear combination of the \( \phi_j \). A basis helps us make the Galerkin equations more manageable in two ways. First, it is enough to verify equation (3.1.1) for all \( \phi_j \) instead of all \( v_h \in V_h \):

\[\int_{0}^{1} u_h'(x)\phi_j'(x) \, dx = \int_{0}^{1} f(x)\phi_j(x) \, dx, \quad j = 1, \ldots, N.\]
Second, since the solution $u_h$ also lies in $V_h$, we can make the ansatz

$$u_h = \sum_{i=1}^{N} q_i \phi_i$$  \hspace{1cm} (3.1.2)$$

and plug it into the Galerkin equations. We obtain

$$\sum_{i=1}^{N} q_i \int_{0}^{1} \phi'_i(x) \phi'_j(x) \, dx = \int_{0}^{1} f(x) \phi_j(x) \, dx, \quad j = 1, \ldots, N;$$

where we have exploited the linearity of the integral to change the order of summation and integration. Note that this is a system of linear equations in the unknowns $q_1, \ldots, q_N$! Defining

$$a_{ij} := \int_{0}^{1} \phi'_i(x) \phi'_j(x) \, dx,$$

$$b_j := \int_{0}^{1} f(x) \phi_j(x) \, dx,$$

we can write the Galerkin equations in the following way.

**Problem 4** (Galerkin equations, algebraic). *Find a $q \in \mathbb{R}^N$ such that $Aq = b$.*

The matrix $A$ is often referred to as *stiffness matrix.*

Finally, let us assemble this linear system for a specific approximation space $V_h$ and basis $\{\phi_1, \ldots, \phi_N\}$. Consider the following set of $N + 2$ equidistant points in $[0, 1]$

$$\Delta = \{ x_i = ih : i = 0, \ldots, N + 1 \},$$

where $h = 1/(N + 1)$. Denote by $S^1(\Delta)$ the space of linear splines on the partition $\Delta$, that is, continuous functions on $[0, 1]$ which are linear (= first-order) polynomials on each subinterval $[x_i, x_{i+1}]$. We now set $V_h = S^1_h(\Delta)$, which is the subset of those linear splines that vanish at the boundary. Linear splines admit a very convenient basis, the *hat functions.* For $S^1_0(\Delta)$ this basis is given by

$$\phi_i(x) = \begin{cases} 
\frac{x - x_{i-1}}{h}, & x \in [x_{i-1}, x_i] \\
\frac{x_{i+1} - x}{h}, & x \in [x_i, x_{i+1}] \\
0, & \text{otherwise},
\end{cases}$$

for $i = 1, \ldots, N$. The diagonal entries of the system matrix $A = (a_{ij})$ are given by

$$a_{i,i} = \int_{0}^{1} \phi'_i(x)^2 \, dx = \int_{x_{i-1}}^{x_i} \frac{1}{h^2} \, dx + \int_{x_i}^{x_{i+1}} \frac{1}{h^2} \, dx = \frac{2}{h}. $$
In the subdiagonals next to the main diagonal we have

\[ a_{i,i+1} = \int_0^1 \phi_i'(x)\phi_{i+1}'(x) \, dx = \int_{x_i}^{x_{i+1}} -\frac{1}{h} \, dx = -\frac{1}{h}. \]

Note that all other entries of \( A \) are zero, since the integrand \( \phi_i'(x)\phi_j'(x) \) vanishes for all \( x \), if \( |i - j| > 1 \). Therefore,

\[
A = \frac{1}{h} \begin{bmatrix}
2 & -1 & -1 & \cdots & -1 \\
-1 & 2 & -1 & \cdots & -1 \\
& \ddots & \ddots & \ddots & \ddots \\
& & -1 & 2 & -1 \\
& & & -1 & 2
\end{bmatrix}.
\]

This matrix is invertible. (In fact, up to a factor of \(-1/h\) it is equal to the matrix (2.3.12) which was derived by discretizing \( u'' \) using centred differences.)

Thus, after computing the vector \( b \) we can solve for the coefficient vector \( q \) and obtain the finite element solution via (3.1.2).

### 3.2 Principles of the Finite Element Method

In this section we discuss the main ingredients of the finite element method from a more general point of view. We will, however, frequently refer to the one-dimensional example considered above. We start by restating the four problems of Section 3.1 using an abstract, more general notation. In doing so we try to highlight the fact the following considerations not only apply to the 1D example above but to a wide range of PDEs.

**Problem 5** (Strong formulation). Find a \( u : \Omega \to \mathbb{R} \) such that

\[ Lu(x) = f(x) \]

for all \( x \in \Omega \).

**Problem 6** (Weak formulation). Find a \( u \in V \) such that

\[ a(u,v) = \ell(v) \]

for all \( v \in V \).

**Problem 7** (Galerkin equations). Find a \( u_h \in V_h \) such that

\[ a(u_h,v_h) = \ell(v_h) \]

for all \( v_h \in V_h \).

**Problem 8** (Galerkin equations, algebraic). Find a \( q \in \mathbb{R}^N \) such that

\[
\sum_{i=1}^N q_i a(\phi_i, \phi_j) = \ell(\phi_j), \quad j = 1, \ldots, N.
\]
Here, $V$ is a space of functions which is determined by the problem at hand and which can be thought of as the largest space of functions for which the weak formulation makes sense. $V_h$ is a finite-dimensional subspace of $V$. The right hand side $\ell : V \rightarrow \mathbb{R}$ is a linear form, and $a : V \times V \rightarrow \mathbb{R}$ is a bilinear form meaning it is linear in both arguments. ("Form" is simply a name for a function from a vector space into the set of scalars.)

### 3.2.1 The Weak Formulation

#### Variational Problems

Frequently, differential equations originate from variational problems. One reason is that physical systems typically arrange themselves in ways that minimize a certain quantity. Think, for instance, of the principle of least action or the principle of minimum energy.

Mathematically, we could model each state of a physical system as a function $v \in V$ to which we assign an "energy" $J(v) \in \mathbb{R}$. The function $J : V \rightarrow \mathbb{R}$ is sometimes referred to as an energy functional. The goal is then to find a state $u \in V$ of minimal energy, that is,

$$J(u) = \min_{v \in V} J(v).$$

Partial differential equations only appear in such a problem when we try to solve it by what is essentially Fermat’s rule: If a function attains an extremum at a certain point, then its derivative must vanish there. This leads to

$$J'(u) = 0.$$

This so-called Euler-Lagrange equation of $J$ is in many cases a differential equation. For example, the ODE of Problem 1 is the Euler-Lagrange equation of the energy functional

$$J(v) = \frac{1}{2} \int_0^1 v'(x)^2 \, dx - \int_0^1 f(x)v(x) \, dx.$$  

More generally, a differential equation with weak formulation $a(u, v) = \ell(v)$ can, under certain conditions, be shown to be the Euler-Lagrange equation of the functional

$$J(v) = \frac{1}{2} a(v, v) - \ell(v).$$  

The weak formulation of a problem basically says that, at a minimum, the variation of $J$ in every direction should be zero. It can be thought of as occupying

---

3 Strictly speaking, $V$ is required to be a Hilbert space for the considerations of Section 3.2 to apply. The space $H^1_0(0, 1)$ mentioned below is an example of such a space, as are many other spaces which naturally turn up in the weak formulation of differential equations. We refrain from delving into mathematical details at this point and refer the interested reader to the excellent treatise [1] instead.

4 "Variational problem" is just the mathematician’s term for “minimization problem.”
The Finite Element Method

the middle ground between minimization problem and Euler-Lagrange equation. Since the finite element method is close in spirit to the minimization problem, we have to revert some of the steps that the mathematician performed in order to derive the Euler-Lagrange equation. Those reverse steps usually are:

1. Multiply the differential equation with a function \( v \) that vanishes on the boundary of \( \Omega \).
2. Integrate the resulting equation over \( \Omega \).
3. Integrate by parts.

Classical and Weak Solutions

The weak and strong formulations of a problem are not equivalent in general. As the names suggest, a solution of the strong formulation, a \textit{classical solution}, is always a solution of the weak formulation too, i.e. a \textit{weak solution}. The reverse statement, however, is not true in many cases. Observe that Problem 1 involves a second derivative and therefore implicitly requires every classical solution to be at least twice differentiable. Problem 2, on the other hand, involves only first derivatives. In addition the derivative only appears inside an integral, and therefore for a weak solution it can be admissible to not even have a first derivative at isolated points. The importance of the weak formulation lies in fact that there are relevant problems, where no classical but only weak solutions exist.

In mathematical terms, the set of candidate solutions for Problem 1 is \( C_0^2[0,1] \), that is, the set of all functions defined on \([0,1]\), which have a continuous second derivative. The subscript 0 indicates that the functions are also required to vanish at the boundary. The solution space \( V \) for Problem 2, however, is the much larger Sobolev space \( H_0^1(0,1) \) consisting of all functions for which the two integrals

\[
\int_0^1 v(x)^2 \, dx \quad \text{and} \quad \int_0^1 v'(x)^2 \, dx
\]

are finite. Sobolev spaces are the natural solution spaces for weak problems and therefore play an important role in finite element theory.

The weak formulation is also useful from a practical point of view. By decreasing the differentiability requirements it gives us more freedom in selecting the approximation space \( V_h \). In Section 3.1 representing \( u_h \) as a linear spline worked only because we considered the weak formulation. (Linear splines do not have a second derivative. In fact, they do not even have a first derivative.)

3.2.2 The Galerkin Method

At first glance it seems like not much has been achieved by Galerkin’s method. We simply change spaces from \( V \) to \( V_h \). It does have, however, several important implications, which we discuss below.
Galerkin Orthogonality

First, we observe that the solution $u$ of the weak form also satisfies the Galerkin equations: Since $V_h \subset V$ we have

$$a(u, v_h) = \ell(v_h)$$  \hspace{1cm} (3.2.1)

for all $v_h \in V_h$. This property, namely that the exact solution of a problem also solves the approximate equation, is called strong consistency. Contrast this to the finite difference method, where the exact solution solves the approximate equation only in the limit $h \to 0$, i.e. the truncation error is nonzero in general but tends to zero.

Next, subtracting (3.2.1) from the Galerkin equations $a(u_h, v_h) = \ell(v_h)$ and using the linearity of $a$ gives

$$a(u_h - u, v_h) = 0$$  \hspace{1cm} (3.2.2)

for all $v_h \in V_h$. This property is referred to as Galerkin orthogonality. The reason for this name is that $a$ can be interpreted as an inner product on $V$ and then, according to this inner product, equation (3.2.2) says that the approximation error $u_h - u$ is orthogonal to the approximation space $V_h$.

Céa’s Lemma

Galerkin orthogonality allows us to obtain an upper bound for the magnitude of the approximation error $\|u_h - u\|$. To derive this estimate we require two more definitions.

**Definition 3.1.** The bilinear function $a(\cdot, \cdot)$ is called

- **bounded**, if there is a constant $C > 0$ such that
  $$a(u, v) \leq C\|u\|\|v\|$$
  holds for all $u, v \in V$, and

- **coercive**, if there is a constant $c > 0$ such that
  $$a(v, v) \geq c\|v\|^2$$
  holds for all $v \in V$.

**Example 3.1.** The left hand side of Problem 2, that is,

$$a(u, v) = \int_0^1 u'(x)v'(x) \, dx$$

is both bounded and coercive with constants $C = c = 1$ for the norm

$$\|v\| = \left( \int_0^1 v'(x)^2 \, dx \right)^{\frac{1}{2}}$$

on $H_0^1(0, 1)$. 
Now suppose that $a$ is a bounded and coercive bilinear form. Then we have
\[ c\|u_h - u\|^2 \leq a(u_h - u, u_h - u) = a(u_h - u, u_h) - a(u_h - u, u). \]

Galerkin orthogonality tells us that the first term in the resulting difference equals zero. Therefore we can replace it by another expression which is also zero.
\[ = a(u_h - u, v_h) - a(u_h - u, u) \]
\[ = a(u_h - u, v_h - u) \leq C\|u_h - u\|\|v_h - u\| \]

In total we have shown
\[ c\|u_h - u\|^2 \leq C\|u_h - u\|\|v_h - u\| \]
for every $v_h \in V$. Dividing both sides by $c\|u_h - u\|$ gives
\[ \|u_h - u\| \leq \frac{C}{c} \|v_h - u\|. \]

Thus we have proved the following fundamental error estimate.

**Theorem 3.1** (Céa’s Lemma). If the bilinear form $a(\cdot, \cdot)$ is bounded and coercive, then the error between the solutions of Problems 6 and 7 admits the following upper bound
\[ \|u_h - u\| \leq \frac{C}{c} \min_{v_h \in V_h} \|v_h - u\|. \]

**Remark 3.1.**

1. Céa’s Lemma relates the approximation error of Galerkin’s method to the “distance” between $u$ and $V_h$, which can be interpreted as the approximation quality of the space $V_h$ for the given problem.

2. The definitions of boundedness and coercivity imply that the fraction $C/c$ cannot be less than 1. If $C/c = 1$, which can be shown in certain cases (see Example 3.1), Céa’s Lemma actually guarantees that
\[ \|u_h - u\| = \min_{v_h \in V_h} \|v_h - u\|. \]

That is, $u_h$ is the best possible approximation to $u$ in $V_h$.

---

5These conditions by the way ensure that Problems 6 and 7 have a unique solution. This result is known as Lax-Milgram Theorem.
3.2.3 Piecewise Polynomials

In the finite difference method one approximately solves the differential equation on grid points. The computed solution is therefore always a grid function \( u_h : \Omega_h \to \mathbb{R} \). A distinguishing feature of the finite element method is that one always represents \( u_h \) as a linear combination of basis functions \( \phi_i : \Omega \to \mathbb{R} \)

\[
  u_h = \sum_{i=1}^{n} q_i \phi_i.
\]

Consequently, the Galerkin equations reduce to a linear system of (algebraic) equations

\[
  \sum_{i=1}^{N} q_i a(\phi_i, \phi_j) = \ell(\phi_j), \quad j = 1, \ldots, N.
\]

In this Section we address the question of how to choose the approximation space \( V_h \) and the basis functions \( \phi_i \). Our guiding principles are accuracy and efficiency of the numerical method.

One option would be to find, among all subspaces \( V_h \subset V \) of dimension \( N \), one which minimizes the “distance” \( \min_{v \in V_h} \| u - v_h \| \). (Recall Céa’s Lemma.) This would allow one to keep the dimension \( N \) of the resulting linear system relatively small while maintaining a good approximation quality. Essentially this line of reasoning leads to so-called spectral methods. They are beyond the scope of this lecture, but a good exposition can be found in [10], for instance.

Another important consideration is computational efficiency. Note that we not only have to solve the system \( Aq = b \) (recall Problem 4), but we also have to compute all the entries of \( A \) and \( b \) in the first place. Therefore we should aim for a system matrix that is as sparse as possible, that is, most entries

\[
  a_{ij} = a(\phi_i, \phi_j)
\]

should vanish. In Section 3.1 this was achieved, because hat functions have a small support. (Recall that the support of a function is that subset of its domain where it is nonzero.) Since the support of the \( i \)-th hat function is \( (x_{i-1}, x_{i+1}) \), most products \( \phi_i \phi_j \) turned out to be zero and so were the entries of the system matrix. In addition, even the nonzero entries of the matrix were very easy to compute. Thus, from the viewpoint of computational efficiency we want to select basis functions which have small support and are easy to integrate. A defining feature of the finite element method is to use piecewise polynomials, which satisfy these two requirements perfectly.

In order to be able to construct a piecewise polynomial basis we first have to partition the domain into “pieces”. In the one-dimensional setting this is rather straightforward. In higher dimensions, however, there are suddenly many possibilities. The finite element method typically uses subdivisions of \( \Omega \) into simple polygonal or polyhedral subsets, also called elements. Common types of elements are triangles or rectangles (in 2D) and tetrahedra or hexahedra (in 3D). How piecewise polynomial basis functions can be constructed on such partitions is discussed in the following section.
3.3 A Two-Dimensional Example

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain. We consider the boundary value problem

$$
\begin{cases}
- \text{div}(\sigma \nabla u) + cu = f, & \text{in } \Omega, \\
u = 0, & \text{on } \partial \Omega,
\end{cases}
$$

where the coefficient function $\sigma : \Omega \rightarrow \mathbb{R}$ is assumed to satisfy $\sigma(x) \geq \sigma_0 > 0$ for some constant $\sigma_0$. This condition implies that the PDE is elliptic and that the bilinear form of the associated weak formulation is coercive.

**Weak Formulation.** In order to derive the weak formulation we first multiply with a function $v$ that vanishes on $\partial \Omega$ and integrate.

$$
- \int_{\Omega} \text{div}(\sigma \nabla u) v + \int_{\Omega} cu v = \int_{\Omega} f v.
$$

Next we use Green’s first identity (see the Appendix) to rewrite the first integral on the left and arrive at

$$
\int_{\Omega} \sigma \nabla u \cdot \nabla v + \int_{\Omega} cu v = \int_{\Omega} f v, \quad (3.3.2)
$$

The weak formulation of problem (3.3.1) therefore consists in finding a $u \in V$ such that (3.3.2) holds for all $v \in V$. The function space $V$ is given by $H^1_0(\Omega)$ consisting of all functions which vanish on the boundary and for which the two integrals

$$
\int_{\Omega} v^2 \quad \text{and} \quad \int_{\Omega} |\nabla v|^2
$$

are finite.

**Piecewise Linear Functions on Triangles.** Next we define an approximation space $V_h$. We choose the most common approach of *piecewise linear functions on triangles*. The first step is to triangulate the domain. Recall that we assumed $\Omega$ to be a polygonal domain. Therefore it can be subdivided into a finite number of triangles $T_1, \ldots, T_M$. We call the set of triangles

$$
\mathcal{T} = \{T_1, \ldots, T_M\}
$$

a *regular triangulation* of $\Omega$, if each pair of adjacent triangles $T_i, T_j$ either touch at only one vertex or along an entire shared edge. In other words, the edges
must not form T-shaped crossings. Denoting by $l_i$ the length of the longest side of $T_i$ we set

$$h = \max_{1 \leq i \leq M} l_i \quad (3.3.3)$$

This number serves as a measure for the fineness of the triangulation and plays a role similar to the stepsizes $h, k$ in finite difference methods.

**Definition 3.2.** The space of piecewise linear functions on a regular triangulation $\mathcal{T}$ is given by

$$S^1(\mathcal{T}) = \left\{ s \in C^0(\Omega) : s\big|_{T_i} \in P_1 \text{ for all } i = 1, \ldots, M \right\}.$$

Here, $C^0(\Omega)$ is the set of all continuous functions on $\Omega$ and $P_1$ consists of all first-order polynomials in two variables. Thus, every $s \in S^1(\mathcal{T})$ is a continuous function on $\Omega$ which, when restricted to any triangle in $\mathcal{T}$, reduces to a linear polynomial in $x$ and $y$. That is, for every triangle $T_i$ there must be a triple of real numbers $(a_i, b_i, c_i)$ such that

$$s(x, y) = a_i x + b_i y + c_i \quad (3.3.4)$$

for all points $(x, y) \in T_i$. Since polynomials are always continuous, the continuity requirement is only relevant across triangles. ($s$ must not make sudden jumps from one triangle to the next.)

Finally, we take the homogeneous Dirichlet boundary condition into account by setting

$$V_h = S^1_0(\mathcal{T}) = \left\{ s \in S^1(\mathcal{T}) : s = 0 \text{ on } \partial \Omega \right\}.$$

**Two-Dimensional Hat Functions.** Having settled for an approximation space we want to construct a basis for this space consisting of very locally supported functions. We start by considering the set

$$\mathcal{P} = \{p_1, \ldots, p_N\}$$

of all vertices of the triangulation $\mathcal{T}$. That is, $p_i \in \mathcal{P}$, if it is a vertex of at least one triangle $T_j \in \mathcal{T}$. Points that are vertices of more than one triangle appear only once in the set $\mathcal{P}$. The set of vertices which do not lie on the boundary of $\Omega$ is denoted by $\mathcal{P}_0$

$$\mathcal{P}_0 = \{p_i \in \mathcal{P} : p_i \notin \partial \Omega\}.$$

For convenience we enumerate the vertices in $\mathcal{P}$ in such a way that

$$\mathcal{P}_0 = \{p_1, \ldots, p_K\}$$

for some integer $K < N$. The importance of $\mathcal{P}$ and $\mathcal{P}_0$ lies in the following statement.

**Lemma 3.1.**
• Every \( s \in S^1(T) \) is uniquely determined by its values at the vertices in \( P \).

• Every \( s \in S^0_0(T) \) is uniquely determined by its values at the vertices in \( P_0 \).

Proof. First, observe that a linear polynomial in two variables is uniquely determined by its values at three points (unless the points lie on one line.) This should be clear considering that (3.3.4) involves three coefficients. It follows that a linear polynomial on a triangle is uniquely determined by its values on the triangle’s vertices. Since the triangulation is regular (!), the claim follows. \( \square \)

Lemma 3.1 guarantees that the following definition is meaningful.

**Definition 3.3.** The two-dimensional hat functions \( \phi_1, \ldots, \phi_N \) associated to \( T \) are defined by

\[
\phi_i(p_j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}
\]

The graph of hat function \( \phi_i \) forms a pyramid with peak above vertex \( p_i \) and base given by the union of triangles adjacent to \( p_i \). As in the one-dimensional setting they span the space of piecewise linear functions.

**Lemma 3.2.**

- The hat functions \( \{ \phi_1, \ldots, \phi_N \} \) form a basis of \( S^1(T) \). In particular, every \( s \in S^1(T) \) can be written as

\[
s = \sum_{i=1}^{N} s(p_i) \phi_i.
\]

- The hat functions \( \{ \phi_1, \ldots, \phi_K \} \) form a basis of \( S^0_0(T) \). In particular, every \( s \in S^0_0(T) \) can be written as

\[
s = \sum_{i=1}^{K} s(p_i) \phi_i.
\]

Since \( V_h = S^0_0(T) \), we are mainly interested in the second statement of Lemma 3.2. It justifies the representation

\[
u_h = \sum_{i=1}^{K} q_i \phi_i,
\]

which turns the problem of determining \( u_h \) into a linear system of equations.
Assembling the Linear System. Next, we demonstrate how to assemble the linear system $Aq = b$ for the important special case $\sigma \equiv 1$ and $c \equiv 0$. We have to calculate

$$a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \nabla \phi_i \cdot \nabla \phi_j,$$

$$b_i = \ell(\phi_i) = \int_\Omega f \phi_i,$$

for $i, j = 1, \ldots, K$.

Regarding the stiffness matrix $A$ we first observe that it is sparse. More precisely, $a_{ij} \neq 0$ only if $p_i$ and $p_j$ are vertices of the same triangle. $A$ is also invertible, since it is an SPD matrix (symmetric and positive definite). Symmetry follows immediately from (3.3.5), while positive definiteness is implied by the coercivity of the bilinear form $a(\cdot, \cdot)$:

$$q^T A q = \sum_{i,j=1}^K q_i q_j a_{ij} = \sum_{i,j=1}^K q_i q_j a(\phi_i, \phi_j) = a\left(\sum_{i=1}^K q_i \phi_i, \sum_{j=1}^K q_j \phi_j\right) \geq c \left\| \sum_{i=1}^K q_i \phi_i \right\|.$$

The norm on the right is nonnegative and equal to zero only if its argument is the zero function, which is equivalent to $q$ being the zero vector.

A systematic way of evaluating the integrals consists in first observing that they can be decomposed into a sum

$$\int_\Omega g = \sum_k \int_{T_k} g$$

and then transforming each of the resulting integrals into one over a reference triangle.

**Lemma 3.3.** Let $T \in T$. Denote its vertices by $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$ and $p_3 = (x_3, y_3)$. Further, let $T_{\text{ref}}$ be the reference triangle with vertices $(0,0)$, $(1,0)$ and $(0,1)$.

Then, the coordinate transformation

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^2, \quad \Phi(\xi, \eta) = p_1 + \xi (p_2 - p_1) + \eta (p_3 - p_1),$$

maps $T_{\text{ref}}$ to $T$. An integral over $T$ transforms according to

$$\int_T g(x, y) \, dx \, dy = |\det J_{\Phi}| \int_{T_{\text{ref}}} g(\Phi(\xi, \eta)) \, d\xi \, d\eta,$$

where the Jacobian determinant is given by

$$\det J_{\Phi} = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1).$$

**Proof.** Equality of the two integrals follows from multidimensional integration by substitution (see the Appendix). \qed
The coordinate transform \( \Phi \) can be expressed in matrix-vector form as

\[
\Phi(\xi, \eta) = \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}
\]

We return to the computation of (3.3.5). Applying the Lemma above gives

\[
\int_{T} \nabla \phi_i \cdot \nabla \phi_j \, dx \, dy = |\det J_{\Phi}| \int_{T_{ref}} (\nabla \phi_i \circ \Phi) \cdot (\nabla \phi_j \circ \Phi) \, d\xi \, d\eta
\]

where we have transformed the gradients in \( x \) and \( y \) to gradients in \( \xi \) and \( \eta \). Recall equation (1.5.4) and that \( A^{-\top} \) denotes \((A^{-1})^\top\). Next, we exploit the fact that hat functions are transformed into hat functions by \( \Phi \) and therefore their gradients are constant on \( T_{ref} \). We obtain

\[
= \frac{|\det J_{\Phi}|}{2} (J_{\Phi}^{-\top} \nabla \phi_i \circ \Phi) \cdot (J_{\Phi}^{-\top} \nabla \phi_j \circ \Phi).
\]

In order to evaluate this expression, it only remains to calculate the gradients. The transformed hat functions associated to the three vertices \( p_1, p_2, p_3 \) of \( T \) are given by

\[
\phi_i(\Phi(\xi, \eta)) = \begin{cases} 1 - \xi - \eta, & i = 1, \\ \xi, & i = 2, \\ \eta, & i = 3, \end{cases}
\]

and therefore

\[
\nabla \phi_i(\Phi(\xi, \eta)) = \begin{cases} (-1, -1)^\top, & i = 1, \\ (1, 0)^\top, & i = 2, \\ (0, 1)^\top, & i = 3. \end{cases}
\]

If we collect the integrals \( \int_{T} \nabla \phi_i \cdot \nabla \phi_j \) for \( i, j = 1, \ldots, 3 \) in a three-by-three matrix \( S_T \) we can write the result as

\[
S_T = \frac{|\det J_{\Phi}|}{2} G J_{\Phi}^{-1} (G J_{\Phi}^{-1})^\top
\]

where \( G \in \mathbb{R}^{3 \times 2} \) contains the gradients \( \nabla \phi_i \circ \Phi \)

\[
G = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.
\]
The matrix \( S_T \) is sometimes referred to as \textit{element stiffness matrix}. In order to complete the assembly of the matrix \( A \) one has to repeat the above procedure and compute \( S_T \) for all other elements \( T \in \mathcal{T} \).

Regarding the right hand side (3.3.6) it is common to replace \( f \) by its piecewise linear interpolant

\[
\sum_{j=1}^{N} f(p_j) \phi_j
\]

which leads to

\[
b_i = \int_{\Omega} f \phi_i = \int_{\Omega} \sum_{j=1}^{N} f(p_j) \phi_j \phi_i = \sum_{j=1}^{N} f(p_j) \int_{\Omega} \phi_j \phi_i, \quad i = 1, \ldots, K,
\]

or in matrix-vector form

\[
b = M \tilde{f},
\]

where \( m_{ij} = \int_{\Omega} \phi_i \phi_j \) and \( \tilde{f} = (f(p_1), \ldots, f(p_N))^T \). The nonzero entries of \( M \) can be found using quadrature rules.

### 3.4 Convergence

Suppose we solve the weak formulation of a PDE on a sequence of approximation spaces

\[
V_{h_1}, V_{h_2}, \ldots \subset V.
\]

Under what conditions do the approximate solutions \( u_{h_i} \) converge to the true solution \( u^* \) of the PDE? We give an answer for two-dimensional problems.

**Theorem 3.2.** Let \( V \) be the solution space of the weak formulation of a PDE of order \( 2\nu \). Let \( p \in \mathbb{N} \) be the order of the piecewise polynomials in \( V \). Suppose the approximation spaces \( V_{h_i} \subset V \) are such that

\[
\dim(V_{h_i}) \to \infty, \quad \text{as} \quad i \to \infty.
\]

In addition assume that

\[
h_i \to 0, \quad \text{as} \quad i \to \infty,
\]

where \( h_i \) is as defined in (3.3.3). Then, there is a \( C > 0 \) such that

\[
\|u_{h_i} - u^*\| \leq Ch_i^{p+1-\nu}.
\]

**Remark 3.2.** In the two-dimensional example of Section 3.3 we have \( p = \nu = 1 \) and therefore

\[
\|u_{h_i} - u^*\| = O(h_i).
\]
Appendix

Big O Notation

Big O notation is a way of describing the limiting behaviour of functions as their argument approaches a certain point. Let $a \in \mathbb{R} \cup \{\pm \infty\}$ and $f, g : \mathbb{R} \to \mathbb{R}$ be two functions. Then we write

$$f(x) = O(g(x)) \quad \text{as} \quad x \to a$$

if there is a constant $C$ such that

$$|f(x)| < C|g(x)|$$

for all $x$ sufficiently close to $a$. So, essentially $f$ can be controlled by $g$ in the vicinity of $a$.

In these notes we will use the big O notation for functions $f, g$ depending on a parameter $h$ that approaches $a = 0$. We often omit the specification “as $h \to 0$” and simply write

$$f(h) = O(g(h)).$$

Most of the time $g(h)$ will be some monomial $h^p$, $p > 0$, in which case $f(h) = O(h^p)$ basically means that $f$ approaches zero about as fast as $h^p$ or faster. In other words, the decay rate of $f(h)$ is $h^p$ or higher. Note that, if a function is a sum of several terms, then its decay rate is dominated by the slowest term, for instance,

$$3h^2 + 7h^3 + h^4 = O(h^2), \quad \text{as} \quad h \to 0.$$

Vector Calculus

Below $\Omega$ denotes a bounded region in $\mathbb{R}^d$ with outward pointing unit normal $n : \partial \Omega \to \mathbb{R}^d$.

**Theorem** (Divergence Theorem). Let $F : \Omega \to \mathbb{R}^d$ be a vector field with continuous partial derivatives. Then

$$\int_{\Omega} \text{div} F = \int_{\partial \Omega} F \cdot n.$$
The divergence theorem can be interpreted as a generalization of the fundamental theorem of calculus. Intuitively, it says that the net flux out of a region (right hand side) is equal to the sum of all sinks $(\text{div} F(x) < 0)$ and sources $(\text{div} F(x) > 0)$ inside that region. An immediate consequence is the following higher-dimensional analogue of integration by parts.

**Theorem (Green’s First Identity).** Let $u : \Omega \to \mathbb{R}$ be a function and $V : \Omega \to \mathbb{R}^d$ a vector field, both with continuous partial derivatives. Then

$$
\int_{\Omega} u \text{div} V = \int_{\partial \Omega} u V \cdot n - \int_{\Omega} \nabla u \cdot V.
$$

**Proof.** We apply the divergence theorem with $F = uV$. It remains to verify that $\text{div}(uV) = \nabla u \cdot V + u \text{div} V$:

$$
\text{div}(uV) = \sum_{i=1}^{d} \frac{\partial}{\partial x_i} (uV_i) = \sum_{i=1}^{d} \left( \frac{\partial u}{\partial x_i} V_i + u \frac{\partial V_i}{\partial x_i} \right) = \sum_{i=1}^{d} \frac{\partial u}{\partial x_i} V_i + u \sum_{i=1}^{d} \frac{\partial V_i}{\partial x_i} = \nabla u \cdot V + u \text{div} V.
$$

Integration by substitution also has a higher-dimensional analogue:

**Theorem.** Let $\Phi : \Omega \to \mathbb{R}^d$ be one-to-one and differentiable and denote its Jacobian matrix by $J_\Phi$. Then

$$
\int_{\Phi(\Omega)} f = \int_{\Omega} (f \circ \Phi) |\det J_\Phi|
$$

for every integrable function $f$. 

Bibliography


