Classical Numerical Methods in Scientific Computing

Jos van Kan, Guus Segal, Fred Vermolen

TU Delft OPEN

Classical Numerical Methods in Scientific Computing

Classical Numerical Methods in Scientific Computing

J. van Kan
A. Segal
F. Vermolen

Delft Institute of Applied Mathematics
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

First Print by Delft Academic Press, 2019
Previously published under the title “Numerical methods for partial differential equations” with Delft Academic Press, 2019
ISBN: 97890-6562-4383

ISBN 978-94-6366-731-9 (paperback)
ISBN 978-94-6366-732-6 (Ebook)
DOI: https://doi.org/10.59490/t.2023.007

Preface

You just opened a book about numerically approximating the solutions to partial differential equations that occur in technological and scientific contexts. Since it is important to judge the quality of obtained numerical approximations of solutions to partial differential equations, one needs to have an impression of the mathematical properties of these solutions. For this reason, the book starts with the treatment of elementary mathematical properties of the solutions to partial equations. These properties entail existence and uniqueness, maximum principles, conservation of mass, conservation of momentum or convergence to steady state. Besides these properties, also some attention is paid to the derivation and motivation of partial differential equations using principles from physics. The connection to physics is crucially important for the analysis of consistency of the models, as well as for the analysis of the quality and fidelity of numerical approximations to solutions.

This book has been written for an audience consisting of engineers and scientists, who are dealing with 'real-world' problems. Most of the treatment is devoted to the actual implementation of the classical discretization methods, such as finite differences, finite volumes and finite elements. Some attention is paid to error analysis of finite difference and finite volume methods. The interested reader is provided with some basic functional analytic theory (like Riesz, Lax-Milgram, Cea's lemmas and theorems) that is needed for understanding existence and uniqueness of the (Galerkin) variational problem and finite element solution, as well as convergence and (apriori) error analysis of finite element solutions, though some important details such as the principles by Aubin-Nietsche and Bramble-Hilbert are omitted. Some error estimating upper bounds for finite element approximations have been listed. Further topics involve time-dependent partial differential equations and an introduction to linear and nonlinear solvers.

We hope that you will enjoy reading this book !

Diepenbeek, August 2023

Jos van Kan
Guus Segal
Fred Vermolen

Preface

This is a book about numerically solving partial differential equations occurring in technical and physical contexts and we (the authors) have set ourselves a more ambitious target than to just talk about the numerics. Our aim is to show the place of numerical solutions in the general modeling process and this must inevitably lead to considerations about modeling itself. Partial differential equations usually are a consequence of applying first principles to a technical or physical problem at hand. That means, that most of the time the physics also have to be taken into account especially for validation of the numerical solution obtained.

This book in other words is especially aimed at engineers and scientists who have 'real world' problems and it will concern itself less with pesky mathematical detail. For the interested reader though, we have included sections on mathematical theory to provide the necessary mathematical background.

This book is an abridged but improved version of our book [15]. The scope corresponds to Chapters 1-4, Section 9.7 and Chapters 10 and 11 from [15]. The material covers the FDM and FVM, but excludes the FEM, and is suitable for a semester course. The improvements will also be implemented in a future edition of the unabridged version [15] of this book.

Delft, August 2019

Jos van Kan
Guus Segal
Fred Vermolen
Hans Kraaijevanger

Contents

Chapter 1

Review of some basic mathematical concepts

1.1 Preliminaries

In this chapter we take a bird's eye view of the contents of the book. Furthermore we establish a physical interpretation of certain mathematical notions, operators and theorems. As a first application we formulate a general conservation law, since conservation laws are the backbone of physical modeling. Finally we treat some mathematical theorems that will be used in the remainder of this book.

1.2 Global contents of the book

First, in Chapter 2, we take a look at second order partial differential equations and their relation with various physical problems. We distinguish between stationary (elliptic) problems and evolutionary (parabolic and hyperbolic) problems.

In Chapters 3 and 4 we look at numerical methods for elliptic equations. Chapter 3 deals with finite difference methods (FDM), of respectable age but still very much in use, while Chapter 4 is concerned with finite volume methods (FVM), a typical engineers option, constructed for conservation laws. In this special version of the book we do not discuss finite element methods (FEM), which have gained popularity over the last decades. These methods are discussed in the unabridged version [15] of the book, however.

Application of the FDM or FVM generally leaves us with a large set of algebraic equations. In Chapter 5 we focus on the difficulties that arise when these equations are nonlinear.

In Chapters 6 and 7 we look at numerical methods for evolutionary problems. Chapter 6 deals with the heat equation (parabolic case), whereas Chapter 7 deals with the wave equation (hyperbolic case).

1.3 Building blocks for mathematical modeling

Several mathematical concepts used in modeling are directly derived from a physical context. We shall consider a few of those and see how they can be used to formulate a fundamental mathematical model: conservation.

1.3.1 Gradient of a scalar

Given a scalar function, u, of two variables, differentiable with respect to both variables, then the gradient is defined as

$$\text{grad}\, u = \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix}. \tag{1.3.1}$$

Instead of the notation grad u also ∇u (pronounce: nabla u) is used. To get to the core of what a gradient really is, think of temperature. If you have a temperature difference between two points, then you get a flow of heat between those points that only will stop when the temperature difference has been annihilated. If the difference is bigger, the flow will be larger. If the points are closer together the flow will be larger.

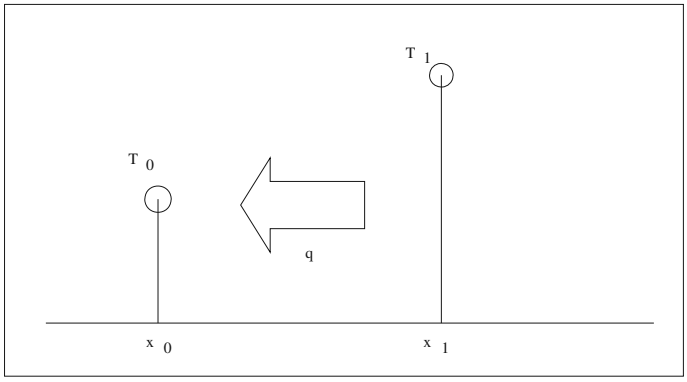

Figure 1.1: One-dimensional heat flow.

The simplest one-dimensional model to reflect this behaviour is the following linear model, illustrated in Figure 1.1. Let q be the generated flow, and assume it is directly proportional to the temperature difference ΔT and inversely proportional to the distance Δx.

Then we obtain the formula

$$q = -\lambda \frac{\Delta T}{\Delta x}, \tag{1.3.2}$$

where λ is a material constant, the *heat conduction* coefficient. The minus sign reflects the facts that

1. heat flows from high to low temperatures;
2. physicists hate negative constants.

In a sufficiently smooth temperature field $T(x)$ we may take limits and obtain a flow that is derived from (driven by) the temperature:

$$q = -\lambda \frac{dT}{dx}. \tag{1.3.3}$$

How is this in more than one dimension? Suppose we have a two-dimensional temperature field $T(x, y)$ which we can represent nicely by considering the contour lines which for temperature are called *isotherms*, lines that connect points of equal temperature (see Figure 1.2).

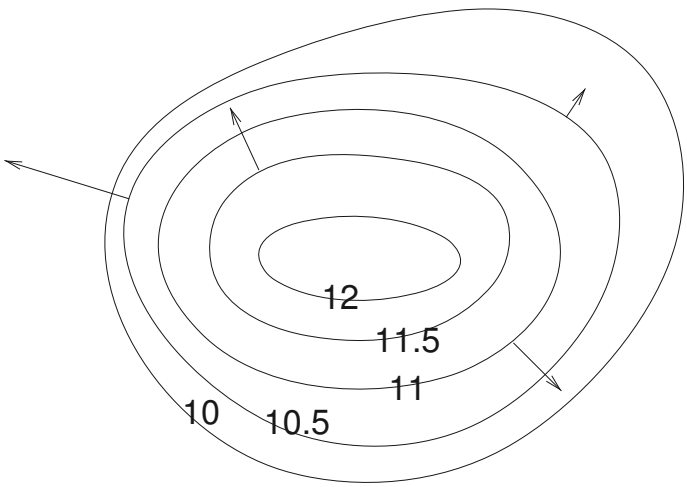

Figure 1.2: Isotherms.

Since there cannot be heat flow between points of equal temperature, the heat flow must be orthogonal to the contour lines at every point. Two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ are orthogonal if their inner product $(\boldsymbol{v}, \boldsymbol{w})$ vanishes. In other words: let $x(s), y(s)$ be a parameterization of a contour line and let $\begin{pmatrix} q_1 \\ q_2 \end{pmatrix}$ be the components of the heat flow field. We then have:

$$q_1 \frac{dx}{ds} + q_2 \frac{dy}{ds} = 0, \tag{1.3.4}$$

at every point $x(s), y(s)$ of the isotherm, for all isotherms. Let us substitute the parameterization of an isotherm into the temperature field: $T(x(s), y(s))$. Doing this makes T a function of s only, *which is constant* because we are on an isotherm. In other words, along an isotherm:

$$\frac{dT}{ds} = \frac{\partial T}{\partial x}\frac{dx}{ds} + \frac{\partial T}{\partial y}\frac{dy}{ds} = 0. \tag{1.3.5}$$

Comparing Equations (1.3.4) and (1.3.5) we see that these can only be satisfied if

$$\boldsymbol{q} = -\lambda \operatorname{grad} T. \tag{1.3.6}$$

For three dimensions you can tell basically the same story that also ends in Equation (1.3.6). This is known as *Fourier's law* and it is at the core of the theory of heat conduction.

Exercise 1.3.1 *(Fick's Law) In diffusion the flow of matter, $\boldsymbol{q}$, is driven by differences in concentration c. Express $\boldsymbol{q}$ in c.* □

Scalar fields like T and c whose gradients drive a flow field, $\boldsymbol{q}$, are called *potentials*.

1.3.2 Directional derivative

In the previous paragraph we saw how the temperature, T, changes along a curve $x(s), y(s)$. The actual value of dT/ds depends on the parameterization. A natural parameterization is the *arc length* of the curve. Note, that in that case

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 = 1.$$

This forms the basis of the following definition:

Definition 1.3.1 *Let $\boldsymbol{n}$ be a* unit *vector, then the directional derivative of T in the direction of $\boldsymbol{n}$ is given by*

$$\frac{\partial T}{\partial n} = \frac{\partial T}{\partial x}n_1 + \frac{\partial T}{\partial y}n_2 = (\operatorname{grad} T, \boldsymbol{n}) = (\boldsymbol{n} \cdot \nabla)T.$$

Exercise 1.3.2 *Compute the directional derivative of $z = x^2 + y^3$ in $(1,1)$ in the direction $(1,-1)$. (Answer: $-\frac{1}{2}\sqrt{2}$).* □

Exercise 1.3.3 *For what value of $\boldsymbol{n}$ is the directional derivative precisely $\frac{\partial T}{\partial x}$?* □

1.3.3 Divergence of a vector field

The mathematical definition of divergence is equally uninspiring. Given a continuously differentiable vector field, $\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, the divergence of $\boldsymbol{v}$ is defined by:

$$\operatorname{div} \boldsymbol{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y}. \tag{1.3.7}$$

For $\mathbb{R}^3$ you have the obvious generalization and there is also a nabla notation: $\operatorname{div} \boldsymbol{v} = \nabla \cdot \boldsymbol{v}$. You will appreciate the correspondence of a genuine inner product of two vectors and the inner product of the "nabla vector" and a vector field. Take care, however. In a genuine inner product you can change the order of the vectors, in the divergence you cannot.

What is the physical meaning of divergence? You could think of a vector field as a river: at any place in the river the water has a certain velocity with direction and magnitude. Now consider a fixed rectangular volume in the river (Figure 1.3).

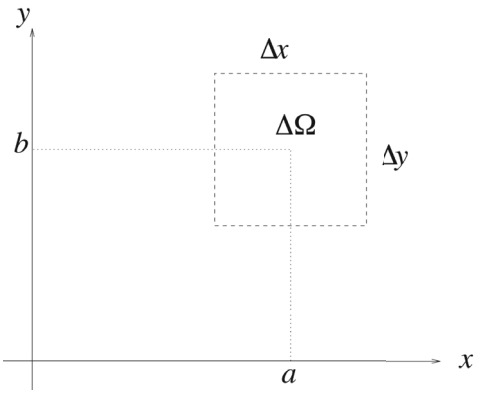

Figure 1.3: Rectangular volume in river.

Water is flowing in through the left and bottom wall and flowing out through the right and top wall. How much is flowing *in* through the left wall? If you think about it, you will notice that the y-component of the velocity gives no contribution to the inflow, because that is parallel to the left wall. So the inflow through the left wall is equal to $v_{1L}\Delta y$, the outflow through the right wall $v_{1R}\Delta y$. By the same reasoning the inflow through the bottom equals $v_{2B}\Delta x$, the outflow through the top equals $v_{2T}\Delta x$. What's left behind? If the net outflow is larger than the net inflow we are losing matter in the volume, if on the other hand the net inflow is larger we're gaining. The net outflow out of

control volume $\Delta\Omega$ in Figure 1.3 is given by

$$\begin{aligned}\Delta\phi(a,b) &= v_1(a+\frac{\Delta x}{2},b)\Delta y - v_1(a-\frac{\Delta x}{2},b)\Delta y \\ &\qquad + v_2(a,b+\frac{\Delta y}{2})\Delta x - v_2(a,b-\frac{\Delta y}{2})\Delta x \\ &= \Delta x\Delta y\frac{v_1(a+\frac{\Delta x}{2},b)-v_1(a-\frac{\Delta x}{2},b)}{\Delta x} \\ &\qquad + \Delta x\Delta y\frac{v_2(a,b+\frac{\Delta y}{2})-v_2(a,b-\frac{\Delta y}{2})}{\Delta y} \\ &= \Delta x\Delta y\left(\frac{\partial v_1}{\partial x}(\xi,b)+\frac{\partial v_2}{\partial y}(a,\eta)\right), \end{aligned} \tag{1.3.8}$$

for a $\xi \in (a-\frac{\Delta x}{2}, a+\frac{\Delta x}{2})$, $\eta \in (b-\frac{\Delta y}{2}, b+\frac{\Delta y}{2})$ from the Mean Value Theorem and continuity of the partial derivatives. This implies

$$\lim_{(\Delta x,\Delta y)\to(0,0)} \frac{\Delta\phi(a,b)}{\Delta x\Delta y} = \operatorname{div} \boldsymbol{v}(a,b). \tag{1.3.9}$$

From this formula, we see that $\operatorname{div} \boldsymbol{v}(a,b)$ is the outflow density (outflow per unit area) at point (a,b). Integration of the outflow density over an entire volume gives the total outflow. Since the total outflow can also be computed from evaluation of the flux over its boundary, we obtain a very important relation between the integral of the divergence of a vector-field over the volume and the integral of the flux over its boundary. This relation is formulated in terms of the divergence theorem, which we shall state in the next subsection.

Definition 1.3.2 *A vector field $\boldsymbol{v}$ that satisfies* $\operatorname{div} \boldsymbol{v} = 0$ *is called divergence-free or solenoidal.*

Exercise 1.3.4 *Explain that for an* incompressible *flow field $\boldsymbol{u}$ we always have* $\operatorname{div} \boldsymbol{u} = 0$. □

Exercise 1.3.5 *Derive in the same way as above that divergence is an outflow density in* $\mathbb{R}^3$. □

1.3.4 Gauss' divergence theorem

In the previous section we informally derived the divergence theorem, which was initially proposed by Gauss. In words: the outflow density integrated over an arbitrary volume gives the total outflow out of this volume. But this is mathematics, so we have to be more precise.

Throughout this book, we will use the more precise concepts 'domain' and 'region' instead of 'volume'. Both are formally defined as a nonempty, open and connected set, and are usually denoted by the Greek letter Ω. A domain

(or region) can be bounded or unbounded. Its boundary is denoted by $\partial\Omega$ or Γ, and its closure by $\overline{\Omega} = \Omega \cup \partial\Omega$. We will often tacitly assume that the boundary $\Gamma = \partial\Omega$ is piecewise smooth, so that at all boundary points (except a finite number of them), the outward normal unit vector is uniquely defined.

Theorem 1.3.1 (Gauss' divergence theorem)
Let Ω be a bounded domain in $\mathbb{R}^2$ ($\mathbb{R}^3$) with piecewise smooth boundary Γ. Let $\boldsymbol{n}$ be the outward normal and $\boldsymbol{v}$ a continuously differentiable vector field. Then

$$\int_{\Omega} \operatorname{div} \boldsymbol{v}\, d\Omega = \int_{\Gamma} \boldsymbol{v} \cdot \boldsymbol{n}\, d\Gamma. \tag{1.3.10}$$

Remarks

1. The expression $\boldsymbol{v} \cdot \boldsymbol{n}$ is the outward normal component of the vector-field, $\boldsymbol{v}$, with respect to the boundary. If this quantity is positive you have outflow, otherwise inflow.
2. Any good book on multivariate analysis will have a proper proof of Gauss' theorem. (See for instance [2] or [12]). A good insight will be obtained however, by subdividing the region Ω in small rectangles and using (1.3.8). Note in particular, that the common side (plane in $\mathbb{R}^3$) of two neighboring volumes cancel: what flows out of one flows into the other. The proof is finalized by taking a limit $\Delta x, \Delta y \to 0$ in the Riemann sum.

Exercise 1.3.6 *Let C be a closed contour in the x-y-plane and $\boldsymbol{q}$ a solenoidal vector field. Show that $\int_C \boldsymbol{q} \cdot \boldsymbol{n}\, d\Gamma = 0$.* □

The divergence theorem has many important implications and these implications are used frequently in various numerical methods, such as the finite volume method and the finite element method. First, one can use the componentwise product rule for differentiation to arrive at the following theorem:

Theorem 1.3.2 *For a continuously differentiable scalar field, c, and vector field, $\boldsymbol{u}$, we have*

$$\operatorname{div}(c\boldsymbol{u}) = \operatorname{grad} c \cdot \boldsymbol{u} + c \operatorname{div} \boldsymbol{u}. \tag{1.3.11}$$

Exercise 1.3.7 *Prove Theorem 1.3.2.*

As a result of this assertion, one can prove the following theorem.

Theorem 1.3.3 (Green's theorem)
For sufficiently smooth c, $\boldsymbol{u}$, we have

$$\int_{\Omega} c \operatorname{div} \boldsymbol{u}\, d\Omega = -\int_{\Omega} (\operatorname{grad} c) \cdot \boldsymbol{u}\, d\Omega + \int_{\Gamma} c\boldsymbol{u} \cdot \boldsymbol{n}\, d\Gamma. \tag{1.3.12}$$

Exercise 1.3.8 *Prove Theorem 1.3.3.*

By the use of Theorem 1.3.3, the following assertion can be demonstrated:

Theorem 1.3.4 *Partial integration in 2 D*
For sufficiently smooth scalar functions ϕ and ψ, we have;

$$\int_\Omega \phi \frac{\partial \psi}{\partial x}\, d\Omega = -\int_\Omega \frac{\partial \phi}{\partial x}\psi\, d\Omega + \oint_\Gamma \phi\psi n_1 d\Gamma, \tag{1.3.13}$$

and

$$\int_\Omega \phi \frac{\partial \psi}{\partial y}\, d\Omega = -\int_\Omega \frac{\partial \phi}{\partial y}\psi\, d\Omega + \oint_\Gamma \phi\psi n_2 d\Gamma. \tag{1.3.14}$$

Exercise 1.3.9 *Prove Theorem 1.3.4.*
Hint: Use Green's theorem (Theorem 1.3.3) with suitable choices for c and $\boldsymbol{u}$. □

1.3.5 Conservation laws

Let us consider some flow field, $\boldsymbol{u}$, in a volume V with boundary Γ. If the net inflow into this volume is positive, *something* in this volume must increase (whatever it is). That is the basic form of a conservation law:

$$\frac{d}{dt}\int_V S\, dV = -\int_\Gamma \boldsymbol{u}\cdot\boldsymbol{n}\, d\Gamma + \int_V f(t,\boldsymbol{x})\, dV. \tag{1.3.15}$$

The term $f(t,\boldsymbol{x})$ is a *production density:* it tells how much S is produced any time, any place within V. The boundary integral describes the net inflow into V (mark the minus sign). The flow field, $\boldsymbol{u}$, is also called the *flux vector* of the model. S just like f has the dimension of a *density.* Since Equation (1.3.15) must hold for every conceivable volume in the flow field, we may formulate a *pointwise* conservation law as follows. First we apply Gauss' divergence theorem 1.3.1 to Equation (1.3.15) to obtain

$$\int_V \frac{\partial}{\partial t} S\, dV = -\int_V \operatorname{div}\boldsymbol{u}\, dV + \int_V f(t,\boldsymbol{x})\, dV. \tag{1.3.16}$$

Note that we also moved the time-derivative d/dt inside the integral, where it has become a partial derivative ($\partial/\partial t$) of course. Subsequently we invoke the mean-value theorem of integral calculus for each integral separately, assuming all integrands are continuous:

$$\frac{\partial S}{\partial t}(\boldsymbol{x}_1) = -\operatorname{div}\boldsymbol{u}(\boldsymbol{x}_2) + f(t,\boldsymbol{x}_3). \tag{1.3.17}$$

Observe that we have divided out a factor $\int_V dV$ and that $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ all lie within V. Finally we let V contract to a single point $\boldsymbol{x}$ to obtain a pointwise conservation law in the form of a PDE:

$$\frac{\partial S}{\partial t} = -\text{div}\, \boldsymbol{u} + f(t, \boldsymbol{x}). \tag{1.3.18}$$

This is all rather abstract, so let us look at an example.

1.3.5.1 Example: Heat flow

In heat flow, conservation law (1.3.18) takes the form

$$\frac{\partial h}{\partial t} = -\text{div}\, \boldsymbol{q} + f(t, \boldsymbol{x}), \tag{1.3.19}$$

in which h is the heat density, $\boldsymbol{q}$ the heat flux vector and f the heat production density. Remember, that all quantities in such a pointwise conservation law are densities. The heat stored in a material can be related to the material's (absolute) temperature T:

$$h = \rho c T, \tag{1.3.20}$$

in which ρ is the mass density and c the heat capacity of the material. These material properties have to be measured. As we already saw in Section 1.3.1, the heat flow, $\boldsymbol{q}$, is driven by the temperature gradient: $\boldsymbol{q} = -\lambda\, \nabla T$. This enables us to formulate everything in terms of temperature. Substituting this all, we get:

$$\frac{\partial \rho c T}{\partial t} = \text{div}\, (\lambda\, \text{grad}\, T) + f(t, \boldsymbol{x}). \tag{1.3.21}$$

If ρ, c are constant throughout the material and if there is no internal heat production this transforms into the celebrated *heat equation*:

$$\frac{\partial T}{\partial t} = \text{div}\, (k\, \text{grad}\, T), \tag{1.3.22}$$

with $k = \lambda/(\rho c)$.

1.4 Preliminaries from linear algebra

In this section we briefly review a number of key concepts from linear algebra that are needed in the forthcoming chapters. Although we assume throughout the book that matrices are real and square, we will consider complex matrices as well in this section. All matrices in this section are therefore square and complex (unless explicitly assumed to be real). We start with a few definitions.

Definition 1.4.1 *A matrix A is called* singular *if there exists a vector $\boldsymbol{v} \in \mathbb{C}^n$ with $A\boldsymbol{v} = \boldsymbol{0}$, $\boldsymbol{v} \neq \boldsymbol{0}$. This is equivalent to the condition* $\det A = 0$, *where* $\det A$ *stands for the determinant of A.*

Definition 1.4.2 *A matrix A is called* nonsingular *(or:* invertible*) if it is not singular. In that case, there exists a unique matrix B with $AB = BA = I$, where I stands for the identity matrix. This matrix B is called the inverse of A and denoted by A^{-1}.*

Of crucial importance are the *eigenvalues* and *eigenvectors* of matrices.

Definition 1.4.3 *Let A be an $n \times n$ matrix. If $\lambda \in \mathbb{C}$ and $\boldsymbol{v} \in \mathbb{C}^n$ satisfy*

$$A\boldsymbol{v} = \lambda \boldsymbol{v}, \quad \boldsymbol{v} \neq \boldsymbol{0}, \tag{1.4.1}$$

then λ is called an eigenvalue *and $\boldsymbol{v}$ an* eigenvector *of A.*

Note that the first part of relation (1.4.1) can be rewritten as $(A - \lambda I)\boldsymbol{v} = \boldsymbol{0}$, showing that the eigenvalues are the values λ for which $A - \lambda I$ is singular, that is,

$$\det(A - \lambda I) = 0. \tag{1.4.2}$$

This is the so-called *characteristic equation*, the roots of which are exactly the eigenvalues of A. The left-hand side of this equation is called the *characteristic polynomial* of A. Since this polynomial has degree n, it follows that the matrix A has exactly n eigenvalues (counted with their multiplicities).

Definition 1.4.4 *Two matrices A and B are called* similar *if there exists a nonsingular matrix V with $A = VBV^{-1}$.*

Exercise 1.4.1 *Show that the matrices A and B have the same eigenvalues if they are similar. Hint: Show that their characteristic polynomials are the same.* □

An important class of matrices consists of the so-called *diagonalizable* matrices.

Definition 1.4.5 *A matrix A is called* diagonalizable *if it is similar to a diagonal matrix.*

Suppose that A is diagonalizable. Then there exists a nonsingular matrix V and a diagonal matrix Λ such that $A = V\Lambda V^{-1}$, and therefore also

$$AV = V\Lambda. \tag{1.4.3}$$

If we denote the diagonal entries of Λ by $\lambda_1, \lambda_2, \ldots, \lambda_n$, and the columns of V by $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$, then relation (1.4.3) can be rewritten as

$$A\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \quad j = 1, 2, \ldots, n. \tag{1.4.4}$$

This simply means that the diagonal entries λ_j are the eigenvalues of A and the columns $\boldsymbol{v}_j$ the corresponding eigenvectors. Since V is nonsingular, we conclude that diagonalizability of A is equivalent to the existence of n linearly independent eigenvectors of A.

We will now turn our attention to symmetric and orthogonal matrices.

Definition 1.4.6 *The transpose of a matrix $A = (a_{ij})$ is the matrix $A^T = (a_{ji})$.*

Definition 1.4.7 *A matrix A is called* symmetric *if $A^T = A$. It is called* real symmetric *if A is real as well.*

Definition 1.4.8 *A matrix A is called* orthogonal *if $A^T A = I$. This is the case if and only if A is invertible with inverse $A^{-1} = A^T$. It is called* real orthogonal *if it is real as well.*

For real symmetric matrices we have the following main result.

Theorem 1.4.1 *A real symmetric matrix A has only real eigenvalues and can be written as*

$$A = Q\Lambda Q^{-1} = Q\Lambda Q^T, \tag{1.4.5}$$

where Λ is a diagonal matrix with the eigenvalues of A on the diagonal, and Q is a real orthogonal matrix with the corresponding eigenvectors of A as columns.

Real symmetric matrices are intimately connected with quadratic forms. The study of quadratic forms requires the introduction of an inner product on $\mathbb{R}^n$. The inner product (or dot product, $\boldsymbol{x} \cdot \boldsymbol{y}$) of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is defined as

$$(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y} = \sum_j x_j y_j. \tag{1.4.6}$$

Note that the inner product is linear in both of its arguments $\boldsymbol{x}$ and $\boldsymbol{y}$, and symmetric, that is, $(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{y}, \boldsymbol{x})$. One further has $(\boldsymbol{x}, \boldsymbol{x}) \geq 0$, and one easily verifies that

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})} \tag{1.4.7}$$

is the well-known *Euclidean norm* on $\mathbb{R}^n$. A useful result is the Cauchy-Schwarz inequality,

$$|(\boldsymbol{x}, \boldsymbol{y})| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|. \tag{1.4.8}$$

Exercise 1.4.2 *Prove the Cauchy-Schwarz inequality.*
Hint: Use that $(\boldsymbol{x} + t\boldsymbol{y}, \boldsymbol{x} + t\boldsymbol{y}) \geq 0$ for all $t \in \mathbb{R}$. □

We consider the following five cases for the quadratic form $(A\boldsymbol{x}, \boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x}$:

Definition 1.4.9 *Let A be a real symmetric matrix.*

- *A is called* positive definite *if $(A\boldsymbol{x}, \boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0}$;*
- *A is called* positive semi-definite *if $(A\boldsymbol{x}, \boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$;*
- *A is called* negative definite *if $(A\boldsymbol{x}, \boldsymbol{x}) < 0$ for all $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0}$;*
- *A is called* negative semi-definite *if $(A\boldsymbol{x}, \boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$;*
- *A is called* indefinite *if none of the above cases applies.*

For the determination of the 'definiteness' of (the quadratic form of) a real symmetric matrix A it is useful to consider its so-called *Rayleigh quotients*.

Definition 1.4.10 *For a real symmetric matrix A, its* Rayleigh quotients *are defined as:*

$$R(A, \boldsymbol{x}) = \frac{(A\boldsymbol{x}, \boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})}, \quad \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x} \neq \boldsymbol{0}. \tag{1.4.9}$$

Theorem 1.4.2 *Let A be real symmetric with eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Then the set of all its Rayleigh quotients is equal to the interval $[\lambda_1, \lambda_n]$, that is,*

$$\{R(A, \boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0}\} = [\lambda_1, \lambda_n]. \tag{1.4.10}$$

Proof
According to Theorem 1.4.1 we can write $A = Q\Lambda Q^T$, where Λ is a diagonal matrix with the eigenvalues λ_j on the diagonal, and Q is a real orthogonal matrix. With the parameterization $\boldsymbol{x} = Q\boldsymbol{y}$, $\boldsymbol{y} \in \mathbb{R}^n$, we find for any $\boldsymbol{y} \neq \boldsymbol{0}$ that

$$R(A, \boldsymbol{x}) = \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \frac{(\boldsymbol{y}^T Q^T)(Q\Lambda Q^T)(Q\boldsymbol{y})}{(\boldsymbol{y}^T Q^T)(Q\boldsymbol{y})} = \frac{\boldsymbol{y}^T \Lambda \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{y}} =$$
$$= \frac{\lambda_1 y_1^2 + \lambda_2 y_2^2 + \ldots + \lambda_n y_n^2}{y_1^2 + y_2^2 + \ldots + y_n^2},$$

from which we immediately see that the set of all Rayleigh quotients of A is equal to the set of all convex combinations (weighted averages) of its eigenvalues, which is exactly $[\lambda_1, \lambda_n]$. □

Corollary 1.4.3 *Let A be real symmetric with eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Then:*

- *A is positive definite if and only if $\lambda_1 > 0$;*
- *A is positive semi-definite if and only if $\lambda_1 \geq 0$;*
- *A is negative definite if and only if $\lambda_n < 0$;*
- *A is negative semi-definite if and only if $\lambda_n \leq 0$;*
- *A is indefinite if and only if $\lambda_1 < 0 < \lambda_n$.*

Exercise 1.4.3 *Prove Corollary 1.4.3.* □

The next topic in this section is matrix norms. Along with the vector norm $\|\cdot\|$ defined in (1.4.7) for vectors in $\mathbb{R}^n$, we introduce the *induced matrix norm* as

$$\|A\| = \max\{\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} : \boldsymbol{x} \in \mathbb{R}^n, \ \boldsymbol{x} \neq 0\}. \tag{1.4.11}$$

Theorem 1.4.4

(i) For a real square matrix A we have $\|A\| = \sqrt{\lambda_{max}(A^TA)}$, where $\lambda_{max}(A^TA)$ denotes the largest eigenvalue of A^TA.

(ii) For a real symmetric matrix A with eigenvalues λ_j we have $\|A\| = \max_j |\lambda_j|$.

Proof
For a real square matrix A and a vector $\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{x} \neq \mathbf{0}$, we have

$$\frac{\|A\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|^2} = \frac{(A\boldsymbol{x}, A\boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})} = \frac{(A\boldsymbol{x})^TA\boldsymbol{x}}{(\boldsymbol{x}, \boldsymbol{x})} = \frac{\boldsymbol{x}^TA^TA\boldsymbol{x}}{(\boldsymbol{x}, \boldsymbol{x})} = \frac{(A^TA\boldsymbol{x}, \boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})} = R(A^TA, \boldsymbol{x}).$$

Since A^TA is real symmetric, the proof of part (i) follows from Theorem 1.4.2.

If A is real symmetric with eigenvalues λ_j, then the matrix $A^TA = A^2$ has eigenvalues λ_j^2. Hence it follows from part (i) that in that case

$$\|A\| = \sqrt{\lambda_{\max}(A^TA)} = \sqrt{\max_j \lambda_j^2} = \max_j |\lambda_j|,$$

which proves part (ii). □

Exercise 1.4.4 *Prove that for any real $n \times n$ matrix A and vector $\boldsymbol{x} \in \mathbb{R}^n$ we have*

$$\|A\boldsymbol{x}\| \leq \|A\|\|\boldsymbol{x}\|. \tag{1.4.12}$$

Exercise 1.4.5 *Prove that the induced matrix norm defined in (1.4.11) is* submultiplicative, *that is, for all real square matrices A, B of the same size one has*

$$\|AB\| \leq \|A\|\|B\|. \tag{1.4.13}$$

□

So far we have only considered the Euclidean vector norm (1.4.7) on $\mathbb{R}^n$, which is also called the L_2 norm. This norm and the corresponding induced matrix norm are often denoted by $\|\cdot\|_2$. Another popular (and useful) vector norm is the so-called *maximum norm*, defined as

$$\|\boldsymbol{x}\|_\infty = \max_j |x_j|. \tag{1.4.14}$$

Exercise 1.4.6 *Prove that the induced matrix norm corresponding to the maximum norm is given by*

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|. \tag{1.4.15}$$

□

We conclude this section with a theorem that can be of great help in estimating bounds for eigenvalues of matrices. This is for example useful in stability analysis.

Theorem 1.4.5 (Gershgorin)
For each eigenvalue λ of an $n \times n$ matrix A there exists an index i such that

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|. \tag{1.4.16}$$

Remark:
Eigenvalues may be complex-valued in general and for complex eigenvalues $\lambda = \mu + i\nu$, the absolute value is the *modulus:* $|\lambda| = \sqrt{\mu^2 + \nu^2}$. So the eigenvalues of A are located within the union of n disks in the complex plane and that is the reason why the theorem is also often referred to as Gershgorin's *disk (or circle)* theorem. But for real symmetric A, the eigenvalues of A are real-valued.

Proof
Let λ be an eigenvalue of A with corresponding eigenvector $\boldsymbol{v}$, that is, $A\boldsymbol{v} = \lambda\boldsymbol{v}$, or equivalently,

$$\sum_j a_{ij} v_j = \lambda v_i, \quad i = 1, \ldots, n. \tag{1.4.17}$$

Let v_i be the component of $\boldsymbol{v}$ with the largest modulus. For the corresponding index i we have

$$\lambda - a_{ii} = \sum_{j:j \neq i} a_{ij} \frac{v_j}{v_i}, \tag{1.4.18}$$

and because $|v_j / v_i| \leq 1$ (for all j), we get

$$|\lambda - a_{ii}| \leq \sum_{j:j \neq i} |a_{ij}|. \tag{1.4.19}$$

This proves the theorem. □

1.5 The Poincaré inequality

The following theorem is used in the proof of Theorem 6.1.1 to show that the Laplace operator is negative definite. As a matter of fact we use a generalization of the theorem dealing with more general boundary conditions. The space of square integrable real funcions on a domain $\Omega \subset \mathbb{R}^m$ is denoted by

$$L^2(\Omega) := \{u | u : \Omega \to \mathbb{R}, \int_\Omega u^2 d\Omega < \infty\},$$

and the corresponding first order Sobolev space by

$$H^1(\Omega) := \{u \in L^2(\Omega) | \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_m} \in L^2(\Omega)\}.$$

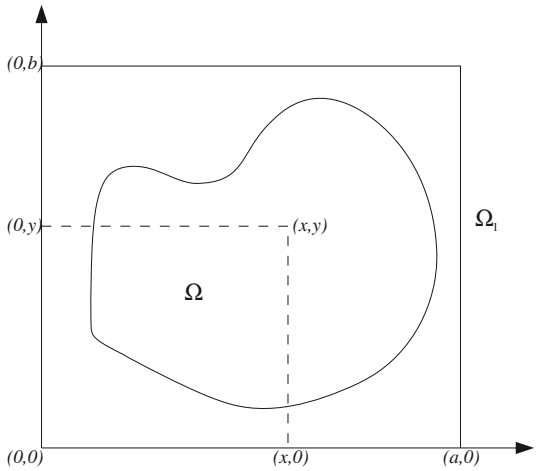

Figure 1.4: 2-dimensional region.

Theorem 1.5.1 *Inequality of Poincaré (Friedrichs)*
Let Ω be a bounded domain in $\mathbb{R}^m$. Then there exists a constant $K > 0$ such that for all $u \in H^1(\Omega)$ with $u|_\Gamma = 0$ we have

$$\int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 d\Omega \geq K \int_\Omega u^2 \, d\Omega. \tag{1.5.1}$$

Proof We shall prove the theorem for $m = 2$.
By shifting coordinates we may assume that $(x, y) \in \Omega$ implies $x > 0$ and $y > 0$. As displayed in Figure 1.4, the region Ω is therefore contained in a rectangular region $\Omega_1 = [0, a] \times [0, b]$. We extend u to a function on the whole domain Ω_1 by defining

$$u(x, y) = 0, \quad (x, y) \in \Omega_1 \backslash \Omega. \tag{1.5.2}$$

Let (x_1, y_1) be an arbitrary point in Ω_1. Then

$$u(x_1, y_1) - u(0, y_1) = \int_0^{x_1} \frac{\partial u(x, y_1)}{\partial x} \, dx, \tag{1.5.3}$$

$$u(0, y_1) = 0 \quad \text{(follows from Figure 1.4).} \tag{1.5.4}$$

According to the Cauchy-Schwarz inequality for the inner product $(f, g) = \int_\alpha^\beta f(x)g(x)dx$ we have:

$$\left(\int_\alpha^\beta f(x)g(x) \, dx \right)^2 \leq \int_\alpha^\beta f(x)^2 \, dx \int_\alpha^\beta g(x)^2 \, dx. \tag{1.5.5}$$

Applying this with $\alpha = 0$, $\beta = x_1$, $f(x) = 1$ and $g(x) = \frac{\partial u(x,y_1)}{\partial x}$ yields

$$\begin{aligned} u^2(x_1,y_1) &= \left(\int_0^{x_1} \frac{\partial u(x,y_1)}{\partial x}\, dx \right)^2 \leq x_1 \int_0^{x_1} \left(\frac{\partial u(x,y_1)}{\partial x} \right)^2 dx \\ &\leq a \int_0^{a} \left(\frac{\partial u(x,y_1)}{\partial x} \right)^2 dx. \end{aligned} \tag{1.5.6}$$

Integration of inequality (1.5.6) over Ω_1 gives

$$\begin{aligned} \int_{\Omega_1} u^2 d\Omega &= \int_0^a \int_0^b u^2(x_1,y_1)\, dy_1\, dx_1 \\ &\leq a \int_0^a \int_0^b \int_0^a \left(\frac{\partial u(x,y_1)}{\partial x} \right)^2 dx\, dy_1\, dx_1 \\ &\leq a^2 \int_0^b \int_0^a \left(\frac{\partial u(x,y_1)}{\partial x} \right)^2 dx\, dy_1 \\ &= a^2 \int_{\Omega_1} \left(\frac{\partial u}{\partial x} \right)^2 d\Omega \\ &\leq a^2 \int_{\Omega_1} \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] d\Omega. \end{aligned}$$

This proves the theorem with $K = 1/a^2$. □

Exercise 1.5.1 *Prove Theorem 1.5.1 with $K = 1/b^2$.* □

We note that it follows from the proof of Theorem 1.5.1 and Exercise 1.5.1 that $K = \max(1/a^2, 1/b^2)$ is a lower bound for the best (largest) possible K.

1.6 Summary of Chapter 1

In this chapter we have seen the importance of conservation laws in the development of models and the role the mathematical operators *divergence* and *gradient* play in that development. We have met the famous divergence theorem of Gauss as an expression of global conservation.

We have looked at various applications deriving from conservation: heat transfer and diffusion. We concluded the chapter with preliminaries from linear algebra and an inequality due to Poincaré.

Chapter 2

A crash course in PDEs

Objectives

In the previous chapter we looked at PDEs from the *modeling* point of view, but now we shall look at them from a *mathematical* angle. Apparently you need partial derivatives and at least *two* independent variables to speak of a PDE (with fewer variables you would have an ordinary differential equation), so the simplest case to consider is a PDE with exactly two independent variables. A second aspect is the *order* of the PDE, that is the order of the highest derivative occurring in it. First order PDEs are a class of their own: the *transport* equations. These equations are beyond the scope of this book, and are only dealt with in the unabridged version [15] of the book. In this chapter we shall concentrate on second order PDEs and show that (for two independent variables) they can be classified into three types. We shall provide boundary and initial conditions that are needed to guarantee a unique solution and we will consider a few properties of the solutions to these PDEs. We conclude the chapter with a few examples of second and fourth order equations that occur in various fields of physics and technology.

2.1 Classification

Consider a linear second order PDE in two independent variables *with constant coefficients*,

$$a_{11}\frac{\partial^2 u}{\partial x^2} + 2a_{12}\frac{\partial^2 u}{\partial x \partial y} + a_{22}\frac{\partial^2 u}{\partial y^2} + b_1\frac{\partial u}{\partial x} + b_2\frac{\partial u}{\partial y} + cu + d = 0. \tag{2.1.1}$$

By *rotating* the coordinate system we can make the term with the mixed second derivative vanish. This is the basis of the classification. To carry out this

rotation, we keep in mind that

$$(\frac{\partial}{\partial x}, \frac{\partial}{\partial y})A \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix} = a_{11}\frac{\partial^2 u}{\partial x^2} + 2a_{12}\frac{\partial^2 u}{\partial x \partial y} + a_{22}\frac{\partial^2 u}{\partial y^2}, \tag{2.1.2}$$

where $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$. Since A is real symmetric, it follows from Theorem 1.4.1 that we can factorize A into $A = Q\Lambda Q^T$, where $\Lambda = \text{diag}\,(\alpha_{11}, \alpha_{22})$, in which α_{11} and α_{22} are eigenvalues of A, and Q is a real orthogonal matrix (cf. Definition 1.4.8) whose columns are the normalized (with length one) eigenvectors of A. Hence one obtains from Equation (2.1.2)

$$\begin{aligned} &a_{11}\frac{\partial^2 u}{\partial x^2} + 2a_{12}\frac{\partial^2 u}{\partial x \partial y} + a_{22}\frac{\partial^2 u}{\partial y^2} = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})Q\Lambda Q^T \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix} = \\ &(\frac{\partial}{\partial \xi}, \frac{\partial}{\partial \eta})\Lambda \begin{pmatrix} \frac{\partial u}{\partial \xi} \\ \frac{\partial u}{\partial \eta} \end{pmatrix} = \alpha_{11}\frac{\partial^2 u}{\partial \xi^2} + \alpha_{22}\frac{\partial^2 u}{\partial \eta^2}. \end{aligned} \tag{2.1.3}$$

The resulting equation will look like:

$$\alpha_{11}\frac{\partial^2 u}{\partial \xi^2} + \alpha_{22}\frac{\partial^2 u}{\partial \eta^2} + \beta_1\frac{\partial u}{\partial \xi} + \beta_2\frac{\partial u}{\partial \eta} + cu + d = 0. \tag{2.1.4}$$

Exercise 2.1.1 *Show that $a_{11}a_{22} - a_{12}^2 > 0$, $a_{11}a_{22} - a_{12}^2 = 0$ and $a_{11}a_{22} - a_{12}^2 < 0$ correspond to $\alpha_{11}\alpha_{22} > 0$, $\alpha_{11}\alpha_{22} = 0$ and $\alpha_{11}\alpha_{22} < 0$, respectively. (These cases correspond to the situations in which the eigenvalues of A have the same sign, one of the eigenvalues of A is zero and opposite signs of the eigenvalues of A, respectively.)* □

There are three possibilities:

1. $\alpha_{11}\alpha_{22} > 0$. (I.e. both coefficients have the same sign) The equation is called *elliptic*. An example of this case is *Poisson's equation*

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f. \tag{2.1.5}$$

2. $\alpha_{11}\alpha_{22} < 0$. (I.e. both coefficients have opposite sign) The equation is called *hyperbolic*. An example of this case is the *wave equation* in one space dimension:

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0. \tag{2.1.6}$$

3. $\alpha_{11}\alpha_{22} = 0$. (I.e. either coefficient vanishes). The equation is called *parabolic*. An example is the *heat equation* in one space dimension:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}. \tag{2.1.7}$$

Exercise 2.1.2 *Let $D = a_{11}a_{22} - a_{12}^2$. Show that the condition for elliptic, parabolic or hyperbolic in the original coefficients a_{ij} is given by $D > 0$, $D = 0$ and $D < 0$, respectively. Use the result of Exercise 2.1.1.* □

For the classification only the second order part of the PDE is important. The three different types have very different physical and mathematical properties. To begin with, elliptic equations are *time-independent* and often describe an *equilibrium* or *steady state*. Parabolic and hyperbolic equations are *time-dependent*: they describe the *time evolution* or *transient behavior* of a process. The difference in nature between parabolic and hyperbolic equations is that the first class describes an evolution towards an equilibrium, whereas the second class mimics wave phenomena.

This classification strictly spoken holds only for equations with constant coefficients. For equations with varying coefficients this classification only holds *locally*. If the coefficients depend on the solution itself, the PDE is called *quasi-linear*, and its type will depend on the solution itself.

2.1.1 Three or more independent variables

In this section, we consider a generalization of the simple classification. The general second order part of a *quasi-linear* PDE in $N > 2$ independent variables is given by:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j}. \tag{2.1.8}$$

Without loss of generality we may assume that $a_{ij} = a_{ji}$ and in a way similar to that in the previous section one may remove the mixed derivatives. This leads to:

$$\sum_{i=1}^{N} \alpha_{ii}\frac{\partial^2 u}{\partial \xi_i^2}. \tag{2.1.9}$$

We treat the following cases in this book:

1. All α_{ii} have the same sign. In this case all independent variables ξ_i are space variables. The equation is called *elliptic*. Example: Laplace's equation in 3D:
$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} - \frac{\partial^2 u}{\partial z^2} = 0. \tag{2.1.10}$$

2. Exactly one α_{ii}, say α_{11}, has different sign from the rest. In this case ξ_1 is a time variable, all other ξ_i are space variables. The equation is called *hyperbolic*. Example: 2D wave equation
$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{2.1.11}$$

3. Exactly one α_{ii} vanishes, say α_{11}, while the other α_{ii} have the same sign. Then ξ_1 is a time variable and the equation is called *parabolic*. Example: 2D heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{2.1.12}$$

Exercise 2.1.3 *If A is a real symmetric $N \times N$ matrix, then (cf. Theorem 1.4.1) there exists a real orthogonal matrix Q such that $Q^T A Q = \Lambda$, where Λ is a diagonal matrix containing the eigenvalues of A on the diagonal. Show that the substitution $\xi = Q^T \boldsymbol{x}$ eliminates the mixed derivatives in the differential operator* div A grad u.

2.2 Boundary and initial conditions

To ensure a unique solution to our PDE we need to prescribe appropriate boundary conditions and for time-dependent problems we need initial conditions too. We will just consider second order PDEs here because the considerations for first order PDEs are very different and beyond the scope of the book.

2.2.1 Boundary conditions

Consider in Figure 2.1 the bounded region $\Omega \subset \mathbb{R}^2$ with boundary Γ .

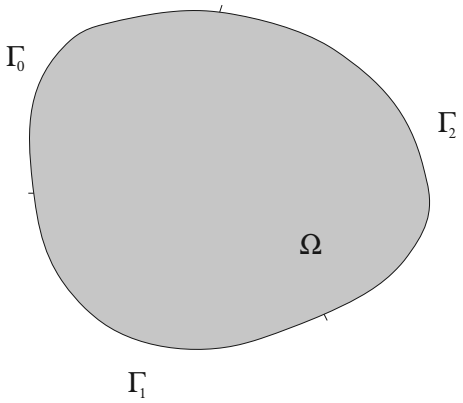

Figure 2.1: The bounded region Ω.

Let Γ consist of three *disjoint* pieces Γ_0, Γ_1 and Γ_2. For an elliptic equation of the form

$$-\text{div}\, k \,\text{grad}\, u = f, \tag{2.2.1}$$

with $k(\boldsymbol{x}) > 0$ (for all $\boldsymbol{x} \in \overline{\Omega}$), the following boundary conditions guarantee a unique solution:

1. the *Dirichlet boundary condition:*

$$u = g_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_0, \tag{2.2.2}$$

2. the *Neumann boundary condition:*

$$k\frac{\partial u}{\partial n} = g_1(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_1, \tag{2.2.3}$$

3. the *Robin*, *radiation*, *kinetic* or *mixed* boundary condition:

$$k\frac{\partial u}{\partial n} + \sigma u = g_2(\boldsymbol{x}), \quad \sigma \geq 0,\ \boldsymbol{x} \in \Gamma_2. \tag{2.2.4}$$

These boundary conditions do not have to occur together, each (but not all) of Γ_0, Γ_1 or Γ_2 could be empty. Because the pieces are disjoint, exactly *one* boundary condition occurs at each point of the boundary. There is a small problem if $\Gamma = \Gamma_1$, in other words, if there is a Neumann boundary condition on all of the boundary. Physically this may be understood as that the *inflow* at each point of the boundary is prescribed. And since we have an equilibrium, the net inflow over the whole region must be annihilated inside or the net outflow must be produced inside. This result is stated in mathematical form in the following theorem.

Theorem 2.2.1 *If a Neumann boundary condition is given on all of* Γ*, then the solution* u *of Equation* (2.2.1) *is determined up to an additive constant only. Moreover the following* compatibility condition *must be satisfied:*

$$-\int_\Gamma g_1 \, d\Gamma = \int_\Omega f \, d\Omega. \tag{2.2.5}$$

Exercise 2.2.1 *Prove Theorem 2.2.1. Use Gauss' divergence theorem on the PDE. It is not necessary to prove the only part.* □

Remarks

1. Only the highest order part of the PDE determines what type of boundary conditions are needed, so the same set of boundary conditions is needed if first and zeroth order terms are added to elliptic equation (2.2.1).

2. On each part of the boundary *precisely one* boundary condition applies. This situation is not restricted to elliptic equations but also applies to the wider class of general second order PDEs.

3. Boundary conditions involving the flux vector (Neumann, Robin) are also called *natural boundary conditions.* Boundary conditions only involving the value of u (Dirichlet) are called *essential boundary conditions.*

4. The boundary conditions needed in parabolic and hyperbolic equations are determined by the spatial part of the equation.

5. If the coefficients of the terms of highest order are *very small* compared to the coefficients of the lower order terms, it is to be expected that the nature of the solution is mostly determined by those lower order terms. Such problems are called *singularly perturbed*. An example is the convection-dominated convection-diffusion equation (see Section 3.3).

2.2.2 Initial conditions

Initial conditions only play a role in time-dependent problems, and we can be very short. If the equation is first order in time, u has to be given on all of Ω at $t = t_0$. If the equation is second order in time, $\frac{\partial u}{\partial t}$ has to be given as well on all of Ω at $t = t_0$.

Exercise 2.2.2 *Consider the transversal vibrations of a membrane that is fixed to an iron ring. These vibrations are described by the wave equation. What is the type of boundary condition? What initial conditions are needed?* □

2.3 Existence and uniqueness of a solution

Physicists and technicians usually consider the mathematical chore of proving existence and uniqueness of a solution a waste of time. 'I know the process behaves in precisely one way', they will claim and of course they are right in that. What they do not know is: if their mathematical model describes their process with any accuracy then existence and uniqueness of a solution is an acid test for that. In the simplest ODEs a practical way to go about this is try and find one. In PDEs this is not much of an option, since solutions in closed form are rarely available.

Proving existence and uniqueness is usually a very difficult assignment, but to get some of the flavor we shall look at a relatively simple example: Poisson's equation (2.1.5). We shall prove that a solution to this equation with Dirichlet boundary conditions on all of Γ is unique.

2.3.1 The Laplace operator

The Laplace operator div grad is such a fundamental operator that it has a special symbol in the literature: Δ. So the following notations are equivalent:

$$\nabla \cdot \nabla u \equiv \text{div grad}\, u \equiv \Delta\, u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{2.3.1}$$

In a technical context div grad is mostly used, in mathematical contexts the other three. The Laplace operator is often referred to as the *Laplacian*.

In a physical context it is clear that if there are no sources, a heat equation in equilibrium takes its minimum and maximum at the boundary. Mathematically this is also true as we shall show in the next subsection.

2.3.2 The maximum principle and uniqueness

Solutions to Laplace's and Poisson's equation satisfy certain properties with respect to existence, uniqueness and the occurrence of extremal values at the boundaries of a bounded domain or in the domain. We note that a function $u(\boldsymbol{x})$ has a local maximum in some point $\boldsymbol{x}_0 \in \Omega$ if there exists a $\delta > 0$ such that $u(\boldsymbol{x}_0) \geq u(\boldsymbol{x})$ for all $\boldsymbol{x}$ with $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta$.

Definition 2.3.1 *The Hessian matrix in $\mathbb{R}^2$ is defined as*

$$H(\boldsymbol{x}_0) = \begin{pmatrix} \frac{\partial^2 u}{\partial x^2}(\boldsymbol{x}_0) & \frac{\partial^2 u}{\partial x \partial y}(\boldsymbol{x}_0) \\ \frac{\partial^2 u}{\partial y \partial x}(\boldsymbol{x}_0) & \frac{\partial^2 u}{\partial y^2}(\boldsymbol{x}_0) \end{pmatrix}. \tag{2.3.2}$$

Theorem 2.3.1 *Assume that the function $u = u(\boldsymbol{x})$ is defined and sufficiently smooth in a neighborhood of the point $\boldsymbol{x}_0$. If u has a local maximum in $\boldsymbol{x}_0$ then the gradient $\nabla u(x_0)$ must be zero and the Hessian matrix $H(\boldsymbol{x}_0)$ must be negative semi-definite.*

Proof Consider the 2-D Taylor expansion of u around $\boldsymbol{x}_0$:

$$\begin{aligned} u(\boldsymbol{x}) = u(\boldsymbol{x}_0) &+ \nabla u(\boldsymbol{x}_0) \cdot (\boldsymbol{x} - \boldsymbol{x}_0) \\ &+ \frac{1}{2}(H(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0) + \mathcal{O}(\|\boldsymbol{x} - \boldsymbol{x}_0\|^3). \end{aligned} \tag{2.3.3}$$

Since u has a local maximum in $\boldsymbol{x}_0$, there exists a $\delta > 0$ such that $u(\boldsymbol{x}_0) \geq u(\boldsymbol{x})$ for all $\boldsymbol{x}$ with $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta$. First we note that we must have $\nabla u(\boldsymbol{x}_0) = \boldsymbol{0}$. This easily follows from (2.3.3) by considering $\boldsymbol{x} = \boldsymbol{x}_0 + t\nabla u(\boldsymbol{x}_0)$ for $t \approx 0$, since for this choice of $\boldsymbol{x}$ we have

$$0 \geq u(\boldsymbol{x}) - u(\boldsymbol{x}_0) = t\|\nabla u(\boldsymbol{x}_0)\|^2 + \mathcal{O}(t^2),$$

which is only possible if $\nabla u(\boldsymbol{x}_0) = \boldsymbol{0}$.
Let $\boldsymbol{v}$ be an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^2$. We show that $(H(\boldsymbol{x}_0)\boldsymbol{v}, \boldsymbol{v}) \leq 0$ by considering $\boldsymbol{x} = \boldsymbol{x}_0 + t\boldsymbol{v}$ for $t > 0$ so small that $t\|\boldsymbol{v}\| < \delta$. For these values of t we have

$$0 \geq u(\boldsymbol{x}) - u(\boldsymbol{x}_0) = \frac{1}{2}t^2(H(\boldsymbol{x}_0)\boldsymbol{v}, \boldsymbol{v}) + \mathcal{O}(t^3), \tag{2.3.4}$$

from which $(H(\boldsymbol{x}_0)\boldsymbol{v}, \boldsymbol{v}) \leq 0$ follows. Hence $H(\boldsymbol{x}_0)$ is negative semi-definite. □

Exercise 2.3.1 *Prove that $H(\boldsymbol{x}_0)$ is positive semi-definite if u has a local minimum in $\boldsymbol{x}_0$.* □

Exercise 2.3.2 *Show that if H is positive semi-definite, then both diagonal elements must be non-negative. Hint: Make special choices for $\boldsymbol{v}$ in $(H\boldsymbol{v}, \boldsymbol{v})$.*

Corollary 2.1 Assume that the function $u = u(\boldsymbol{x})$ is defined and sufficiently smooth in a neighborhood of the point $\boldsymbol{x}_0$.

(i) If u has a local maximum in $\boldsymbol{x}_0$ then $\nabla u(\boldsymbol{x}_0) = \boldsymbol{0}$ and $-\Delta u(\boldsymbol{x}_0) \geq 0$;

(ii) If u has a local minimum in $\boldsymbol{x}_0$ then $\nabla u(\boldsymbol{x}_0) = \boldsymbol{0}$ and $-\Delta u(\boldsymbol{x}_0) \leq 0$.

Proof Combine Theorem 2.3.1 and Exercises 2.3.1, 2.3.2. □

Exercise 2.3.3 *Give an alternative (simpler) proof of Corollary 2.1 by only considering $u(x, y)$ in the x-direction and y-direction.* □

Next we are going to consider solutions to Laplace's equation, $-\Delta u = 0$.

Definition 2.3.2 *A function satisfying Laplace's equation $-\Delta u = 0$ in Ω is called* harmonic *in Ω.*

Definition 2.3.3 *A function satisfying $-\Delta u \leq 0$ in Ω is called* subharmonic *in Ω.*

Definition 2.3.4 *A function satisfying $-\Delta u \geq 0$ in Ω is called* superharmonic *in Ω.*

Theorem 2.3.2 *(Maximum principle)*
Let Ω be a bounded domain with boundary Γ and closure $\overline{\Omega}$, that is $\overline{\Omega} = \Omega \cup \Gamma$.
If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is subharmonic in Ω, then

(i) (Weak maximum principle)
At no point in Ω can the value of u exceed the maximum value of u on Γ.

(ii) (Strong maximum principle)
If there is a point $\boldsymbol{x}_0$ in Ω where u reaches its maximum, i.e., $u(\boldsymbol{x}_0) = \max_{\boldsymbol{x} \in \overline{\Omega}} u$, then u is constant on $\overline{\Omega}$, that is $u(\boldsymbol{x}) = u(\boldsymbol{x}_0)$ on $\overline{\Omega}$.

This theorem is formulated and proved in Evans [7] among others. To prove the maximum principle, we shall use the arguments given in Protter and Weinberger [11]. Theorem 2.3.2 says that the maximum of a subharmonic function is only found on the boundary Γ unless the function is constant. By replacing u by $-u$, we recover for superharmonic functions u similar assertions as in Theorem 2.3.2 with *min* replacing *max*. Before we prove the theorem we give several corollaries.

Theorem 2.3.3 *Laplace's equation in Ω with a homogeneous Dirichlet boundary condition, that is $u = 0$ on Γ, has only the trivial solution, that is $u = 0$ in Ω.*

Exercise 2.3.4 *Prove Theorem 2.3.3.* □

Theorem 2.3.4 *(uniqueness) Let Ω be a bounded region in $\mathbb{R}^2$ with boundary Γ. Then the problem*

$$-\Delta u = f(x,y), \qquad (x,y) \in \Omega, \tag{2.3.5}$$
$$u = g(x,y), \qquad (x,y) \in \Gamma \tag{2.3.6}$$

has at most one solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$.

Exercise 2.3.5 *Prove Theorem 2.3.4.*
Hint: assume that there are two solutions u_1 and u_2 and consider the difference. □

Next we prove part (i) of Theorem 2.3.2.

Proof of Theorem 2.3.2.
We prove the theorem for $\Omega \subset \mathbb{R}^2$. Any dimensionality is dealt with analogously.
Let u_m be the maximum on Γ, that is $u \leq u_m$ on Γ. We introduce the function

$$v(x,y) = u(x,y) + \epsilon(x^2+y^2), \text{ with } \epsilon > 0 \text{ arbitrary.} \tag{2.3.7}$$

Since u is subharmonic, this implies

$$-\Delta v \leq -4\epsilon < 0, \quad \text{in } \Omega. \tag{2.3.8}$$

Suppose that v has a local maximum in the domain Ω, then according to Corollary 2.1 we have $-\Delta v \geq 0$. This contradicts with the strict inequality (2.3.8), and hence v cannot have a local maximum in Ω. Since Ω is a bounded domain in $\mathbb{R}^2$, there exists a finite radius R such that

$$R = \max_{\boldsymbol{x} \in \Gamma} \|\boldsymbol{x}\| = \max_{\boldsymbol{x} \in \Gamma} \sqrt{x^2+y^2}. \tag{2.3.9}$$

This implies $v(x,y) \leq u_m + \epsilon R^2$ on Γ. Since v does not have a maximum within the interior Ω, we deduce

$$u(\boldsymbol{x}) \leq v(\boldsymbol{x}) \leq u_m + \epsilon R^2, \quad \text{in } \overline{\Omega} = \Omega \cup \Gamma. \tag{2.3.10}$$

Since $\epsilon > 0$ can be taken arbitrarily small, we get $u \leq u_m$ in $\overline{\Omega}$. Hence at no point in Ω, the value of u can exceed the maximum value of u on Γ.
For the proof of (ii) we refer to [11]. □

Uniqueness for the solution to the Poisson equation with Robin conditions can also be proved easily.

Theorem 2.3.5 *(uniqueness) Let Ω be a bounded domain in $\mathbb{R}^2$ with boundary Γ. Then the problem*

$$-\Delta u = f(x,y), \qquad (x,y) \in \Omega, \tag{2.3.11}$$
$$\sigma u + \frac{\partial u}{\partial n} = g(x,y), \qquad (x,y) \in \Gamma, \tag{2.3.12}$$

with $\sigma > 0$, has at most one solution $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$.

Exercise 2.3.6 *Prove Theorem 2.3.5.*
Hints: Assume that there are two solutions u_1 and u_2 and consider the difference $v = u_1 - u_2$. Use multiplication by v and integration by parts to conclude that $v = 0$ on $\overline{\Omega}$. □

Theorem 2.3.6 *Let $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfy*

$$-\Delta u \geq 0, \qquad \text{in } \Omega, \tag{2.3.13}$$
$$u = 0, \qquad \text{on } \Gamma, \tag{2.3.14}$$

where Ω is a bounded domain with boundary Γ. Then $u \geq 0$ in Ω.

Exercise 2.3.7 *Show that Theorem 2.3.6 is a corollary of Theorem 2.3.2.*

Exercise 2.3.8 *Let a, b, c be given constants with $ac - b^2 > 0$. Show that the elliptic operator $au_{xx} + 2bu_{xy} + cu_{yy}$ satisfies the same maximum principle as the Laplace operator.*
Hint: Use scaling and rotation of the coordinates. □

Qualitative properties of the solutions to Poisson's or Laplace's equation like the maximum principle are an important tool to evaluate the quality of numerical solutions. Indeed we want our numerical solution to inherit these properties.

2.3.3 Existence

To prove *existence* of a solution of Poisson's equation is very hard. In general one needs extra requirements on the smoothness of the boundary. This is far outside the scope of this book, the interested reader may look at [5].

2.4 Examples

In this section we give a few examples of PDEs that describe physical and technical problems. For all problems we consider a bounded region $\Omega \subset \mathbb{R}^2$ with boundary Γ.

2.4.1 Flows driven by a potential

Flows driven by a potential we already met in Chapter 1. They all have the form

$$\frac{\partial c(u)}{\partial t} = \operatorname{div} \lambda \operatorname{grad} u + f(t, \boldsymbol{x}, u). \tag{2.4.1}$$

For uniqueness c must be a monotone function of u and for stability it must be non-decreasing. In ordinary heat transfer and diffusion, c is linear. In phase transition problems and diffusion in porous media it is non-linear. If f depends on u, the function f may influence the stability of the equation.

2.4.1.1 Boundary conditions

In Section 2.2 three types of linear boundary conditions have been introduced. These conditions may occur in any combination. This is not a limitative enumeration, there are other ways to couple the heat flow at the boundary to the temperature difference one way or another, mostly non-linear.

2.4.1.2 Initial condition

To guarantee that Problem (2.4.1) with boundary conditions (2.2.2) to (2.2.4) has a unique solution $u(\boldsymbol{x}, t)$, it is necessary that u is prescribed at $t = t_0$: $u(\boldsymbol{x}, t_0) = u_0(\boldsymbol{x}), \forall \boldsymbol{x} \in \Omega$.

2.4.1.3 Equilibrium

An equilibrium of Equation (2.4.1) is reached when all temporal dependence has disappeared. But this problem can also be considered in its own right:

$$-\operatorname{div} \lambda \operatorname{grad} u = f(\boldsymbol{x}, u), \tag{2.4.2}$$

with boundary conditions (2.2.2) to (2.2.4).

2.4.2 Convection-Diffusion

The *convection-diffusion* equation describes the transport of a pollutant with concentration, c, by a transporting medium with given velocity, $\boldsymbol{u}$. The equation is

$$\frac{\partial c}{\partial t} + \boldsymbol{u} \cdot \operatorname{grad} c = \operatorname{div} \lambda \operatorname{grad} c + f(t, \boldsymbol{x}, c). \tag{2.4.3}$$

Comparing Equation (2.4.3) with (2.4.1) shows that a *convection term* $\boldsymbol{u} \cdot \operatorname{grad} c$ has been added. Boundary and initial conditions are the same as for the potential-driven flows.

In cases where the diffusion coefficient, λ, is small compared to the velocity, $\boldsymbol{u}$, the flow is *dominated* by the convection. The problem then becomes *singularly perturbed* and in these cases the influence of the second order term is mostly felt at the boundary in the form of *boundary layers*. This causes specific difficulties in the numerical treatment, see for example Section 3.3.

2.4.3 Navier-Stokes equations

The Navier-Stokes equations describe the dynamics of material flow. The momentum equations are given by:

$$\rho\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y}\right) = \operatorname{div} \boldsymbol{s}_x + \rho b_x, \tag{2.4.4a}$$

$$\rho\left(\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y}\right) = \operatorname{div} \boldsymbol{s}_y + \rho b_y. \tag{2.4.4b}$$

We shall not derive the equations (see for instance [3]), but we will say a few things about their interpretation. The equations describe Newton's second law on a small volume V of fluid with density, ρ, and velocity, $\boldsymbol{u} = \begin{pmatrix} u \\ v \end{pmatrix}$, *moving along with the flow*. Thus, a particle $P \in V$ with coordinates $\boldsymbol{x}$ at time t has at time $t + \Delta t$, with $\Delta t \to 0$, coordinates $\boldsymbol{x} + \boldsymbol{u}\Delta t$. Therefore the change in velocity of a *moving* particle is described by

$$\Delta \boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x} + \boldsymbol{u}\Delta t, t + \Delta t) - \boldsymbol{u}(\boldsymbol{x}, t). \tag{2.4.5}$$

We recall Taylor's theorem in three variables:

$$f(x+h, y+k, t+\tau) = f(x,y) + h\frac{\partial f}{\partial x} + k\frac{\partial f}{\partial y} + \tau\frac{\partial f}{\partial t} + \mathcal{O}(h^2 + k^2 + \tau^2). \tag{2.4.6}$$

Applying this to Equation (2.4.5) we get:

$$\Delta u = u\Delta t\frac{\partial u}{\partial x} + v\Delta t\frac{\partial u}{\partial y} + \Delta t\frac{\partial u}{\partial t}, \tag{2.4.7a}$$

$$\Delta v = u\Delta t\frac{\partial v}{\partial x} + v\Delta t\frac{\partial v}{\partial y} + \Delta t\frac{\partial v}{\partial t}. \tag{2.4.7b}$$

If we divide both sides by Δt and let $\Delta t \to 0$ we find the *material derivative*

$$\frac{Du}{Dt} = u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + \frac{\partial u}{\partial t}, \tag{2.4.8a}$$

$$\frac{Dv}{Dt} = u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + \frac{\partial v}{\partial t}. \tag{2.4.8b}$$

The right hand side of Equations (2.4.4) consists of the forces exerted on a (small) volume of fluid. The first term describes surface forces like viscous friction and pressure, the second term describes body forces like gravity. The symmetric 2×2-matrix

$$\Sigma = \begin{pmatrix} \boldsymbol{s}_x^T \\ \boldsymbol{s}_y^T \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & \tau_{xy} \\ \tau_{xy} & \sigma_{yy} \end{pmatrix} \tag{2.4.9}$$

is called the *stress tensor*.

The form of the stress tensor depends on the fluid. A *Newtonian fluid* has a stress tensor of the form:

$$\sigma_{xx} = -p + 2\mu\frac{\partial u}{\partial x}, \tag{2.4.10a}$$

$$\sigma_{yy} = -p + 2\mu\frac{\partial v}{\partial y}, \tag{2.4.10b}$$

$$\tau_{xy} = \mu(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}), \tag{2.4.10c}$$

in which p is the pressure and μ the dynamic viscosity. The minimum configuration to be of practical importance requires a mass conservation equation in addition to (2.4.4):

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \boldsymbol{u}) = 0, \tag{2.4.11}$$

and a functional relation between ρ and p like for instance *Boyle's law.*

An important special case is where ρ is constant and Equation (2.4.11) changes into the *incompressibility condition*

$$\operatorname{div} \boldsymbol{u} = 0. \tag{2.4.12}$$

In this case ρ can be scaled out of Equation (2.4.4) and together with (2.4.10) and (2.4.12) we obtain

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + \frac{\partial \bar{p}}{\partial x} = \nu \Delta u + b_x, \tag{2.4.13a}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + \frac{\partial \bar{p}}{\partial y} = \nu \Delta v + b_y, \tag{2.4.13b}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{2.4.13c}$$

with $\nu = \frac{\mu}{\rho}$ the kinematic viscosity and $\bar{p} = \frac{p}{\rho}$ the kinematic pressure. In this case $\bar{p}$ is determined by the equations.

Exercise 2.4.1 *Derive Equation* (2.4.13). □

2.4.3.1 Boundary conditions

On each boundary *two* boundary conditions are needed, one in the normal direction and one in the tangential direction. This can be either the velocity or the stress. The tangential stress is computed by $(\boldsymbol{t}, \Sigma \boldsymbol{n})$ for given unit tangent vector, $\boldsymbol{t}$, and unit normal vector, $\boldsymbol{n}$. For reasons that go beyond the scope of this book, no boundary conditions for the pressure are required. For an extensive treatment of the Navier-Stokes equations we refer to [14] and [6].

2.4.4 Plane stress

Consider the flat plate in Figure 2.2.

The plate is fixed along side ABC but forces are applied along the free boundary ADC as a consequence of which the plate deforms in the x-y-plane. We are interested in the stresses $\Sigma = \begin{pmatrix} \sigma_{xx} & \tau_{xy} \\ \tau_{xy} & \sigma_{yy} \end{pmatrix}$ and the *displacements* $\boldsymbol{u} = \begin{pmatrix} u \\ v \end{pmatrix}$. The differential equations for the stresses (compare also (2.4.4)) are

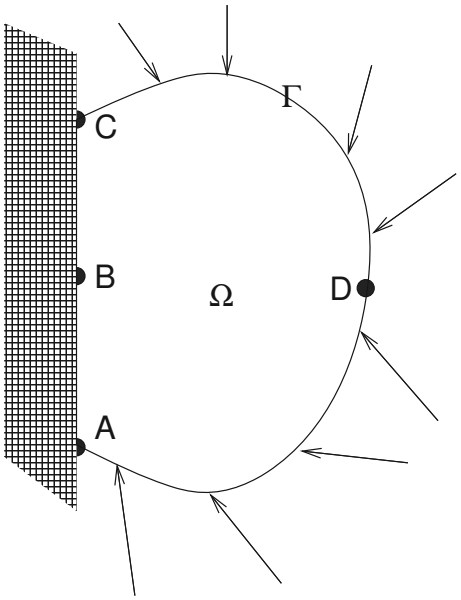

Figure 2.2: Fixed plate with forces applied along the boundary.

given by

$$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + b_1 = 0, \tag{2.4.14a}$$

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + b_2 = 0, \tag{2.4.14b}$$

in which $\boldsymbol{b}$ is the (given) body force per unit volume. Usually only gravity contributes to the body force term. We transform Equations (2.4.14) in two stages into a set of PDEs in the displacements. If the medium is *isotropic* we have a a very simple form of *Hooke's Law* relating stresses and strains:

$$E\varepsilon_x = \sigma_{xx} - \nu\sigma_{yy}, \tag{2.4.15a}$$

$$E\varepsilon_y = -\nu\sigma_{xx} + \sigma_{yy}, \tag{2.4.15b}$$

$$E\gamma_{xy} = 2(1+\nu)\tau_{xy}. \tag{2.4.15c}$$

E, the *modulus of elasticity*, and ν, *Poisson's ratio*, are material constants. Furthermore, for infinitesimal strains, there is a relation between strain and dis-

placement:

$$\varepsilon_x = \frac{\partial u}{\partial x}, \tag{2.4.16a}$$

$$\varepsilon_y = \frac{\partial v}{\partial y}, \tag{2.4.16b}$$

$$\gamma_{xy} = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}. \tag{2.4.16c}$$

This leads to the following set of PDEs in the displacements $\boldsymbol{u}$:

$$\frac{E}{1-\nu^2}\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x} + \nu\frac{\partial v}{\partial y}\right) + \frac{E}{2(1+\nu)}\frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) = -b_1, \tag{2.4.17a}$$

$$\frac{E}{2(1+\nu)}\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) + \frac{E}{1-\nu^2}\frac{\partial}{\partial y}\left(\nu\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = -b_2. \tag{2.4.17b}$$

Exercise 2.4.2 *Derive Equations* (2.4.17) □

2.4.4.1 Boundary conditions

The boundary conditions are comparable to those of the Navier-Stokes equations. At each boundary point we need a normal and a tangential piece of data, either the displacement or the stress.

Exercise 2.4.3 *Formulate the boundary conditions along ABC.* □

Exercise 2.4.4 *Along ADC the force per unit length is given:* $\boldsymbol{f}$. *Show that*

$$\sigma_{xx}n_x + \tau_{xy}n_y = f_1, \tag{2.4.18a}$$

$$\tau_{xy}n_x + \sigma_{yy}n_y = f_2, \tag{2.4.18b}$$

and hence:

$$\frac{n_x E}{1-\nu^2}\left(\frac{\partial u}{\partial x} + \nu\frac{\partial v}{\partial y}\right) + \frac{n_y E}{2(1+\nu)}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) = f_1, \tag{2.4.19a}$$

$$\frac{n_x E}{2(1+\nu)}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) + \frac{n_y E}{1-\nu^2}\left(\nu\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = f_2. \tag{2.4.19b}$$

□

2.4.5 Biharmonic equation

The prototype of a fourth order PDE is the biharmonic equation on a bounded region $\Omega \subset \mathbb{R}^2$ with boundary Γ:

$$\Delta\Delta w = f. \tag{2.4.20}$$

It describes the vertical displacement w of a flat plate in the x-y-plane, loaded perpendicularly to that plane with force f. To this problem belong three sets of physical boundary conditions:

1. *Clamped boundary*

$$w = 0, \quad \frac{\partial w}{\partial n} = 0, \quad \boldsymbol{x} \in \Gamma. \tag{2.4.21}$$

2. *Freely supported boundary*

$$w = 0, \quad \frac{\partial^2 w}{\partial n^2} + \nu \frac{\partial^2 w}{\partial t^2} = 0, \quad \boldsymbol{x} \in \Gamma. \tag{2.4.22}$$

3. *Free boundary*

$$\frac{\partial^2 w}{\partial n^2} + \nu \frac{\partial^2 w}{\partial t^2} = 0, \quad \frac{\partial^3 w}{\partial n^3} + (2-\nu) \frac{\partial^3 w}{\partial t^3} = 0, \quad \boldsymbol{x} \in \Gamma. \tag{2.4.23}$$

$\frac{\partial}{\partial n}$ and $\frac{\partial}{\partial t}$ stand for the *normal* and *tangential* derivative, respectively. Further ν is Poisson's ratio, which depends on the material. In the biharmonic equation the natural boundary conditions contain derivatives of second order or higher, all other boundary conditions are essential.

2.5 Summary of Chapter 2

In this chapter we obtained a classification of second order PDEs into *hyperbolic*, *parabolic* and *elliptic* equations. We formulated appropriate initial and boundary conditions to guarantee a unique solution. We obtained a maximum principle for subharmonic (and superharmonic) functions and used this to prove uniqueness for elliptic equations. We looked at a few examples of partial differential equations in various fields of physics and technology.

Chapter 3

Finite difference methods

Objectives

In this chapter we shall look at the form of discretization that has been used since the days of Euler (1707-1783): finite difference methods. To grasp the essence of the method we shall first look at some one-dimensional examples. After that we consider two-dimensional problems on a *rectangle* because that is a straightforward generalization of the one-dimensional case. We take a look at the discretization of the three classical types of boundary conditions. After that we consider more general domains and the specific problems at the boundary. Finally we shall turn our attention to the solvability of the resulting discrete systems and the convergence towards the exact solution.

3.1 The cable equation

As an introduction we consider the displacement y of a cable under a vertical load (see Figure 3.1).

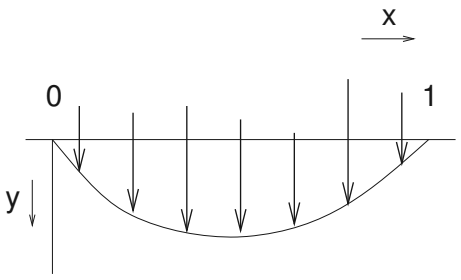

Figure 3.1: Loaded cable.

This problem is described mathematically by the second order ordinary differential equation

$$-\frac{d^2y}{dx^2} = f, \tag{3.1.1}$$

and since the cable has been fixed at both ends we have a Dirichlet boundary condition at each boundary point:

$$y(0) = 0, \quad y(1) = 0. \tag{3.1.2}$$

Note that, also here, *one* boundary condition is necessary for each point of the boundary, which just consists of two points.

3.1.1 Discretization

We divide the interval $(0,1)$ into N subintervals with length $h = 1/N$ (see Figure 3.2). We introduce the notation $x_i = ih$, $y_i = y(x_i)$ and $f_i = f(x_i)$.

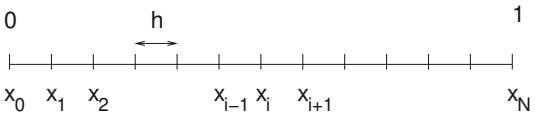

Figure 3.2: Subdivision of the interval $(0,1)$.

In the *nodes* or *nodal points* x_i we have

$$-\frac{d^2y}{dx^2}(x_i) = f_i, \tag{3.1.3}$$

and we shall try to derive an equation that connects the three variables y_{i-1}, y_i and y_{i+1} with the aid of Equation (3.1.3). We recall Taylor's formula for sufficiently smooth y:

$$y_{i+1} = y_i + h\frac{dy}{dx}(x_i) + \frac{h^2}{2!}\frac{d^2y}{dx^2}(x_i) + \frac{h^3}{3!}\frac{d^3y}{dx^3}(x_i) + \mathcal{O}(h^4), \tag{3.1.4a}$$

$$y_{i-1} = y_i - h\frac{dy}{dx}(x_i) + \frac{h^2}{2!}\frac{d^2y}{dx^2}(x_i) - \frac{h^3}{3!}\frac{d^3y}{dx^3}(x_i) + \mathcal{O}(h^4). \tag{3.1.4b}$$

When we sum Equations (3.1.4) together, the odd order terms drop out, which gives us

$$y_{i+1} + y_{i-1} = 2y_i + h^2\frac{d^2y}{dx^2}(x_i) + \mathcal{O}(h^4). \tag{3.1.5}$$

Rearranging and dividing by h^2 finally gives us the *second divided difference* approximation to the second derivative:

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = \frac{d^2y}{dx^2}(x_i) + \mathcal{O}(h^2). \tag{3.1.6}$$

The $\mathcal{O}(h^2)$ error term is called the *truncation error*, caused by truncating the Taylor series.

Exercise 3.1.1 *Show by the same method that for sufficiently smooth y the* forward divided difference $(y_{i+1} - y_i)/h$ *satisfies*

$$\frac{y_{i+1} - y_i}{h} = \frac{dy}{dx}(x_i) + \mathcal{O}(h). \tag{3.1.7}$$

Show that the backward divided difference $(y_i - y_{i-1})/h$ *satisfies*

$$\frac{y_i - y_{i-1}}{h} = \frac{dy}{dx}(x_i) + \mathcal{O}(h). \tag{3.1.8}$$

□

Exercise 3.1.2 *Show by the same method that for sufficiently smooth y the* central divided difference $(y_{i+1} - y_{i-1})/2h$ *satisfies*

$$\frac{y_{i+1} - y_{i-1}}{2h} = \frac{dy}{dx}(x_i) + \mathcal{O}(h^2). \tag{3.1.9}$$

□

Subsequently, we apply Equation (3.1.6) to *every internal node* of the interval, i.e. $x_1, x_2, \ldots, x_{N-1}$, neglecting the $\mathcal{O}(h^2)$ error term. Of course by doing so, we only get an approximation (that we denote by u_i) to the exact solution y_i. So we get

$$\begin{aligned}
h^{-2}(-u_0 + 2u_1 - u_2) &= f_1, && \text{(3.1.10a)}\\
h^{-2}(-u_1 + 2u_2 - u_3) &= f_2, && \text{(3.1.10b)}\\
\ddots\ \ddots\ \ddots &\ \ \vdots\\
h^{-2}(-u_{N-2} + 2u_{N-1} - u_N) &= f_{N-1}. && \text{(3.1.10c)}
\end{aligned}$$

Taking into account the boundary values $y(0) = y(1) = 0$ we find that $u_0 = u_N = 0$. These values are substituted into Equations (3.1.10a) and (3.1.10c) respectively. Hence the system becomes

$$\begin{aligned}
h^{-2}(2u_1 - u_2) &= f_1, && \text{(3.1.11a)}\\
h^{-2}(-u_1 + 2u_2 - u_3) &= f_2, && \text{(3.1.11b)}\\
\ddots\ \ddots\ \ddots &\ \ \vdots\\
h^{-2}(-u_{N-2} + 2u_{N-1}) &= f_{N-1}. && \text{(3.1.11c)}
\end{aligned}$$

Or in matrix-vector notation:

$$A\boldsymbol{u} = \boldsymbol{f}, \tag{3.1.12}$$

with A an $(N-1)\times(N-1)$ matrix:

$$A = h^{-2}\begin{pmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 2 \end{pmatrix}. \tag{3.1.13}$$

Exercise 3.1.3 *Show that in case of the non-homogeneous Dirichlet boundary conditions $y(0) = a$, $y(1) = b$, the matrix A is given by (3.1.13) and that the first and last element of the right-hand side $\boldsymbol{f}$ are given by $f_1 + h^{-2}a$ respectively $f_{N-1} + h^{-2}b$.* □

The solution of this system can be found by *LU-decomposition*. Since the matrix A is symmetric positive definite, also *Cholesky decomposition* (see [9]) can be used. The proof of positive definiteness will be given in the next section.

3.1.2 Properties of the discretization matrix A

From Expression (3.1.13) it is clear that the matrix A is symmetric. It is easy to prove that the $(N-1)\times(N-1)$ matrix A is positive semi-definite.

Exercise 3.1.4 *Show that matrix A is positive semi-definite.*
Hint: Use Gershgorin's theorem 1.4.5. □

There are several methods to prove that the matrix A is positive definite. The first one is by showing that the inner product $(A\boldsymbol{x}, \boldsymbol{x})$ can be written as a sum of squares.

Exercise 3.1.5 *Show that*

$$h^2(A\boldsymbol{x}, \boldsymbol{x}) = x_1^2 + \sum_{k=1}^{N-2}(x_{k+1} - x_k)^2 + x_{N-1}^2. \tag{3.1.14}$$

Derive from this result that A is positive definite. □

Another method is to estimate the eigenvalues of the matrix. In this simple case it is possible to compute the eigenvalues of the matrix A, which is more accurate than using the bounds that follow from Gershgorin's theorem.
To that end we consider the eigenvalue problem corresponding to the Laplace equation,

$$-\frac{d^2\varphi}{dx^2} = \lambda\varphi, \qquad \varphi(0) = \varphi(1) = 0, \tag{3.1.15}$$

which is a special case of the set of Sturm-Liouville problems.

Theorem 3.1.1 *The eigenvalues of Equation (3.1.15) form an infinite set given by $\lambda = k^2\pi^2$ with k any positive integer. Hence the smallest eigenvalue is exactly π^2.*

Proof
If we disregard the boundary conditions $\varphi(0) = \varphi(1) = 0$, any function of the form $\varphi(x) = e^{i\mu x}$ with $\mu \in \mathbb{R}$ is an eigenfunction with corresponding eigenvalue $\lambda = \mu^2$.
Taking the boundary conditions into account and requiring the eigenfunctions to be real we arrive at the eigenvalues λ_k and eigenfunctions φ_k defined by

$$\lambda_k = k^2\pi^2, \quad \varphi_k(x) = \sin(k\pi x) = \mathrm{Im}(e^{ik\pi x}), \quad k = 1, 2, \ldots \tag{3.1.16}$$

□

For the discrete problem (3.1.12), the corresponding eigenvalue problem $A\boldsymbol{v} = \lambda\boldsymbol{v}$ can be solved using a discrete version of the above harmonic $\varphi(x) = e^{i\mu x}$. If we extend the vector $\boldsymbol{v}$ to a grid function with values v_j for all $j \in \mathbb{Z}$ and disregard the boundary conditions

$$v_0 = v_N = 0, \tag{3.1.17}$$

then any vector of the form $v_j = e^{i\mu jh}$ with $\mu \in \mathbb{R}$ is a solution to the discrete eigenvalue problem

$$\frac{1}{h^2}(-v_{j-1} + 2v_j - v_{j+1}) = \lambda v_j, \quad j \in \mathbb{Z} \tag{3.1.18}$$

with corresponding eigenvalue

$$\lambda = \frac{1}{h^2}(2 - e^{-i\mu h} - e^{i\mu h}) = \frac{2}{h^2}(1 - \cos(\mu h)) = \frac{4}{h^2}\sin^2(\mu h/2). \tag{3.1.19}$$

Taking the boundary conditions (3.1.17) into account, and requiring the values v_j to be real, we arrive at the eigenvalues λ_k and eigenvectors $\boldsymbol{v}_k$ defined by

$$\begin{aligned} \lambda_k &= \frac{4}{h^2}\sin^2(k\pi h/2), && k = 1, 2, \ldots, N-1, \\ (\boldsymbol{v}_k)_j &= \sin(k\pi jh) = \mathrm{Im}(e^{ik\pi jh}), && k, j = 1, 2, \ldots, N-1. \end{aligned} \tag{3.1.20}$$

Exercise 3.1.6 *Use* (3.1.20) *to show that the smallest eigenvalue of the symmetric matrix A is approximately π^2.* □

Since the smallest eigenvalue of the symmetric matrix A is positive, it follows from Corollary 1.4.3 that A is positive definite.

The above method is only applicable for simple cases with constant coefficients, like the one treated here.

3.1.3 Global error

We will estimate the order of the error in our approximate solution u. From Equation (3.1.6) we know that each of the equations of the set (3.1.11) contains a truncation error of order $\mathcal{O}(h^2)$, provided that y is sufficiently smooth. Suppose that this error in the k-th equation, E_k, is given by $E_k = h^2 p_k$. We know that p_k remains bounded as $h \to 0$ by the definition of $\mathcal{O}$. Now let $\Delta y_k = y_k - u_k$, where y_k is the exact solution and u_k our numerical approximation. Then

$$A\boldsymbol{y} = \boldsymbol{f} + h^2\boldsymbol{p}, \tag{3.1.21}$$

and

$$A\boldsymbol{u} = \boldsymbol{f}. \tag{3.1.22}$$

We subtract (3.1.22) from (3.1.21) to obtain a set of equations for the global error $\Delta\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{u}$:

$$A\Delta\boldsymbol{y} = h^2\boldsymbol{p}. \tag{3.1.23}$$

We shall show that the global error is of order $\mathcal{O}(h^2)$ when measured in the scaled L_2-norm defined by $\|\boldsymbol{x}\|_{2,h} = \sqrt{h}\|\boldsymbol{x}\|_2$.

Theorem 3.1.2 *The discretization of the Poisson equation (3.1.1) with boundary conditions (3.1.2) by Equation (3.1.6) gives a global error $\Delta\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{u}$ satisfying $\|\Delta\boldsymbol{y}\|_{2,h} = \mathcal{O}(h^2)$.* □

Proof
From Equation (3.1.23) we obtain $\Delta\boldsymbol{y} = h^2 A^{-1}\boldsymbol{p}$, which implies (cf. Exercise 1.4.4)

$$\|\Delta\boldsymbol{y}\|_2 \leq h^2\|A^{-1}\|_2\|\boldsymbol{p}\|_2. \tag{3.1.24}$$

Since it follows from Theorem 1.4.4 that for a real symmetric positive definite matrix A, the induced matrix norm $\|A^{-1}\|_2$ is equal to the reciprocal of its smallest eigenvalue, λ_1, we get

$$\|\Delta\boldsymbol{y}\|_2 \leq \frac{h^2}{\lambda_1}\|\boldsymbol{p}\|_2 \approx \frac{h^2}{\pi^2}\|\boldsymbol{p}\|_2. \tag{3.1.25}$$

Multiplication by $\sqrt{h}$ and using that $\|\boldsymbol{p}\|_{2,h}$ remains bounded gives the required result. □

In this special case it is also possible to estimate the error in the maximum norm as will be shown in the following theorem.

Theorem 3.1.3 *Let $E_k = h^2 p_k$ denote the truncation errors defined above, and let $\boldsymbol{p}$ be the vector with components p_k. Then the global error $\Delta\boldsymbol{y}$ of Theorem 3.1.2 satisfies*

$$\|\Delta\boldsymbol{y}\|_\infty \leq \frac{h^2}{8}\|\boldsymbol{p}\|_\infty.$$

□

The above theorem will be proved in the remainder of this section.

Exercise 3.1.7 *Let $\boldsymbol{e}$ be the vector with components $e_k = 1, k = 1, 2, \ldots, N-1$. Show by substitution that the solution $\boldsymbol{v}$ of the set of equations $A\boldsymbol{v} = \boldsymbol{e}$ has components $v_k = \frac{1}{2}h^2(N-k)k, k = 1, 2, \ldots, N-1$. Show that this implies $\|\boldsymbol{v}\|_\infty \leq 1/8$. (Hint: Recall the definition* (1.4.14) *of the maximum norm and use $Nh = 1$.)* □

In Chapter 2, we saw that the smooth solutions of Laplace's equation satisfy a maximum principle. This should also hold for the numerical solution, which is obtained after the discretization. The following theorem represents the discrete version of the maximum principle. The vector inequality $\boldsymbol{y} \geq \boldsymbol{x}$ means that the inequality is valid for *every* component.

Theorem 3.1.4 *(Discrete Maximum Principle)*
Let A be the discretization matrix defined in (3.1.13). *Then $A\boldsymbol{u} \geq \boldsymbol{0}$ implies $\boldsymbol{u} \geq \boldsymbol{0}$.*
□

Exercise 3.1.8 *Prove Theorem 3.1.4. Reason by contradiction and assume that $\boldsymbol{u}$ has a negative minimum for some component u_k. Now consider the k-th equation and show that this is impossible.* □

The next important property is the existence and uniqueness of a numerical solution. This is formulated in the following theorem:

Theorem 3.1.5 *(Existence and uniqueness)*

1. *Let A be the matrix defined in* (3.1.13). *Then $A\boldsymbol{u} = \boldsymbol{0}$ implies $\boldsymbol{u} = \boldsymbol{0}$.*

2. *The set of equations $A\boldsymbol{u} = \boldsymbol{f}$ has a unique solution for every $\boldsymbol{f}$.*

Exercise 3.1.9 *Prove Theorem 3.1.5. First use Theorem 3.1.4 to prove assertion 1, and then use assertion 1 to prove assertion 2.* □

Exercise 3.1.10 *With the definitions as in Exercise 3.1.7, show that*

$$-h^2\|\boldsymbol{p}\|_\infty \boldsymbol{v} \leq \Delta\boldsymbol{y} \leq h^2\|\boldsymbol{p}\|_\infty \boldsymbol{v}. \tag{3.1.26}$$

Show that therefore

$$\|\Delta\boldsymbol{y}\|_\infty \leq \frac{h^2}{8}\|\boldsymbol{p}\|_\infty. \tag{3.1.27}$$

Hint: use Theorem 3.1.4. □

This concludes the proof of Theorem 3.1.3.

3.2 Some simple extensions of the cable equation

The Poisson equation (3.1.1) is a special case of the diffusion equation

$$-\frac{d}{dx}(\kappa(x)\frac{d\varphi}{dx}) = f, \tag{3.2.1}$$

with boundary conditions

$$\varphi(0) = a, \quad \varphi(1) = b, \tag{3.2.2}$$

and $\kappa(x)$ a positive function of x.

3.2.1 Discretization of the diffusion equation

There are several possibilities to discretize Equation (3.2.1) with an accuracy of $\mathcal{O}(h^2)$. The first one is to rewrite Equation (3.2.1) as

$$-\kappa(x)\frac{d^2\varphi}{dx^2} - \frac{d\kappa(x)}{dx}\frac{d\varphi}{dx} = f. \tag{3.2.3}$$

However, if we apply central differences to discretize (3.2.3), the symmetry that is inherent to Equation (3.2.1) is lost.
One could use Taylor expansion to derive a $\mathcal{O}(h^2)$ symmetric discretization of (3.2.1). Unfortunately, such an approach is quite complicated.

A better method is to use the central divided differences of Exercise 3.1.2 repeatedly. Define

$$y(x) = \kappa(x)\frac{d\varphi}{dx} \tag{3.2.4}$$

and use central differences based on the midpoints $x_{i-\frac{1}{2}}$, $x_{i+\frac{1}{2}}$ (see Figure 3.3).

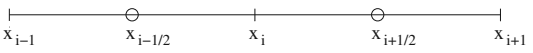

Figure 3.3: Position of discretization points.

This leads to

$$\frac{y_{i+\frac{1}{2}} - y_{i-\frac{1}{2}}}{h} = \frac{dy}{dx} + \mathcal{O}(h^2). \tag{3.2.5}$$

Substitution of (3.2.4) into (3.2.5) gives

$$-\frac{\kappa(x_{i+\frac{1}{2}})\frac{d\varphi}{dx}(x_{i+\frac{1}{2}}) - \kappa(x_{i-\frac{1}{2}})\frac{d\varphi}{dx}(x_{i-\frac{1}{2}})}{h} = -\frac{d}{dx}(\kappa(x)\frac{d\varphi}{dx}) + \mathcal{O}(h^2). \tag{3.2.6}$$

Next use central differences to discretize $\frac{d\varphi}{dx}$ to get the final expression

$$-\kappa(x_{i+\frac{1}{2}})\frac{\varphi_{i+1} - \varphi_i}{h^2} + \kappa(x_{i-\frac{1}{2}})\frac{\varphi_i - \varphi_{i-1}}{h^2} = f_i. \tag{3.2.7}$$

Exercise 3.2.1 *Use Taylor series expansion to prove that*

$$\kappa(x_{i+\frac{1}{2}}) = \kappa + \frac{h}{2}\kappa' + \frac{h^2}{8}\kappa'' + \mathcal{O}(h^3). \tag{3.2.8}$$

Derive a similar expression for $\kappa(x_{i-\frac{1}{2}})$.
Use Taylor series expansion to prove that

$$-\frac{1}{h}\left[\kappa(x_{i+\frac{1}{2}})\frac{\varphi_{i+1}-\varphi_i}{h} - \kappa(x_{i-\frac{1}{2}})\frac{\varphi_i-\varphi_{i-1}}{h}\right] = -\frac{d}{dx}\left[\kappa(x_i)\frac{d\phi}{dx}(x_i)\right] + \mathcal{O}(h^2). \tag{3.2.9}$$

Hint: Use Equation (3.2.3). □

This discretization matrix is clearly symmetric and one can prove that it is also positive definite. Hence the original properties of Equation (3.2.1) are kept.

3.2.2 Boundary conditions

The treatment of Dirichlet boundary conditions is trivial as shown in the previous section. In case the boundary condition contains derivatives, getting an $\mathcal{O}(h^2)$ accuracy requires a thorough discretization.
Consider the Poisson equation (3.1.1) with boundary conditions

$$y(0) = a, \qquad \frac{dy}{dx}(1) = c. \tag{3.2.10}$$

If we use the subdivision of Figure 3.2, then the value of y_N is unknown. Since the discretization (3.1.6) is only applicable to internal points (why?), we need an extra equation to get a square matrix. The most simple method is to use a backward difference to discretize the Neumann boundary condition. This introduces an extra equation, but the truncation error is only $\mathcal{O}(h)$ according to Exercise 3.1.1. A better method is to introduce an extra *virtual point*, x_{N+1}, outside the domain. This implies that the discretization (3.1.6) can be extended to node x_N. The Neumann boundary condition in $x = 1$ can be discretized by central differences. So $y(x_{N+1})$ can be expressed into $y(x_N)$ and $y(x_{N-1})$, and this can be substituted in the discretization of the differential equation in $x = 1$. In fact the virtual point is eliminated in this way. The error in each of the steps is $\mathcal{O}(h^2)$, but unfortunately the symmetry of the matrix is lost. Another option is to let the boundary $x = 1$ be in the middle of the interval (x_{N-1}, x_N) as in Figure 3.4. If we omit the truncation error, Equation (3.1.6) for $i = N - 1$ leads to

$$-\frac{y_{N-2} - 2y_{N-1} + y_N}{h^2} = f_{N-1}. \tag{3.2.11}$$

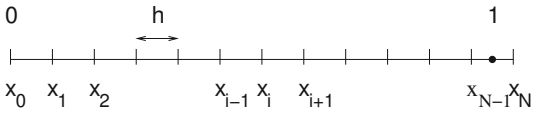

Figure 3.4: Subdivision with virtual point.

Central difference discretization of $\frac{dy}{dx}(1) = c$ gives

$$\frac{y_N - y_{N-1}}{h} = c, \tag{3.2.12}$$

and substitution of (3.2.12) in (3.2.11) results in

$$\frac{-y_{N-2} + y_{N-1}}{h^2} = f_{N-1} + \frac{c}{h}. \tag{3.2.13}$$

Remark
A simpler way to get a symmetric matrix would be to use the original matrix and to divide the last row of matrix and right-hand side by 2. However, such an approach is only applicable for constant coefficients.

Although in each step of the derivation $\mathcal{O}(h^2)$ approximations are used, still the local truncation error of Equation (3.2.13) is $\mathcal{O}(h)$, see Exercise 3.2.2.

Exercise 3.2.2 *Show that the Taylor series expansion around x_{N-1} of the left-hand side of Equation (3.2.13) can be written as*

$$\frac{y'}{h} - \frac{y''}{2} + \frac{h}{6}y''' + \mathcal{O}(h^2), \tag{3.2.14}$$

where $y = y(x_{N-1}) = y(1 - \frac{h}{2})$.
Show, using a Taylor series around x_{N-1}, that the first derivative of $y(x)$ in point $x = 1$ can be written as

$$y'(1) = y' + \frac{h}{2}y'' + \frac{h^2}{8}y''' + \mathcal{O}(h^3). \tag{3.2.15}$$

Show by substitution of (3.2.15) in (3.2.14) and the boundary condition (3.2.10) that the local truncation error of (3.2.13) is $\mathcal{O}(h)$. □

It is rather disappointing that the local truncation error is $\mathcal{O}(h)$, despite the fact that we used $\mathcal{O}(h^2)$ approximations in each step. Fortunately it is possible to prove that the global error is still $\mathcal{O}(h^2)$. For that purpose we write the truncation error for the complete system as $h^2\boldsymbol{p} + h\boldsymbol{q}$, where $\boldsymbol{p}$ is defined as in (3.1.21) and $\boldsymbol{q}$ is a vector that is completely zero except for the last component which is equal to q_{N-1}, so

$$\boldsymbol{q} = (0, 0, ..., 0, q_{N-1})^T. \tag{3.2.16}$$

The global error $\Delta \boldsymbol{y}$ can be split into $\Delta \boldsymbol{y} = \Delta \boldsymbol{y}_1 + \Delta \boldsymbol{y}_2$, with

$$A\Delta \boldsymbol{y}_1 = h^2 \boldsymbol{p}, \tag{3.2.17}$$
$$A\Delta \boldsymbol{y}_2 = h\boldsymbol{q}. \tag{3.2.18}$$

Exercise 3.2.3 *Prove analogous to Section 3.1.2 that the smallest eigenvalue of the matrix A is approximately* $\pi^2/4$.

Exercise 3.2.4 *Show the matrix A satisfies a discrete maximum principle similar to that in Theorem 3.1.4.*

From Exercises 3.2.3 and 3.2.4 it follows that $\|\Delta \boldsymbol{y}_1\| = \mathcal{O}(h^2)$ both in the maximum norm and the scaled L_2-norm. The exact solution of (3.2.18) is $(\Delta \boldsymbol{y}_2)_i = h^2 q_{N-1} x_i$, hence the global error $\|\Delta \boldsymbol{y}\|$ is also $\mathcal{O}(h^2)$.

Exercise 3.2.5 *Show that* $\varphi(x) = hq_{N-1}x$ *is the solution of*

$$-\frac{d^2\varphi}{dx^2} = 0, \quad \varphi(0) = 0, \quad \frac{d\varphi}{dx}(1) = hq_{N-1}.$$

Deduce from this result that $(\Delta \boldsymbol{y}_2)_i = h^2 q_{N-1} x_i$, *and hence* $\|\Delta \boldsymbol{y}_2\| \leq |q_{N-1}|h^2$. □

Periodic boundary conditions require a slightly different approach. This type of boundary condition is for example used in case the solution repeats itself endlessly. Consider for example the Poisson equation

$$-\frac{d^2u}{dx^2} = f(x), \quad x \in [0,1], \tag{3.2.19}$$

where $u(x)$ and $f(x)$ are periodic functions with period 1.
Periodicity implies

$$u(x) = u(x+L) \tag{3.2.20}$$

with L the length of the interval. Therefore the trivial boundary condition is

$$u(0) = u(1). \tag{3.2.21}$$

However, since a second order elliptic equation requires a boundary condition for the whole boundary, two boundary conditions are needed. The second boundary condition one can use is

$$\frac{du}{dx}(0) = \frac{du}{dx}(1). \tag{3.2.22}$$

Exercise 3.2.6 *Derive* (3.2.22). *Hint: use* (3.2.20). □

To discretize Equation (3.2.19) we use the grid of Figure 3.2. The discretization of the differential equation is standard. The discretization of the boundary condition (3.2.21) is trivial. It is sufficient to identify the unknowns u_0 and

u_N and represent them by one unknown only (say u_N). To discretize boundary condition (3.2.22) one could use divided differences for both terms in the equation. A more natural way of dealing with this boundary condition is to use the periodicity explicitly by discretizing the differential equation (3.2.19) in $x = 1$ and using the fact that the next point is actually x_1. So we use condition (3.2.20). Hence

$$\frac{-u_{N-1} + 2u_N - u_1}{h^2} = f_N. \tag{3.2.23}$$

Exercise 3.2.7 *Why is it sufficient to apply (3.2.23) only for $x = 1$ (and not for $x = 0$)?* □

Exercise 3.2.8 *Show that the discretization of (3.2.19) using (3.2.23) gives the following system of equations:*

$$h^{-2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & -1 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -1 & 2 & -1 \\ -1 & 0 & \dots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix}.$$

3.3 Singularly perturbed problems

Singularly perturbed problems occur when the coefficient of the highest order derivative is very small compared to the other coefficients. A common example is the *convection-diffusion* equation

$$-\varepsilon \frac{d^2c}{dx^2} + v\frac{dc}{dx} = 0, \qquad c(0) = 0, \quad c(L) = 1, \tag{3.3.1}$$

which describes the transport of a pollutant with concentration c by a convecting medium with known velocity v.

3.3.1 Analytical solution

For constant velocity v and diffusion coefficient ε there is a solution in closed form:

$$c(x) = \frac{\exp(vx/\varepsilon) - 1}{\exp(vL/\varepsilon) - 1}. \tag{3.3.2}$$

For $L = 1$ and $vL/\varepsilon = 40$ the solution has been plotted in Figure 3.5.
The dimensionless quantity vL/ε that occurs regularly in convection diffusion problems is called the *Péclet number Pe*. It is a measure for by how much the convection dominates the diffusion. Note that if the Péclet number is large

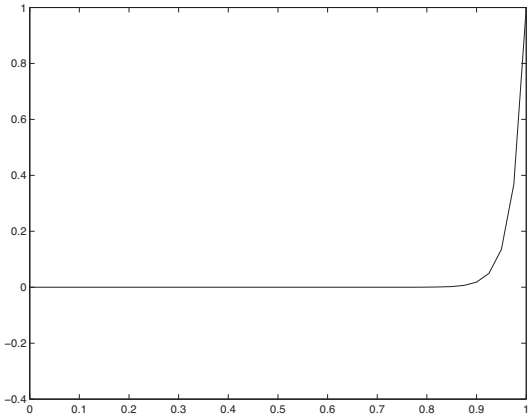

Figure 3.5: Analytic solution.

(say $|Pe| > 10$), there is a boundary layer at $x = L$: the right-hand side boundary condition makes itself felt only very close to the boundary. This boundary layer will cause problems in the numerical treatment.

3.3.2 Numerical approximation

Let us take central differences for the first derivative to provide us with an $\mathcal{O}(h^2)$ consistent scheme. This gives us a set of equations

$$A\boldsymbol{c} = \boldsymbol{f}, \tag{3.3.3}$$

where A and $\boldsymbol{f}$ are given by

$$A = h^{-2}\begin{pmatrix} 2 & -1+p_h & 0 & \dots & \dots & 0 \\ -1-p_h & 2 & -1+p_h & 0 & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -1-p_h & 2 & -1+p_h \\ 0 & \dots & \dots & 0 & -1-p_h & 2 \end{pmatrix},$$

$$\boldsymbol{f} = \frac{1}{h^2}\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1-p_h \end{pmatrix},$$

in which $p_h = \frac{vh}{2\varepsilon}$ is the so-called *mesh Péclet number*.

Exercise 3.3.1 *Derive the above expressions for the matrix A and vector* $\boldsymbol{f}$ *in Equation (3.3.3).* □

In Figures 3.6 and 3.7 you see the numerical solution for $Pe = 40$ and $h = 0.1$ and $h = 0.025$ respectively. In Figure 3.6 we observe wiggles and negative concentrations. These oscillations are unacceptable from a physical point of view. The wiggles have disappeared in Figure 3.7.

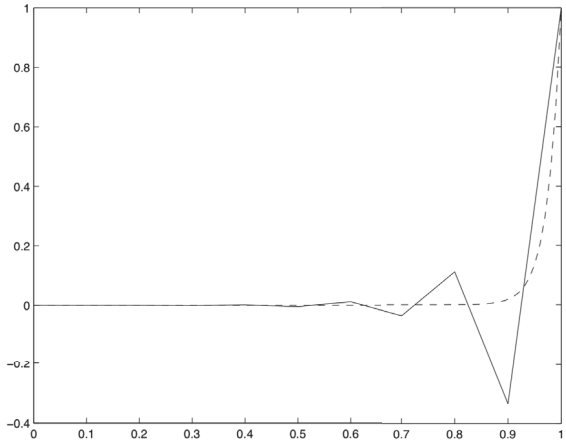

Figure 3.6: Numerical (solid) and exact (dotted) solution, coarse grid.

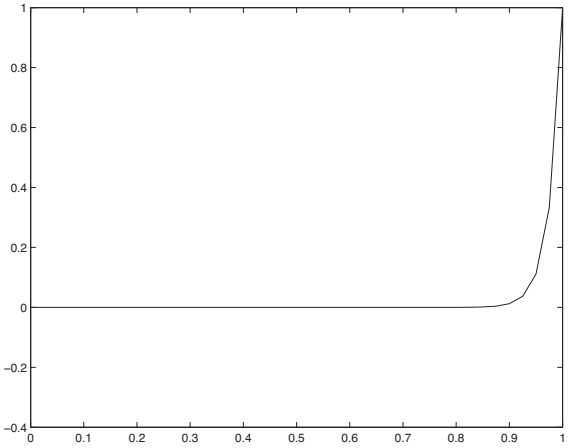

Figure 3.7: Numerical solution, fine grid.

3.3.2.1 Explanation

To explain this phenomenon we consider the following set of *linear difference equations:*

$$bu_{k-1} - (b+a)u_k + au_{k+1} = 0, \quad u_0 = 0, \quad u_n = 1. \tag{3.3.4}$$

This system can be solved by substituting $u = r^k$. From Equation (3.3.4) it follows that

$$b - (b+a)r + ar^2 = 0, \tag{3.3.5}$$

with solutions $r = 1$ and $r = b/a$. The general solution of (3.3.4) can now be written as

$$u_k = A + B\left(\frac{b}{a}\right)^k. \tag{3.3.6}$$

After application of the boundary conditions we find

$$u_k = \frac{\left(\frac{b}{a}\right)^k - 1}{\left(\frac{b}{a}\right)^n - 1}. \tag{3.3.7}$$

Apparently it is necessary that $\frac{b}{a} \geq 0$ to have a monotone, increasing solution.

3.3.2.2 Upwind differencing

For the mesh Péclet number p_h we need the condition $|p_h| \leq 1$ to have a monotone solution. This follows directly from the result of the previous section. To satisfy this inequality we need a condition on the stepsize h: apparently we must have $\frac{h}{L} \leq \frac{2}{|Pe|}$. This condition may lead to unrealistically small stepsizes, because in practice Pe can be as large as 10^6. To overcome this you often see the use of *backward* differences for $v > 0$ and *forward* differences for $v < 0$. This is called *upwind differencing*.

Exercise 3.3.2 *Show that taking a backward difference leads to a three-term recurrence relation of the form:*

$$(-1 - 2p_h)u_{k-1} + (2 + 2p_h)u_k - u_{k+1} = 0. \tag{3.3.8}$$

Show that this recurrence relation has a monotone solution if $p_h > 0$. □

Exercise 3.3.3 *Give the three-term recurrence relation for $v < 0$. Show that this also has a monotone solution.* □

Upwind differencing has a big disadvantage: the accuracy of the solution drops an order and in fact you're having the worst of two worlds: your approximation is bad and you will not be warned that this is the case. See Figure 3.8.

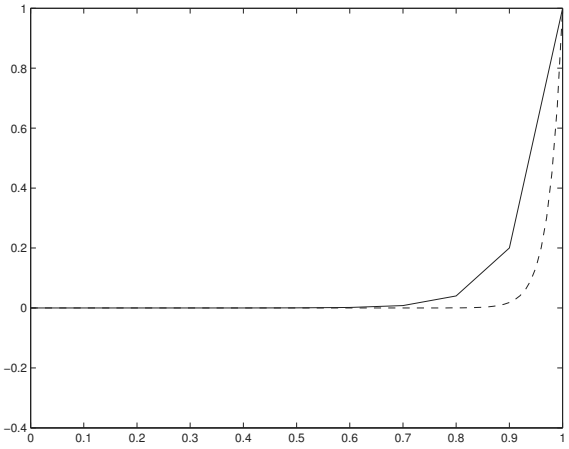

Figure 3.8: Upwind (solid) and exact (dotted) solution.

Why is this approximation so bad? The first order approximation of the first order derivative introduces an artificial diffusion term to suppress the wiggles. This artificial diffusion is an order of magnitude larger than the physical diffusion. So in fact you solve a different problem. See Exercise 3.3.4.

Exercise 3.3.4 *Show that*

$$\frac{c_k - c_{k-1}}{h} = c'_k - \frac{h}{2}c''_k + \mathcal{O}(h^2). \tag{3.3.9}$$

Show that this approximation reduces the Péclet number to

$$\widehat{Pe} = \frac{Pe}{1 + p_h}. \tag{3.3.10}$$

Deduce from this that $\hat{p}_h < 1$ *for* $v > 0$*. Give analogous relations for* $v < 0$ *and explain why it is necessary to take a forward difference in this case.* □

Effectively, using upwind differencing, you are approximating the solution of

$$-(\varepsilon + \frac{vh}{2})\frac{d^2c}{dx^2} + v\frac{dc}{dx} = 0. \tag{3.3.11}$$

It is clear that for a good accuracy $\frac{vh}{2}$ must be small compared to ε. Hence upwind differencing produces nice pictures, but if you need an accurate solution, then, central differences with small h are preferred.

A better way to handle the boundary layer is *mesh refinement* in the boundary layer itself. The boundary layer contains large gradients and to resolve these you need a sufficient number of points. Actual practice shows that taking sufficient points in the boundary layer suppresses the wiggles. In Figure 3.9 the

solution is calculated with 10 points only, but at nodes 0.5, 0.8, 0.85, 0.88, 0.91, 0.93, 0.95, 0.97, 0.99 and 1.

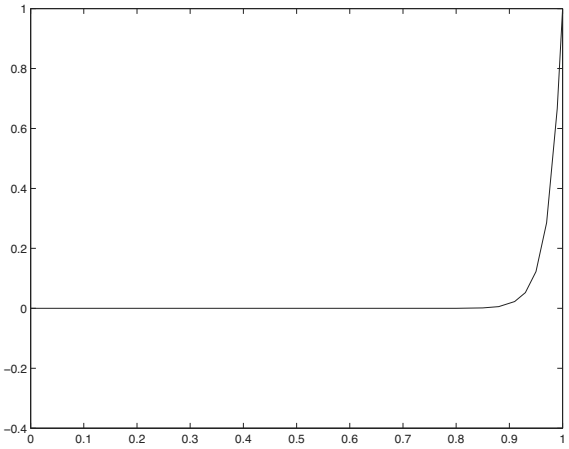

Figure 3.9: Non-equidistant nodal points.

In favor of the upwind differencing method it has to be said that it is the only course of action available in the neighborhood of shocks. As a result you often see methods with a higher accuracy in smooth regions of the solution that fall back on the first order upwind scheme close to shocks.

3.3.2.3 Source terms

If *source terms* in the equation suppress the boundary layer there will be no wiggles in the numerical solution, even if the matrix does not satisfy the *mesh Péclet condition* $|p_h| \leq 1$.

Exercise 3.3.5 *Calculate with central differences the numerical solution of*

$$-y'' + vy' = \pi^2 \sin \pi x + v\pi \cos \pi x, \quad y(0) = y(1) = 0. \tag{3.3.12}$$

Take $v = 40$ *and* $h = 0.1$. □

Remark
The use of the previous upwind differencing, also called *first order upwind*, may be inaccurate, but it usually produces nice pictures. This makes the method attractive from a selling point of view. In the literature more accurate higher order upwind schemes can be found. Treatment of these schemes goes beyond the scope of this textbook.

3.4 Poisson's equation on a rectangle

We now generalize our procedure to two dimensions. Consider a rectangle Ω with length L and width W. In this rectangle we consider *Poisson's equation*

$$-\Delta u = f \tag{3.4.1}$$

with *homogeneous boundary conditions* $u = 0$ on Γ.

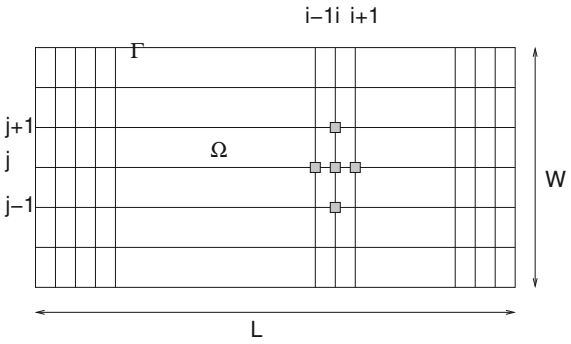

Figure 3.10: Rectangular grid with 5 point molecule.

We divide Ω into small rectangles with sides Δx and Δy such that $M\Delta x = L$ and $N\Delta y = W$. At the intersections of the grid lines we have *nodes* or *nodal points* where we shall try to find approximations of the unknown u. The unknown at node (x_i, y_j) (or (i, j) for short) we denote by $u_{i,j}$. In the same way as in Section 3.1 we replace the differential equation in this node by

$$\frac{-u_{i-1,j} + 2u_{i,j} - u_{i+1,j}}{\Delta x^2} + \frac{-u_{i,j-1} + 2u_{i,j} - u_{i,j+1}}{\Delta y^2} = f_{i,j}. \tag{3.4.2}$$

Exercise 3.4.1 *Use Taylor expansion in two variables to show that the truncation error in* (3.4.2) *is given by*

$$E_{ij} = -\frac{1}{12}\left(\Delta x^2 \frac{\partial^4 u}{\partial x^4}(x_i, y_j) + \Delta y^2 \frac{\partial^4 u}{\partial y^4}(x_i, y_j)\right). \tag{3.4.3}$$

In this expression terms of order 5 and higher in the Taylor expansion have been neglected. □

Writing down Equation (3.4.2) for every internal nodal point (i, j) with $i = 1, 2, \ldots, M-1$, $j = 1, 2, \ldots, N-1$, presents us with a set of $(M-1) \times (N-1)$ equations with just as many unknowns.

Exercise 3.4.2 *Give the equation with node (1,5) as central node. Substitute the homogeneous boundary conditions.* □

Exercise 3.4.3 *Give the equation with node* $(M-1, N-1)$ *as central node. Substitute the homogeneous boundary conditions.* □

3.4.1 Matrix vector form

Since the system we obtained is a linear system we can represent it in matrix vector form $A\boldsymbol{u} = \boldsymbol{f}$. This is not exactly a trivial task, because we have a vector of unknowns with a double index and the conventional matrix vector representation uses a single index. We shall show how to do this in a specific example, $M = 6, N = 4$. First of all we show how to convert the double index (i, j) into a single index α. As we will see below, this can be done in a number of different ways. Each of these numbering schemes will give rise to a so-called *band matrix*.

Definition 3.4.1 *A matrix* $A = (a_{ij})$ *is called a* band matrix *if all elements outside a certain band are equal to zero. In formula:* $a_{ij} = 0$ if $i - j > b_1$ or $j - i > b_2$. *The* bandwidth *of the matrix is in that case* $b_1 + b_2 + 1$.

We will now discuss three numbering schemes. They are most easily represented in a picture.

3.4.1.1 Horizontal numbering

The nodes are numbered sequentially in horizontal direction (see Figure 3.11).

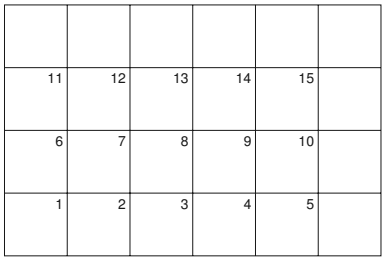

Figure 3.11: Horizontal numbering.

The conversion formula from double index (i, j) to single index α is straightforward:

$$\alpha = i + (j-1) * (M-1). \tag{3.4.4}$$

Exercise 3.4.4 *Show that* A *is a* 3×3 *block matrix in which each block is* 5×5. *What is the bandwidth of* A*?* □

The diagonal blocks are tridiagonal, the sub- and super-diagonal blocks are diagonal and all other blocks are 0.

3.4.1.2 Vertical numbering

The nodes are numbered sequentially in vertical direction (see Figure 3.12).

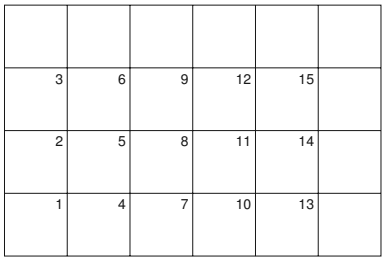

Figure 3.12: Vertical numbering.

The conversion formula from double index (i, j) to single index α is straightforward:

$$\alpha = (i-1) * (N-1) + j. \tag{3.4.5}$$

Exercise 3.4.5 *Show that A is a 5×5 block matrix in which each block is 3×3. What is the bandwidth of A?* □

The diagonal blocks are tridiagonal, the sub- and super-diagonal blocks are diagonal and all other blocks are 0.

3.4.1.3 Oblique numbering

The nodes are numbered sequentially along lines $i + j = k, k = 2, \ldots, 8$ (see Figure 3.13).

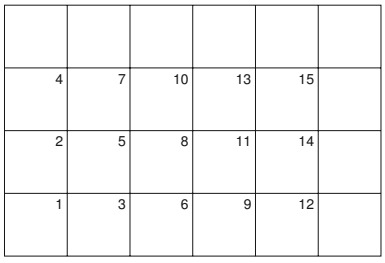

Figure 3.13: Oblique numbering.

The conversion formula from double index (i, j) to single index α is not so straightforward. A is still a block matrix, in which the diagonal blocks increase in size from 1×1 to 3×3. The diagonal blocks are diagonal, the sub- and super-diagonal blocks are bidiagonal and all other blocks are 0.

Exercise 3.4.6 *What is the bandwidth of A?* □

3.5 Boundary conditions extended

3.5.1 Natural boundary conditions

Basically *natural boundary conditions* (i.e. Neumann or Robin boundary conditions) involve a flow condition. The treatment in 2D is similar to 1D (see Section 3.2.2). Since these conditions are dealt with in a natural way by Finite Volume Methods we postpone a more detailed discussion of that subject until the next chapter.

3.5.2 Dirichlet boundary conditions on non-rectangular regions

Unfortunately on non-rectangular regions the boundary does not coincide with the grid, see Figure 3.14.

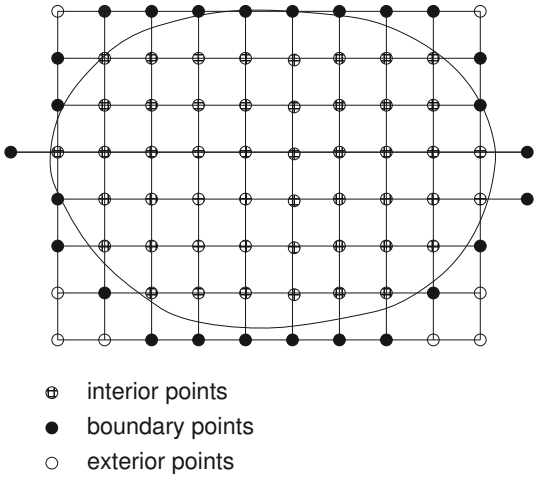

Figure 3.14: Grid on non-rectangular region.

For each interior point we have an equation involving function values in five nodes. The black points in Figure 3.14 have to be determined by the Dirichlet boundary condition. It is acceptable to express a black point in a nearby boundary value and the function values in one or more interior points (interior variables). The idea is to end up with a system of equations that only contains interior variables. In this way we can guarantee that we have as many equations as unknowns. We explain the way to proceed by an example. Consider the situation in Figure 3.15.

In this figure we have to express u_S in the known value u_B and the interior variable u_C. Let h be the distance between grid points and sh the fraction that separates the boundary from the S-point. By linear interpolation we have

$$u_B = (1-s)u_S + su_C + \mathcal{O}(h^2), \tag{3.5.1}$$

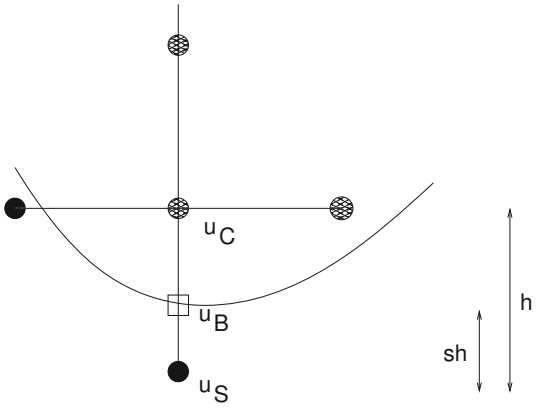

Figure 3.15: Boundary molecule, u_S is an exterior boundary point.

and that gives us the relation that we can substitute into the equation:

$$u_S = \frac{u_B - s u_C}{1 - s}. \tag{3.5.2}$$

If s is close to 1 this procedure may lead to an unbalanced set of equations. For that reason we usually consider a point that is closer than say $\frac{1}{4}h$ to the boundary as a *boundary point* even if it belongs to the interior. In that case u_S falls in between u_B and u_C (see Figure 3.16) and the formula changes correspondingly.

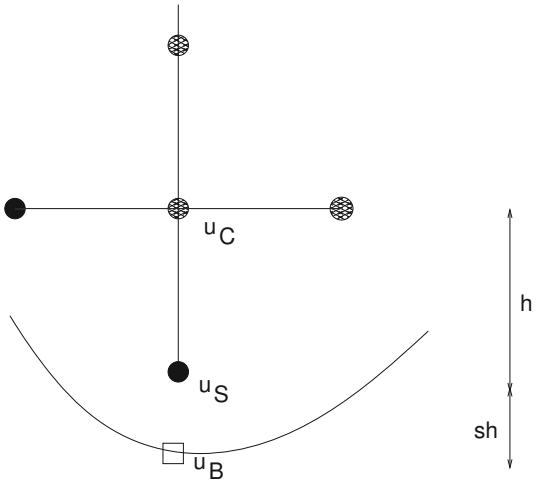

Figure 3.16: Boundary molecule, u_S is an interior boundary point.

Here we have

$$u_S = \frac{u_B + s u_C}{1 + s}. \tag{3.5.3}$$

Remark
The method treated here is quite old-fashioned. It is better to use a coordinate transformation (Section 3.7).

3.6 Global error estimate

We shall try to get some of the flavor of global error estimates for numerical solutions of Problem (3.4.1). The scaled L_2 error estimate can be derived in the same way as in Theorem 3.1.2. Here we shall concentrate ourselves on pointwise estimates. In order to do so we need to develop some properties for the discrete Laplace operator. These properties also hold in 3 dimensions, so in a certain way this is a generic treatment of the problem. We will do the estimate on a rectangle with homogeneous Dirichlet boundary conditions, but in subsequent sections we shall hint at ways to apply the theory to more general domains and boundary conditions.

3.6.1 The discrete maximum principle

If the $N \times N$ system of equations $A\boldsymbol{u} = \boldsymbol{f}$ is a Finite Difference discretization of Problem (3.4.1) with Dirichlet boundary conditions then A has the following properties:

$$a_{jk} \leq 0, \quad \text{if } j \neq k, \tag{3.6.1a}$$

$$a_{kk} > 0, \quad \text{for all } k, \tag{3.6.1b}$$

$$|a_{kk}| \geq \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{kj}|, \quad k = 1, \ldots, N. \tag{3.6.1c}$$

We first formulate some definitions and after that an important theorem.

Definition 3.6.1 *A is a* Z-matrix *if*

$$a_{ij} \leq 0, \quad \textit{for all } i, j \textit{ with } i \neq j.$$

Definition 3.6.2 *A is an* L-matrix *if A is a Z-matrix and*

$$a_{ii} > 0, \quad \textit{for all } i.$$

Definition 3.6.3 *A is* diagonally dominant *if*

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{ij}|, \quad i = 1, 2, \ldots, N. \tag{3.6.2}$$

Definition 3.6.4 *A is* reducible *if there exists a permutation matrix P such that*

$$P^T AP = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \tag{3.6.3}$$

where A_{11} and A_{22} are square matrices of order less than N. If no such P exists, then A is called irreducible.

Exercise 3.6.1 *For a given $N \times N$ matrix $A = (a_{ij})$ we say that one can "step" from row i to row j if and only if $a_{ij} \neq 0$.*
Show that the matrix A is irreducible if and only if you can "walk" (in one or more steps) from any row i to any other row j. □

In most practical cases our discretization matrices are irreducible.

Definition 3.6.5 *A is* irreducibly diagonally dominant *if it is irreducible, diagonally dominant, and we have strict inequality in (3.6.2) for at least one row index i.*

The following theorem is a generalization of Theorems 3.1.4 and 3.1.5.

Theorem 3.6.1 *(Discrete Maximum Principle)*
If A is an irreducibly diagonally dominant L-matrix, then

(i) $A\boldsymbol{u} \geq \mathbf{0} \Rightarrow \boldsymbol{u} \geq \mathbf{0}$.

(ii) A is non-singular.

Proof
First we prove (i). Suppose $A\boldsymbol{u} \geq 0$ but that $\boldsymbol{u} \geq 0$ does not hold. Then we define $M > 0$ and the non-empty index set K by

$$-M = \min_{1 \leq i \leq N} u_i, \quad \mathcal{K} = \{k | u_k = -M\}. \tag{3.6.4}$$

Let k be an arbitrary member of $\mathcal{K}$. By noting that A is an L-matrix, it follows from (3.6.4) and the assumption $A\boldsymbol{u} \geq 0$ that

$$|a_{kk}|M = -a_{kk}u_k \leq \sum_{\substack{j=1 \\ j \neq k}}^{N} a_{kj}u_j = \sum_{\substack{j=1 \\ j \neq k}}^{N} -|a_{kj}|u_j \leq \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{kj}|M. \tag{3.6.5}$$

Since $M > 0$ we can divide by M in (3.6.5), which leads to a contradiction with the diagonal dominance of A unless

$$|a_{kk}| = \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{kj}|. \tag{3.6.6}$$

Note that (3.6.6) implies that both inequalities in (3.6.5) are equalities, implying that $u_j = -M$ for all j with $a_{kj} \neq 0$. This means that $j \in \mathcal{K}$ for all j with $a_{kj} \neq 0$. Hence all rows j to which you can "step" from row k (cf. Exercise 3.6.1) also belong to $\mathcal{K}$. Since A is irreducible we can repeat the above argument and "walk" (in one or more steps) to any other row and hence $\mathcal{K} = \{1, 2, \ldots, N\}$. But this means that (3.6.6) must hold for all rows k of A, which contradicts the assumption that (3.6.2) holds strictly for at least one row i. This proves part (i) of the theorem.
The non-singularity of A is proven in Exercise 3.6.2. □

Exercise 3.6.2 *Prove, under the hypothesis of Theorem 3.6.1, that* $A\boldsymbol{u} \leq \mathbf{0}$ *implies* $\boldsymbol{u} \leq \mathbf{0}$ *(Hint: consider* $-\boldsymbol{u}$*). Use this result to prove that* $A\boldsymbol{u} = \mathbf{0}$ *implies* $\boldsymbol{u} = \mathbf{0}$. □

According to Theorem 2.2.1 the solution of the Poisson equation with Neumann boundary conditions is not unique. In that case the row sum of each row of the matrix is equal to 0. In Exercise 3.6.3 it is shown that also the numerical solution is not unique.

Exercise 3.6.3 *Show that if equality holds in Equation* (3.6.1c) *for all k the system* $A\boldsymbol{u} = \mathbf{0}$ *(A being an L-matrix) has a nontrivial solution. Determine that solution.* □

Exercise 3.6.4 *Use Theorem 3.6.1 to prove that if A is an irreducibly diagonally dominant L-matrix and* $A\boldsymbol{u} = \boldsymbol{f}$ *and* $A\boldsymbol{w} = |\boldsymbol{f}|$*, then* $|\boldsymbol{u}| \leq \boldsymbol{w}$.
Hint: also consider $A(-\boldsymbol{u})$. □

3.6.1.1 Discrete harmonics and linear interpolation

We show an important consequence of the discrete maximum principle. This theorem is in fact the discrete equivalent of the weak maximum principle (Theorem 2.3.2).

Theorem 3.6.2 *A discrete solution to Laplace's equation with Dirichlet boundary conditions has its maximum and minimum on the physical boundary, provided the boundary conditions have been approximated by linear interpolation.*

Proof
We only sketch the proof, the reader will have no difficulty in filling in the details. The ordinary five point molecule to approximate the Laplace operator generates an irreducibly diagonally dominant L-matrix, and application of linear interpolation does not alter that. The inequality (3.6.1c) is only strict for those molecules (rows) that contain a Dirichlet boundary condition. So the maximum M will, by a now familiar reasoning, be attained by an interior point that is one cell away from the boundary, like u_C in Figure 3.16. This equation has been modified into:

$$-u_W - u_N - u_E + (3 + \frac{1}{1+s})u_C = \frac{1}{1+s}u_B. \tag{3.6.7}$$

But since u_N, u_W and u_E are not greater than M this means

$$-3M + (3 + \frac{1}{1+s})u_C \leq \frac{1}{1+s}u_B, \tag{3.6.8}$$

or since $u_C = M$ by assumption

$$M \leq u_B. \tag{3.6.9}$$

An analogous reasoning shows that the minimum m is attained at the physical boundary. □

Exercise 3.6.5 *Derive Equation (3.6.7).* □

3.6.2 Super solutions

The (discrete) maximum principle is used to *bound* (discrete) solutions to Poisson's equation. Why would we want to do such a thing? Remember that we have an error estimate in the form:

$$A\boldsymbol{\varepsilon} = h^2\boldsymbol{p}, \tag{3.6.10}$$

in which the vector $\boldsymbol{p}$ is uniformly bounded as $h \to 0$. Suppose we had a solution $\boldsymbol{q}$ to the equation $A\boldsymbol{q} = \boldsymbol{p}$; we would then have an error estimate $\boldsymbol{\varepsilon} = h^2\boldsymbol{q}$. Usually this is asking too much. But if we are able to *bound* the vector $\boldsymbol{p}$ by a vector $\boldsymbol{r} \geq \boldsymbol{p}$ then the solution $\boldsymbol{s}$ to $A\boldsymbol{s} = \boldsymbol{r}$ bounds $\boldsymbol{q}$ by the discrete maximum principle: $\boldsymbol{q} \leq \boldsymbol{s}$. This gives us an error estimate as well: $\boldsymbol{\varepsilon} \leq h^2\boldsymbol{s}$. Such a *super solution* $\boldsymbol{s}$ is obtained by solving the Laplacian for a specific right-hand side that has the properties:

- the solution can be easily obtained;
- it dominates the right-hand side of the equation that we are interested in.

An obvious choice for the vector $\boldsymbol{r}$ would be the constant vector $h^2\|\boldsymbol{p}\|_\infty$. We will show that to get the solution $\boldsymbol{s}$, it is sufficient to consider the equation $-\Delta u = 1$.

3.6.2.1 A discrete solution to $-\Delta u = 1$

Consider the problem $-\Delta v = 1$ on a disk of radius 1 and the origin as its midpoint with homogeneous Dirichlet boundary conditions. By substitution it is easily verified that $v = \frac{1}{4}(1 - x^2 - y^2)$ is the solution of this problem. But since second divided differences are *exact* for polynomials of degree 2 (why?) the discrete function $v_{ij} = \frac{1}{4}(1 - x_i^2 - y_j^2)$ is a solution to the discretized equation $A\boldsymbol{u} = \boldsymbol{e}$ in which $\boldsymbol{e}$ contains all ones and the single index vector $\boldsymbol{u}$ is an appropriate remap of the double index vector v_{ij}. That is, if we disregard the approximation to the boundary conditions for the moment.

Exercise 3.6.6 *Show that* $\|\boldsymbol{u}\|_\infty = \frac{1}{4}$. □

Exercise 3.6.7 *Give the solution of* $-\Delta u = 1$ *with homogeneous Dirichlet boundary conditions on a disk D with midpoint* $(0,0)$ *and radius R. Show that this is a super solution to the same problem on an arbitrary G region wholly contained in D. Hint: consider the difference of the two solutions and show that they satisfy a Laplace equation with nonnegative boundary conditions. Use Theorem 3.6.2 to conclude that the difference must be nonnegative also.* □

3.6.2.2 Pesky mathematical details: the boundary condition

To develop our train of thoughts unhampered in the previous section we overlooked a pesky mathematical detail. At a boundary point we used linear interpolation and that has influenced our equation somewhat. As a result, the function v_{ij} as introduced in the previous paragraph is not really the solution of $A\boldsymbol{u} = \boldsymbol{e}$ but rather of a perturbed system $A\tilde{\boldsymbol{u}} = \boldsymbol{e} + \boldsymbol{e}_b$. The vector $\boldsymbol{e}_b$ contains the interpolation error of $\mathcal{O}(h^2)$ at the boundary.

Exercise 3.6.8 *Consider the discretization of* $-\Delta u = 1$ *with homogeneous Dirichlet boundary conditions on the disk with radius 1 in the neighborhood of the boundary as in Figure 3.16. Show that this discretization is given by:*

$$-u_W - u_N - u_E + (3 + \frac{1}{1+s})u_C = h^2. \tag{3.6.11}$$

Verify, that the discrete solution $v_{ij} = \frac{1}{4}(1 - x_i^2 - y_j^2)$ *does not satisfy* this *equation, but rather the equation:*

$$-u_W - u_N - u_E + (3 + \frac{1}{1+s})u_C = h^2 + \frac{s}{4}h^2. \tag{3.6.12}$$

*(*Hint: $1 - x_i^2 - (y_j - (1+s)h)^2 = 0$.*)*
Show that this is equivalent with an error in the boundary condition Δu_B *of* $\mathcal{O}(h^2)$.
□

Exercise 3.6.9 *Show by using Theorem 3.6.2 and the result of Exercise 3.6.8 that* $\tilde{\boldsymbol{u}} - \boldsymbol{u} = \mathcal{O}(h^2)$. □

In the sequel we shall neglect the influence of linear interpolation error on the boundary conditions.

3.6.2.3 A pointwise error estimate to the discrete solution

Let us apply the results of the previous sections to our error estimate. We have the following theorem:

Theorem 3.6.3 *Let $A\boldsymbol{u} = \boldsymbol{f}$ be the discretization of the Poisson equation with homogeneous Dirichlet boundary conditions on a region G wholly contained in a disk with radius R. Let the discretization error be given by $A\boldsymbol{\varepsilon} = h^2\boldsymbol{p}$ such that $\|\boldsymbol{p}\|_\infty$ is bounded as $h \to 0$. Then*

$$\|\boldsymbol{\varepsilon}\|_\infty \leq \frac{1}{4}R^2h^2\|\boldsymbol{p}\|_\infty \tag{3.6.13}$$

Exercise 3.6.10 *Explain why the midpoint of the disk does not play a role in Theorem 3.6.3. Is it true that we can take the smallest disk that wholly contains G?* □

Exercise 3.6.11 *Show that if $A\boldsymbol{w} = \|\boldsymbol{p}\|_\infty\boldsymbol{e}$, then $|\varepsilon| \leq h^2\boldsymbol{w}$.* □

Exercise 3.6.12 *Prove Theorem 3.6.3.* □

3.7 Boundary fitted coordinates

In Section 3.5 we looked at boundary conditions on general domains. A different approach is the use of *boundary fitted coordinates* that make the boundary of the domain a coordinate line. This usually leads to a reformulation of the problem in *general curvilinear coordinates*. This solves one problem, but introduces another because usually the PDE (even a simple PDE like the Laplacian) can easily become very complex. This approach can also be used if one wants to apply a local grid refinement. We will explain the principle for a one-dimensional problem. Suppose that one has to solve the following problem:

$$-\frac{d}{dx}\left(D(x)\frac{du}{dx}\right) = f(x), \text{ with } u(0) = 0 \text{ and } u(1) = 1. \tag{3.7.1}$$

Here $D(x)$ and $f(x)$ are given functions. For specific choices of $D(x)$ and $f(x)$ a local grid refinement is desirable at positions where the magnitude of the second derivative is large. One can use a coordinate transformation such that the grid spacing is uniform in the transformed co-ordinate. Let this coordinate be given by ξ, then in general, the relation between x and ξ can be written as

$$x = \Gamma(\xi), \tag{3.7.2}$$

where Γ represents the function for the coordinate transformation and we require that Γ is a *bijection* (that is, Γ is *one-to-one* and *onto*). Then, differentiation with respect to x yields

$$1 = \Gamma'(\xi)\frac{d\xi}{dx}, \tag{3.7.3}$$

so $\frac{d\xi}{dx} = \frac{1}{\Gamma'(\xi)}$ and this implies, after using the Chain Rule for differentiation

$$\frac{du}{dx} = \frac{1}{\Gamma'(\xi)}\frac{du}{d\xi}. \tag{3.7.4}$$

Hence, the differential equation (3.7.1) in x transforms into the following differential equation for ξ

$$\begin{aligned} -\frac{1}{\Gamma'(\xi)}\frac{d}{d\xi}\left[\frac{D(\Gamma(\xi))}{\Gamma'(\xi)}\frac{du}{d\xi}\right] &= f(\Gamma(\xi)). \\ u(\xi_L) = 0, \qquad u(\xi_R) &= 1, \end{aligned} \tag{3.7.5}$$

where $0 = \Gamma(\xi_L)$ and $1 = \Gamma(\xi_R)$. The above differential equation is much more complicated than E quation (3.7.1), but it can be solved on an equidistant grid. After the equation is solved, the solution is mapped onto the grid nodes on the x-number line. In practice, one often does not know the function $\Gamma(\xi)$ in an explicit form, so one has to use a numerical approximation for the derivative of $\Gamma(\xi)$. We will return to this subject in Section 4.3.1.

Exercise 3.7.1 *Consider equation* (3.7.1)*, where*

$$f(x) = \begin{cases} 256\,(x-1/4)^2(x-3/4)^2, & \text{for } 1/4 < x < 3/4 \\ 0, & \text{elsewhere.} \end{cases}$$

Suppose that we prefer to discretize such that the mesh is refined at positions where the error is maximal. Then, one has to use a local mesh refinement near $x = 1/2$. Therefore, we use the transformation $x = \Gamma(\xi) = 3\xi - 2\xi^2(3-2\xi)$. Show that this transformation yields a mesh refinement at $x = 1/2$, and give the transformed differential equation expressed in ξ, in which one will use an equidistant grid. □

The extension to two dimensions is quite simple. Consider for example Poisson's equation on a disk,

$$-\text{div grad } u = f(x,y), \text{ for } (x,y) \in \Omega. \tag{3.7.6}$$

In order to get a rectangular grid we map the disk onto a rectangle in (r,θ) space, i.e. we transform to polar coordinates. This transformation is defined by

$$x = r\cos\theta, \qquad y = r\sin\theta. \tag{3.7.7}$$

Exercise 3.7.2 *Express the derivatives of u with respect to x and y in $\frac{\partial u}{\partial r}$ and $\frac{\partial u}{\partial \theta}$.* □

Exercise 3.7.3 *Show that the derivatives, $\frac{\partial r}{\partial x}$, $\frac{\partial r}{\partial y}$, $\frac{\partial \theta}{\partial x}$ and $\frac{\partial \theta}{\partial y}$ are given by*

$$\begin{pmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{pmatrix} = \frac{1}{r}\begin{pmatrix} r\cos\theta & r\sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}. \tag{3.7.8}$$

□

Exercise 3.7.4 *Use the results of Exercises 3.7.2 and 3.7.3 to prove that the Poisson equation (3.7.6) in polar coordinates is defined by*

$$-\left(\frac{\partial^2 u}{\partial r^2}+\frac{1}{r}\frac{\partial u}{\partial r}+\frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}\right)=f(r\cos\theta, r\sin\theta). \tag{3.7.9}$$

□

Remark
Note that $r = 0$ is a singular line in Equation (3.7.9).

Exercise 3.7.5 *Which boundary conditions are needed to get rid of the singularity?* □

Exercise 3.7.6 *Discretize Poisson's equation on a disk of radius 1 in the (r, θ)-plane. Use homogeneous Dirichlet boundary conditions on the disk. Formulate boundary conditions for $r = 0$, $\theta = 0$ and $\theta = 2\pi$.* □

3.8 Summary of Chapter 3

In this chapter we have seen finite difference methods in one and two dimensions. We have looked at the effect of a boundary layer on numerical approximations. We have derived pointwise error estimates for problems with homogeneous Dirichlet boundary conditions using a discrete maximum principle. A method to include Dirichlet boundary conditions on more general regions has been shown and finally we have presented the formula of the Laplacian operator in general coordinates.

Chapter 4

Finite volume methods

Objectives

In the previous chapter we got to know discretization by finite differences. This discretization has two major disadvantages: it is not very clear how to proceed with non-equidistant grids; moreover natural boundary conditions are very hard to implement, especially in two or three dimensions. The finite volume discretization that we are about to introduce do not possess these disadvantages. But they apply only to differential operators in *divergence* or *conservation form*. For physical problems this is rather a feature than a bug: usually the conservation property of the continuous model will be inherited by the discrete numerical model.

We shall start out with a one-dimensional example that we left dangling in our previous chapter: a second order equation on a non-equidistant grid. We shall pay attention to Neumann and Robin boundary conditions too. Subsequently we shall turn our attention to two dimensions and discretize the Laplacian in general coordinates. Then we will look at problems with two components: fluid flow and plane stress. We shall introduce the concept of *staggered grids* and show that that is a natural way to treat these problems. There will be a problem at the boundaries in this case that we have to pay attention to.

4.1 Heat transfer with varying coefficient

We consider the diffusion equation on the interval $(0, 1)$:

$$-\frac{d}{dx}\left(\lambda\frac{dT}{dx}\right) = f, \quad \lambda\frac{dT}{dx}(0) = 0, \quad -\lambda\frac{dT}{dx}(1) = \alpha(T(1) - T_R). \tag{4.1.1}$$

In this equation λ may depend on the space coordinate x. T_R is a (given) reference temperature and as you see we have natural boundary conditions

on both sides of the interval. We divide the interval in (not necessarily equal) subintervals $e_k, k = 1, \ldots, N$, where e_k is bounded by the nodal points x_{k-1}, x_k. See Figure 4.1.

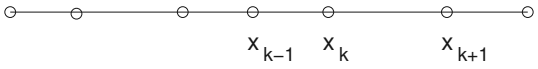

Figure 4.1: Non-equidistant grid.

To derive a discrete equation to this problem we consider three subsequent nodes x_{k-1}, x_k and x_{k+1} in isolation, see Figure 4.2.

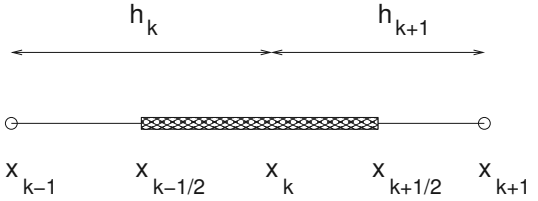

Figure 4.2: Control volume.

We let $h_k = x_k - x_{k-1}$, $h_{k+1} = x_{k+1} - x_k$ and define $x_{k-1/2} = x_k - \frac{1}{2}h_k$ and $x_{k+1/2} = x_k + \frac{1}{2}h_{k+1}$. We now integrate Equation (4.1.1) over the *control volume* $(x_{k-1/2}, x_{k+1/2})$ to obtain

$$\int_{x_{k-1/2}}^{x_{k+1/2}} -\frac{d}{dx}\left(\lambda \frac{dT}{dx}\right)\, dx = \int_{x_{k-1/2}}^{x_{k+1/2}} f\, dx,$$

which gives

$$-\lambda \frac{dT}{dx}\bigg|_{x_{k+1/2}} + \lambda \frac{dT}{dx}\bigg|_{x_{k-1/2}} = \int_{x_{k-1/2}}^{x_{k+1/2}} f\, dx. \tag{4.1.2}$$

Equation (4.1.2) represents the physical conservation law: the net outflow through the left and right boundary of the control volume is equal to the production in the control volume. We may approximate the derivatives on the left-hand side by central divided differences and the integral on the right by one-point integration to obtain:

$$\lambda_{k-1/2}\frac{T_k - T_{k-1}}{h_k} - \lambda_{k+1/2}\frac{T_{k+1} - T_k}{h_{k+1}} = \frac{1}{2}(h_k + h_{k+1})f_k + E_T, \tag{4.1.3}$$

which after rearrangement becomes:

$$-\frac{\lambda_{k-1/2}}{h_k}T_{k-1}+\left(\frac{\lambda_{k-1/2}}{h_k}+\frac{\lambda_{k+1/2}}{h_{k+1}}\right)T_k-\frac{\lambda_{k+1/2}}{h_{k+1}}T_{k+1} = \frac{1}{2}(h_k+h_{k+1})f_k+E_T. \quad (4.1.4)$$

The structure of the error term E_T will be considered in Exercises 4.1.2 and 4.1.3. To get a set of discrete equations we drop the error term.

Exercise 4.1.1 *Show that in case of an equidistant grid Equation (4.1.4) without the error term is identical to the finite difference discretization of (4.1.1) multiplied by the length h. (Hint: check Equation* (3.2.7)*.)* □

The error E_T in Equation (4.1.4) consist of two terms: one part of the error, E_1, originates from the use of one point integration instead of exact integration, the other part, E_2, originates from the use of central differences instead of derivatives. In the following exercises it is shown that both error terms E_1 and E_2 are of the order $\mathcal{O}(h_{k+1}^2-h_k^2)+\mathcal{O}(h_{k+1}^3+h_k^3)$. Further, if the grid spacing satisfies $h_{k+1}=h_k(1+\mathcal{O}(h))$, where h denotes the maximum h_k, then it is shown in a subsequent exercise that both error terms are of the order $\mathcal{O}(h^3)$. The global error is one order lower, that is $\mathcal{O}(h^2)$, since compared to the finite difference method all equations are multiplied by the length h.

Exercise 4.1.2 *Show that the error that originates from the one-point integration is given by* $E_1=\mathcal{O}(h_{k+1}^2-h_k^2)+\mathcal{O}(h_{k+1}^3+h_k^3)$.
Hint: Assume that $f(x)$ *is the derivative of* $F(x)$*. Express the integral in terms of F and use Taylor series expansion.* □

Exercise 4.1.3 *Show that the error from the use of central differences is given by* $E_2=\mathcal{O}(h_{k+1}^2-h_k^2)+\mathcal{O}(h_{k+1}^3+h_k^3)$*. You may assume that* λ *does not depend on* x*.* □

Exercise 4.1.4 *Show that if* $h_{k+1}=h_k(1+\mathcal{O}(h)), k=1,\ldots,N-1$*, then* $h_{k+1}-h_k=\mathcal{O}(h^2), k=1,\ldots,N-1$*, and therefore* $E_1=\mathcal{O}(h^3)$ *and* $E_2=\mathcal{O}(h^3)$*.* □

4.1.1 The boundaries

At the left-hand boundary we take $(x_0, x_{1/2})$ as control volume and we integrate to get:

$$\lambda\frac{dT}{dx}\bigg|_{x_0}-\lambda\frac{dT}{dx}\bigg|_{x_{1/2}}=\int_{x_0}^{x_{1/2}} f\,dx. \quad (4.1.5)$$

The left-hand boundary condition can be substituted directly:

$$-\lambda\frac{dT}{dx}\bigg|_{x_{1/2}}=\int_{x_0}^{x_{1/2}} f\,dx. \quad (4.1.6)$$

Application of central differences and one-point integration gives:

$$\frac{\lambda_{1/2}}{h_1}T_0 - \frac{\lambda_{1/2}}{h_1}T_1 = \frac{1}{2}h_1 f_0 + E_T. \tag{4.1.7}$$

Exercise 4.1.5 *Show that the truncation error E_T is $\mathcal{O}(h_1^2)$ in the above equation.* □

At the right-hand boundary we take $(x_{N-1/2}, x_N)$ as control volume and integrate to get:

$$\lambda \frac{dT}{dx}\bigg|_{x_{N-1/2}} - \lambda \frac{dT}{dx}\bigg|_{x_N} = \int_{x_{N-1/2}}^{x_N} f\, dx. \tag{4.1.8}$$

On substitution of the right-hand boundary condition this becomes:

$$\lambda \frac{dT}{dx}\bigg|_{x_{N-1/2}} + \alpha T_N = \int_{x_{N-1/2}}^{x_N} f\, dx + \alpha T_R. \tag{4.1.9}$$

Application of central differences and one-point integration gives:

$$-\frac{\lambda_{N-1/2}}{h_N}T_{N-1} + \left(\frac{\lambda_{N-1/2}}{h_N} + \alpha\right) T_N = \frac{1}{2}h_N f_N + \alpha T_R + E_T. \tag{4.1.10}$$

Remark
If we would have a Dirichlet boundary condition, for example $T = T_0$ at the left-hand boundary, there is no need to use the control volume $(x_0, x_{1/2})$. We treat this boundary condition like in Chapter 3, i.e., we substitute the given value and no extra equation is required.

4.1.2 Conservation

Finite volume schemes are often described as *conservative schemes* for the following reason. When we write the finite volume equations in terms of *fluxes* by applying Fick's (Darcy's, Ohm's, Fourier's) law for each finite volume (x_L, x_R), each equation looks like:

$$q_R - q_L = \int_{x_L}^{x_R} f\, dx, \tag{4.1.11}$$

or in words: the local production in a control volume is equal to the net outflow through its boundary points. This will be true *regardless of the numerical approximation to the fluxes*. If the production is zero, there will be no generation of mass (energy, momentum) by the numerical scheme. The only error that will be made in the fluxes will be caused by the error in approximating the production term.
In the following exercises we shall prove that the error in the flux is equal to the error in the inflow flux at the left boundary $x = 0$ plus the maximum error in the production, provided the flux itself is not discretized.

Exercise 4.1.6 *Show that if the equation*

$$-\left(\lambda y'\right)' = 0 \tag{4.1.12}$$

is discretized on the interval $(0,1)$ *by the Finite Volume Method, necessarily* $q_0 = q_N$ *with* $q = -\lambda y'$*, regardless of the number of steps* N*.* □

Exercise 4.1.7 *Show that if the equation*

$$-\left(\lambda y'\right)' = 1 \tag{4.1.13}$$

is discretized on the interval $(0,1)$ *by the Finite Volume Method, necessarily* $q_N = q_0 + 1$ *with* $q = -\lambda y'$*, regardless of the number of steps* N*.* □

Let the vector $\boldsymbol{q}$ be the vector of fluxes defined for all boundary points of the control volumes. Hence $\boldsymbol{q} = (q_0, q_{1/2}, q_{3/2}, \ldots, q_{N-1/2}, q_N)^T$. The error in the fluxes is denoted by $\boldsymbol{dq}$, and its components can be solved from the following relations, where the right-hand sides are the production errors:

$$dq_{1/2} - dq_0 = \tfrac{1}{2} h_1 E_0, \tag{4.1.14a}$$

$$dq_{k+1/2} - dq_{k-1/2} = \tfrac{1}{2}(h_k + h_{k+1}) E_k \quad (k = 1, \ldots, N-1), \tag{4.1.14b}$$

$$dq_N - dq_{N-1/2} = \tfrac{1}{2} h_N E_N, \tag{4.1.14c}$$

Exercise 4.1.8 (Propagation of production error)
Show that it follows from (4.1.14) *that*

$$|dq_{k+1/2}| \leq |dq_0| + \max_{0 \leq j \leq k} |E_j|, \quad k = 0, 1, \ldots, N-1, \tag{4.1.15a}$$

$$|dq_N| \leq |dq_0| + \max_{0 \leq j \leq N} |E_j|. \tag{4.1.15b}$$

Hint: use $\sum_k h_k = 1$*.* □

Exercise 4.1.9 (Propagation of boundary error)
Show that it follows from (4.1.14) *with* $E_k = 0$ *(*$k = 0, 1, \ldots, N$*) that all components of* $\boldsymbol{dq}$ *are equal to* dq_0*.* □

4.1.3 Error in the temperatures

The error in the fluxes is in general of the same order as the error in the production terms (see Exercise 4.1.8). Since we have approximated this term with one-point integration, we may expect an error of magnitude $\mathcal{O}(h^2)$ in the fluxes, $q_{k+1/2}$, for smoothly varying stepsizes. By the same reasoning as in Exercise 4.1.8 we may now show, that the error in the temperatures *remains* $\mathcal{O}(h^2)$, because if

$$-\lambda \tilde{T}'(x_{k+1/2}) = q_{k+1/2} + \mathcal{O}(h^2), \tag{4.1.16}$$

the approximation with central differences *also* generates an $\mathcal{O}(h^2)$ error term and we get for the error dT_k:

$$\lambda_{k+1/2}\frac{dT_k - dT_{k+1}}{h_{k+1}} = E_{k+1}, \tag{4.1.17}$$

where $E_{k+1} = \mathcal{O}(h^2)$. Now defining *the error in temperature* $\boldsymbol{dT}$ in much the same way as in Exercise 4.1.8 we can show that

$$|dT_k| \leq |dT_N| + \sup_{j>k} |E_j|/\lambda_{j-1/2}. \tag{4.1.18}$$

However, since the numerical approximation to q_N has an error of $\mathcal{O}(h^2)$, it follows from the right-hand-boundary condition $q_N = \alpha(T_N - T_R)$ that $dT_N = \mathcal{O}(h^2)$. Backsubstitution into inequality (4.1.18) proves the result.

4.2 The stationary diffusion equation in 2 dimensions

The Finite Volume approximation of the stationary diffusion equation in two dimensions is a straightforward generalization of the previous section. Let us consider:

$$-\operatorname{div} \lambda \operatorname{grad} u = f, \qquad \boldsymbol{x} \in \Omega, \tag{4.2.1a}$$

$$-\lambda\frac{\partial u}{\partial n} = \alpha(u - u_0), \quad \boldsymbol{x} \in \Gamma. \tag{4.2.1b}$$

Both λ and f are functions of the coordinates x and y. In the boundary condition the radiation coefficient $\alpha > 0$ and the reference temperature u_0 are known functions of $\boldsymbol{x}$. We subdivide the region Ω into cells like in Figure 4.3.

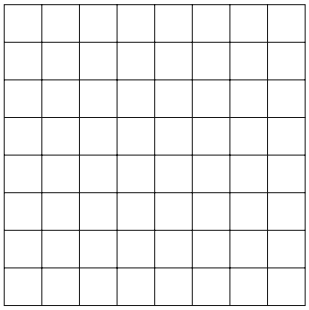

Figure 4.3: Subdivision of rectangular region into cells.

Usually these cells are rectangles, but also quadrilaterals or even triangles are allowed. In the literature one can find two ways of positioning the unknowns.

The first one is to place the unknowns in the nodes of the grid. This is called the *vertex-centered* approach. The other one is to put the unknowns in the centers of the cells (*cell-centered*). These methods only differ at the boundary of the domain. For the moment we restrict ourselves to the vertex-centered method, and a rectangular equidistant grid.

We use the same (i, j) notation for the nodes as in Chapter 3. In the literature a node x_{ij} somewhere in the interior of Ω is also denoted by x_C and the surrounding neighbors by their compass names in capitals: N, E, S, W. Cell quantities and quantities on the cell edges are denoted with lower case subscripts: n, s, e, w. If appropriate we shall also apply this notation. We construct a control volume V with edges half way between two nodes, like in Figure 4.4.

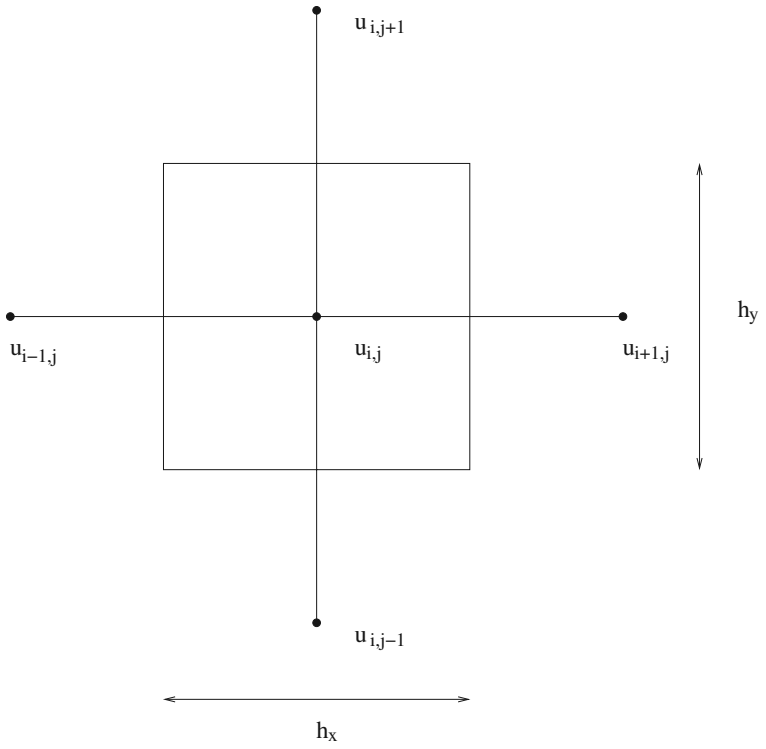

Figure 4.4: Control volume for the diffusion equation.

We integrate the equation over the control volume V to obtain:

$$\int_V -\text{div}\, \lambda \,\text{grad}\, u \; dV = \int_V f \; dV,$$

which we can rewrite by using the divergence theorem (Theorem 1.3.1) as

$$\int_{\partial V} -\lambda \frac{\partial u}{\partial n}\, d\Gamma = \int_{V} f\, dV. \tag{4.2.2}$$

Using central differences for $\frac{\partial u}{\partial n}$ and one-point integration for the left-hand-side edges and the right-hand-side volume we get the interior molecule:

$$-\lambda_{i-1/2,j} h_y \frac{u_{i-1,j} - u_{i,j}}{h_x} - \lambda_{i,j-1/2} h_x \frac{u_{i,j-1} - u_{i,j}}{h_y} - \lambda_{i+1/2,j} h_y \frac{u_{i+1,j} - u_{i,j}}{h_x}$$
$$- \lambda_{i,j+1/2} h_x \frac{u_{i,j+1} - u_{i,j}}{h_y} = h_x h_y f_{i,j}. \tag{4.2.3}$$

Note that Equation (4.2.3) is identical to the finite difference Equation (3.4.2) if $\lambda = 1$.

Exercise 4.2.1 *Derive the finite volume discretization of (4.2.1) for non-uniform step-sizes.* □

Exercise 4.2.2 *Apply the finite volume method to the* convection-diffusion equation *with incompressible flow:*

$$\text{div}\ (-\varepsilon(\text{grad}\, c) + c\boldsymbol{u}) = 0, \tag{4.2.4}$$

with ε and $\boldsymbol{u}$ constant. Show that the contribution of the convection term is non-symmetric. □

4.2.1 Boundary conditions in case of a vertex-centered method

The treatment of boundary conditions is usually the most difficult part of the finite volume method. Dirichlet boundary conditions are treated in the same way as in 1D. The Robin boundary condition (4.2.1b) requires a special approach. For simplicity we restrict ourselves to the east boundary. All other boundaries can be dealt with in the same way. Since the nodes on the boundary correspond to the unknown function u, it is necessary to define a control volume around these points. The common approach is to take only the *half part* inside the domain as sketched in Figure 4.5. Integration of the diffusion equation (4.2.1a) over the control volume gives Equation (4.2.2). The integral over the west edge is treated as for the internal points. The integral over the north and south edges are also treated in the same way, but their length is multiplied by $\frac{1}{2}$. On the east edge boundary condition (4.2.1b) is applied to get

$$\int_{\Gamma_e} -\lambda \frac{\partial u}{\partial n}\, d\Gamma \;=\; \int_{\Gamma_e} \alpha(u - u_0)\, d\Gamma. \tag{4.2.5}$$

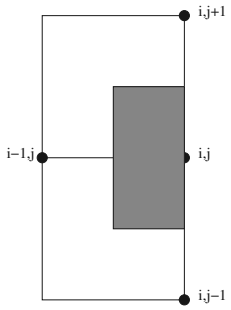

Figure 4.5: Half cell control volume for the vertex-centered Robin b.c.

Discretization of the right-hand side of (4.2.5) gives

$$\int_{\Gamma_e} \alpha(u - u_0)\, d\Gamma \approx h_y \alpha_{i,j}(u_{i,j} - (u_0)_{i,j}), \tag{4.2.6}$$

so the complete discretization for a point (i, j) at the east boundary becomes

$$-\lambda_{i-1/2,j} h_y \frac{u_{i-1,j} - u_{i,j}}{h_x} - \lambda_{i,j-1/2} h_x \frac{u_{i,j-1} - u_{i,j}}{2h_y} - \lambda_{i,j+1/2} h_x \frac{u_{i,j+1} - u_{i,j}}{2h_y}$$
$$+ h_y \alpha_{i,j} u_{i,j} = h_y \alpha_{i,j} (u_0)_{i,j} + \frac{h_x h_y}{2} f_{i,j}. \tag{4.2.7}$$

Exercise 4.2.3 *Suppose we want to solve the diffusion equation (4.2.1a) over the square $\Omega = (0,1) \times (0,1)$. Let λ and f be periodic in x-direction. Assume we have periodic boundary conditions at the boundaries $x = 0$ and $x = 1$. Furthermore boundary condition (4.2.1b) holds for the other two boundaries.*

(i) Formulate the periodic boundary conditions at $x = 0$ and $x = 1$. Motivate why the number of boundary conditions is correct.

(ii) Derive the finite volume discretization of the equation at the periodic boundaries. Use an equidistant grid with $h_x = h_y$. □

4.2.2 Boundary conditions in case of a cell-centered method

If a cell-centered method is applied, cells and control volumes coincide. All unknowns are positioned in the centers of the cells, which implies that there are no unknowns on the boundary.

Exercise 4.2.4 *Show that the discretization of Equation (4.2.1a) for all internal cells (which have no common edge with the boundary), is given by Equation (4.2.3).* □

The absence of unknowns on the boundary has its effect on the treatment of boundary conditions. Neumann boundary conditions of the type

$$-\lambda \frac{\partial u}{\partial n} = g \text{ at } \Gamma \tag{4.2.8}$$

are the easiest to implement since (4.2.8) can be substituted immediately in the boundary integrals.

Exercise 4.2.5 *Derive the discretization for a boundary cell with boundary condition (4.2.8).* □

In case of a Dirichlet boundary condition $u = g_2$ on the south boundary, one may introduce a virtual point $i, j-1$ like in Figure 3.15. The value of $u_{i,j-1}$ can be expressed in terms of $u_{i,j}$ and the boundary value $u_{i,j-1/2}$ using linear extrapolation. Substitution in the 5-point molecule results in a 4-point stencil.

The introduction of such a virtual point can actually be avoided by the following equivalent approach, where the boundary integral

$$\int_{\Gamma_S} -\lambda \frac{\partial u}{\partial n}\, d\Gamma \tag{4.2.9}$$

is approximated with the midpoint rule, where the value of the integrand in the south boundary point B (with indices $i, j-1/2$) is approximated by

$$-\lambda \frac{\partial u}{\partial n} \approx -\lambda_B \frac{u_B - u_C}{\frac{1}{2}\Delta y} = -\lambda_{i,j-1/2} \frac{(g_2)_{i,j-1/2} - u_{i,j}}{\frac{1}{2}\Delta y}. \tag{4.2.10}$$

Exercise 4.2.6 *Derive the discretization of Equation (4.2.1a) in a cell adjacent to the Dirichlet boundary.* □

The Robin boundary condition (4.2.1b) is the most difficult to treat. On the boundary we have to evaluate the integral

$$\int_{\partial V} \alpha(u - u_0)\, d\Gamma \tag{4.2.11}$$

while u is unknown, and not present on the boundary either. In order to keep the second order accuracy, the best way is to express u using linear extrapolation from two internal points. Consider for example the south boundary in Figure 4.6. We can express u_B in terms of u_C and u_N using linear extrapolation, resulting again in a 4-point molecule.

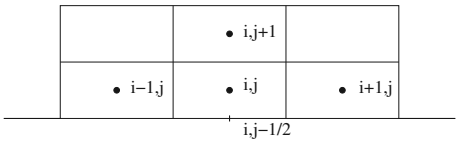

Figure 4.6: Control volume for the cell-centered Robin boundary condition.

Exercise 4.2.7 *Derive the 4-point molecule.* □

4.2.3 Boundary cells in case of a skewed boundary

This section applies to both the vertex- and the cell-centered method.

The best way to treat a *skewed boundary* is to make sure that unknowns fall *on* the boundary. This leads to triangular grid-cells at the boundary, see Figure 4.7.

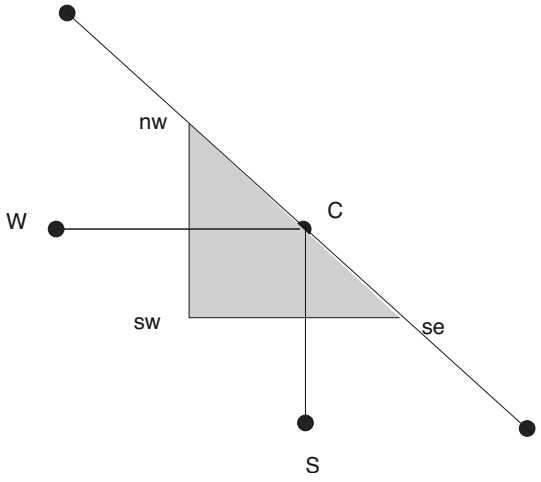

Figure 4.7: Triangular boundary cell.

Integration over the triangle and substitution of central differences give with the notations of Figure 4.7:

$$-\beta_W u_W - \beta_S u_S + (\beta_W + \beta_S) u_C + \int_{hyp} -\lambda \frac{\partial u}{\partial n}\, d\Gamma = \frac{1}{2} h_x h_y f_C, \qquad (4.2.12)$$

where the integral has to be taken over the hypotenuse of the triangle. Writing h_h for the length of the hypotenuse and substituting the boundary condition (4.2.1b) we get:

$$-\beta_W u_W - \beta_S u_S + (\beta_W + \beta_S + \alpha_C h_h) u_C = \alpha_C h_h u_{0C} + \frac{1}{2} h_x h_y f_C. \qquad (4.2.13)$$

Of course a Dirichlet boundary condition is trivial to implement if the unknowns are on the boundary.

Remark 4.2.1 *(Symmetry and diagonal dominance)*

1. *The discretization matrix generated by the FVM is symmetric;*
2. *The numerical approximation of Problem (4.2.1) with the FVM leads to an irreducibly diagonally dominant L-matrix.*

Exercise 4.2.8 *Prove the first statement in Remark 4.2.1.*
Hint: across a volume edge between, say, volumes $V_{i+1,j}$ and $V_{i,j}$ the flux is approximated in the same way for the equations of $u_{i+1,j}$ and $u_{i,j}$. □

Exercise 4.2.9 *Prove the second statement of Remark 4.2.1.* □

Theorem 4.2.1 *Consider the Finite Volume discretization in this section for Problem* (4.2.1). *If $f \geq 0$ (for $\boldsymbol{x} \in \Omega$) and $u_0 \geq 0$ (for $\boldsymbol{x} \in \Gamma$), then the solution of the discrete problem is non-negative.*

Exercise 4.2.10 *Prove Theorem 4.2.1. (Hint: use the second statement of Remark 4.2.1.)* □

Theorem 4.2.2 *Consider the Finite Volume discretization in this section for Problem* (4.2.1). *The solution of the discrete problem with $f = 0$ has a maximum and minimum on the boundary.*

Exercise 4.2.11 *Prove Theorem 4.2.2. (Hint: use the second statement of Remark 4.2.1.)* □

If the boundary is curved, then the discretization with a rectangular Cartesian grid is toilsome. An alternative could be to introduce boundary fitted coordinates.

4.2.4 Error considerations in the interior

We shall not go into great detail in error analysis, but indicate sources of error. We started out by integrating the conservation law of the flux vector *exactly*:

$$\Phi_w + \Phi_n + \Phi_e + \Phi_s = \int_V f \, dV, \tag{4.2.14}$$

where Φ stands for the net outflow through that particular edge of the control volume. After that we made a number of approximations:

1. Approximate integrals over the sides by one-point integration:
   $\mathcal{O}(h^2)$ accurate for smoothly changing stepsizes, otherwise $\mathcal{O}(h)$.
2. Approximate derivatives by central differences:
   $\mathcal{O}(h^2)$ accurate for smoothly changing stepsizes, otherwise $\mathcal{O}(h)$.
3. Approximate the right-hand side by one-point integration:
   $\mathcal{O}(h^2)$ accurate for smoothly changing stepsizes, otherwise $\mathcal{O}(h)$.

It gets monotonous. From finite difference approximations we already know that *uniform* stepsizes lead to overall $\mathcal{O}(h^2)$ accuracy. The same accuracy can be obtained with smoothly varying stepsizes, by which we mean that $h_{k+1} = h_k(1 + \mathcal{O}(h))$, where h denotes the maximum h_k. Smoothly varying stepsizes were also considered in Section 4.1, and give still pretty much leeway in stretching grids, so that should not be regarded as too restrictive.

4.2.5 Error considerations at the boundary

At the boundary one-point integration of the right-hand side is always $\mathcal{O}(h)$, because the integration point has to be the gravicenter for order $\mathcal{O}(h^2)$ accuracy, whereas the integration point is always on the edge. (Note that in fact the absolute magnitude of the error is $\mathcal{O}(h^3)$, but that is because the volume of integration is itself $\mathcal{O}(h^2)$.)

So the situation looks grim, but in fact there is nothing that should worry us. And that is because of the following phenomenon: for the solution $\boldsymbol{u}$ of the discrete equations with $f = 0$, we have

$$\|\boldsymbol{u}\|_\infty \leq \sup_{\boldsymbol{x}\in\Gamma} |u_0|. \tag{4.2.15}$$

Exercise 4.2.12 *Prove Inequality* (4.2.15)*. Use the results of Exercise 4.2.9 and following items.* □

Exercise 4.2.13 *Prove that if $\tilde{u}_0 = u_0 + \varepsilon_0$ then the perturbation ε in the solution of the homogeneous discrete problem is less than* $\sup|\varepsilon_0|$ *for all components of ε. (Hint: subtract the equations and boundary conditions of u and $\tilde{u}$ to obtain an equation and boundary condition for ε. Then use* (4.2.15)*)* □

From all this we see that a perturbation of $\mathcal{O}(h^3)$ in the right-hand side of equations for the boundary cells leads to an error of $\mathcal{O}(h^2)$ in the solution. But one-point integration of the right-hand side *also* gives a perturbation of $\mathcal{O}(h^3)$. So the effect on the solution should *also* be no worse than $\mathcal{O}(h^2)$.

4.3 Laplacian in general coordinates

4.3.1 Transformation from Cartesian to General coordinates

Consider a region in the x-y-plane as in Figure 4.8 that we want to transform into a rectangular region in the ξ-η-plane.

We assume that there is a formal mapping $x(\xi,\eta)$ and $y(\xi,\eta)$ and its inverse $\xi(x,y)$ and $\eta(x,y)$ exists. Coordinate lines in the ξ-η-plane transform to the curves $\boldsymbol{x}(\xi_0,\eta)$ and $\boldsymbol{x}(\xi,\eta_0)$ respectively. Such a transformation is called regular if it has an inverse, otherwise it is singular. Sufficient conditions for regularity is, that the *Jacobian matrix* exists and is non-singular. The Jacobian matrix consists of the partial derivatives of x and y with respect to ξ and η,

$$J = \begin{pmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{pmatrix} \tag{4.3.1}$$

and its inverse consists of the partial derivatives of ξ and η with respect to x and y,

$$J^{-1} = \begin{pmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{pmatrix}. \tag{4.3.2}$$

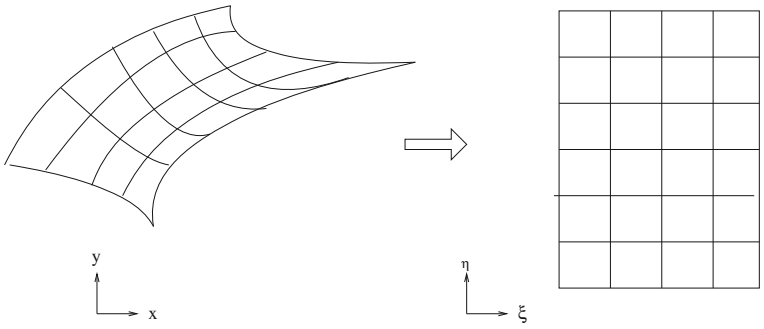

Figure 4.8: General region transformation.

Usually the mapping is only known in the cell vertices. This means that we do not have an analytical expression for the derivatives and we must compute them by finite differences. Unfortunately not all derivatives are easily available. Take a look at a cell in the ξ-η-plane (Figure 4.9): Given the configuration

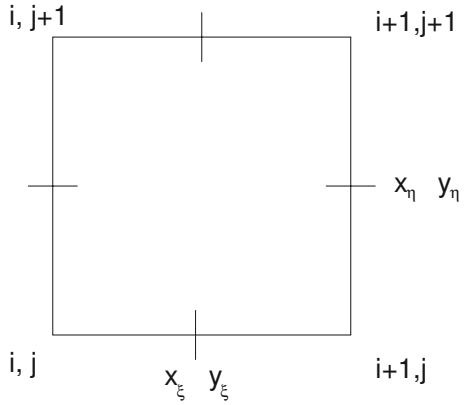

Figure 4.9: Cell in the ξ-η-plane with natural place of coordinate derivatives.

in Figure 4.9, central differences can be applied to compute x_ξ and y_ξ at the midpoints of the horizontal cell boundaries. Analogously, central differences are applied to compute x_η and y_η at the vertical cell boundaries. Everything else has to be computed by averaging over the neighbors. The quantities ξ_x, ξ_y etcetera have to be calculated by inverting J.

Exercise 4.3.1 *Consider Figure 4.9. Explain how to express $x_\xi, x_\eta, y_\xi, y_\eta$ in the cell center in the ξ-η-plane in terms of the cell coordinates in the x-y-plane. Explain how to calculate ξ_x, ξ_y, η_x and η_y.* □

In the Finite Volume Method, we consider integration of a function, or of a differential expression. If a regular transformation is applied from (x, y) to

(ξ, η), then the *Jacobian* enters the picture. Suppose that we integrate over a domain Ω_{xy} defined in (x, y)-space, and that Ω_{xy} is mapped onto $\Omega_{\xi\eta}$ in (ξ, η)-space Then, from Calculus, it follows that

$$\int_{\Omega_{xy}} f(x,y) d\Omega_{xy} = \int_{\Omega_{\xi\eta}} f(x(\xi,\eta), y(\xi,\eta)) \left| \frac{\partial(x,y)}{\partial(\xi,\eta)} \right| d\Omega_{\xi\eta}, \tag{4.3.3}$$

where the Jacobian is defined as the determinant of the Jacobian matrix,

$$\frac{\partial(x,y)}{\partial(\xi,\eta)} = \det(J),$$

which is expressed in the coordinate framework (ξ, η). We use the notation $d\Omega_{xy}$ and $d\Omega_{\xi\eta}$ to emphasize that the integral is in (x, y)-space and (ξ, η)-space respectively. For the derivation of this procedure, we refer to a textbook on Calculus, like Stewart [12] or Adams [1]. This procedure is applied in general to all integrals that are involved in the Finite Volume discretization. We will illustrate how the finite volume method works in a polar coordinate system.

4.3.2 An example of finite volumes in polar coordinates

We consider an example on a cut piece of cake, on which Poisson's equation is imposed,

$$-\text{div grad } u = f(x,y), \text{ on } \Omega, \tag{4.3.4}$$

where Ω is described in polar coordinates by

$$\Omega_{r\theta} = \{(r,\theta) \in \mathbb{R}^2 : 1 < r < 3,\ 0 < \theta < \pi/4\}.$$

To solve the above equation by Finite Volumes, the equation is integrated over a control volume V, to obtain

$$-\int_V \text{div grad } u \, d\Omega_{xy} = \int_V f(x,y) d\Omega_{xy}. \tag{4.3.5}$$

From Equation (3.7.9), we know that the above PDE (4.3.4) is transformed into

$$-\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} \right) = f(r\cos\theta, r\sin\theta). \tag{4.3.6}$$

Note that $\Omega_{r\theta}$ is a rectangular domain in (r, θ)-space. The Jacobian of the transformation from polar coordinates to Cartesian coordinates is given by

$$\frac{\partial(x,y)}{\partial(r,\theta)} = r. \tag{4.3.7}$$

Exercise 4.3.2 *Prove the above formula.* □

Next, we integrate the transformed PDE (4.3.6) over the untransformed control volume in (x, y)-space, and rewrite it to an integral over the transformed control volume in (r, θ)-space, which is rectangular and hence much easier to work with, to get

$$\int_V -\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}\right) r d\Omega_{r\theta} = \int_V f(r\cos\theta, r\sin\theta) r d\Omega_{r\theta}. \tag{4.3.8}$$

Note that the Jacobian has been implemented on both sides of the above equation. The integral of the left-hand side of the above equation can be worked out as

$$\int_V -\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) - \frac{1}{r}\frac{\partial^2 u}{\partial \theta^2} d\Omega_{r\theta} = -\int_V (\frac{\partial}{\partial r}, \frac{\partial}{\partial \theta}) \cdot (r\frac{\partial u}{\partial r}, \frac{1}{r}\frac{\partial u}{\partial \theta}) d\Omega_{r\theta}. \tag{4.3.9}$$

The integrand in the right-hand side of the above equation consists of an inner product of the divergence operator and a vector field. Both vectors are in the (r, θ) frame. The domain over which the integral is determined is closed and hence the divergence theorem can be applied in this volume with piecewise straight boundaries. This implies that Equation (4.3.8) can be written as

$$-\int_{\partial V} (n_r, n_\theta) \cdot (r\frac{\partial u}{\partial r}, \frac{1}{r}\frac{\partial u}{\partial \theta}) d\Gamma = \int_V f(r\cos\theta, r\sin\theta) r d\Omega_{r\theta}. \tag{4.3.10}$$

This equation contains a volume integral with the function f over a control volume and a line integral related to the Laplacian over the boundary of the control volume. The treatment of both integrals is analogous to the Cartesian case: Consider the control volume, with length Δr and $\Delta\theta$, around C, with coordinates (r_C, θ_C) in Figure 4.10. The integral at the right-hand side in the above equation is approximated by

$$\int_V f(r\cos\theta, r\sin\theta) r d\Omega_{r\theta} \approx f_C r_C \Delta r \Delta\theta, \tag{4.3.11}$$

where $f_C = f(r_C\cos\theta_C, r_C\sin\theta_C)$. The boundary integral is replaced by the sum of the approximations of the integrals over all the boundary segments. Substitution of these approximations into (4.3.10) gives the final result for an internal control volume:

$$-\left\{\frac{1}{r_C}\frac{u_S - u_C}{\Delta\theta}\Delta r + r_e\frac{u_E - u_C}{\Delta r}\Delta\theta + \frac{1}{r_C}\frac{u_N - u_C}{\Delta\theta}\Delta r + r_w\frac{u_W - u_C}{\Delta r}\Delta\theta\right\} = f_C r_C \Delta r \Delta\theta. \tag{4.3.12}$$

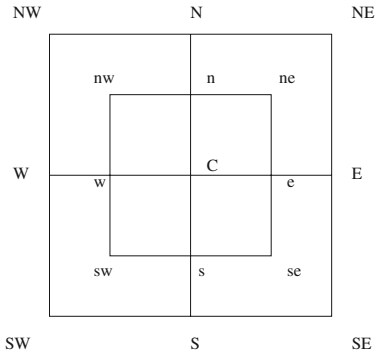

Figure 4.10: General control volume.

4.3.3 Boundary conditions

Boundary conditions of Dirichlet type do not present any problem, so we shall turn our attention to radiation boundary conditions of the form

$$\frac{\partial u}{\partial n} = \alpha(u_0 - u),$$

where we assume for simplicity that α and u_0 are constant. From an implementation point of view, it is easiest to take the nodal points *on* the boundary, which gives us a half cell control volume at the boundary like in Figure 4.11.

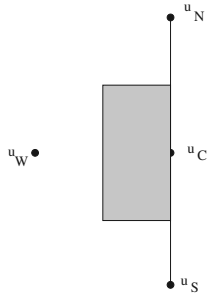

Figure 4.11: Boundary cell.

Integrating over the half volume and applying the divergence theorem we get:

$$-\left\{\frac{1}{r_C}\frac{u_S - u_C}{\Delta\theta}\frac{\Delta r}{2} + r_C\alpha(u_0 - u_C)\Delta\theta + \frac{1}{r_C}\frac{u_N - u_C}{\Delta\theta}\frac{\Delta r}{2} + r_w\frac{u_W - u_C}{\Delta r}\Delta\theta\right\} = f_C r_C\frac{\Delta r}{2}\Delta\theta, \quad (4.3.13)$$

where the radiation boundary condition has been substituted into the boundary integral of the right (east) boundary of the control volume.

4.4 Finite volumes on two component fields

We shall show an example of an application of the FVM on a two component field. We recall the problem for plane stress from Section 2.4.4. We consider a rectangular plate fixed at the sides ABC and subject to a body force $\boldsymbol{b}$ inside $\Omega = ABCD$ and boundary stresses t at the two free sides CDA, see Figure 4.12.

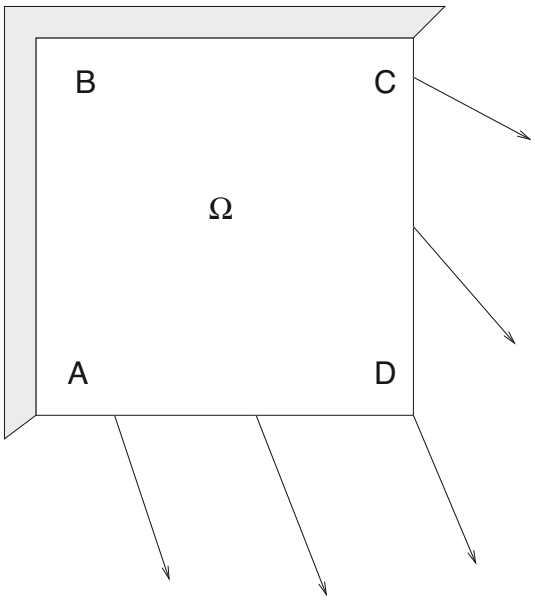

Figure 4.12: Square plate.

The equations for the stresses, already presented in (2.4.14), are:

$$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + b_1 = 0, \tag{4.4.1a}$$

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + b_2 = 0. \tag{4.4.1b}$$

We integrate the first equation over a control volume V_1 and the second one over a control volume V_2. We define

$$\boldsymbol{s}_x = \begin{pmatrix} \sigma_{xx} \\ \tau_{xy} \end{pmatrix} \quad \text{and} \quad \boldsymbol{s}_y = \begin{pmatrix} \tau_{xy} \\ \sigma_{yy} \end{pmatrix}. \tag{4.4.2}$$

After application of Gauss' divergence theorem we obtain:

$$\oint_{\Gamma_1} \boldsymbol{s}_x \cdot \boldsymbol{n}\, d\Gamma + \int_{V_1} b_1\, dV = 0, \tag{4.4.3a}$$

$$\oint_{\Gamma_2} \boldsymbol{s}_y \cdot \boldsymbol{n}\, d\Gamma + \int_{V_2} b_2\, dV = 0, \tag{4.4.3b}$$

or

$$\int_{e_1} \sigma_{xx}\, dy - \int_{w_1} \sigma_{xx}\, dy + \int_{n_1} \tau_{xy}\, dx - \int_{s_1} \tau_{xy} dx = -h_x h_y b_1, \tag{4.4.4a}$$

$$\int_{e_2} \tau_{xy}\, dy - \int_{w_2} \tau_{xy}\, dy + \int_{n_2} \sigma_{yy}\, dx - \int_{s_2} \sigma_{yy} dx = -h_x h_y b_2. \tag{4.4.4b}$$

It is not self-evident that the control volumes for the two force components should be the same for Equation (4.4.4a) and Equation (4.4.4b) and in fact we shall see that a very natural choice will make them different.

4.4.1 Staggered grids

We apply the finite volume method with volume V_1 to Equation (4.4.4a) and we express the stress tensor components in the *displacements* u and v. In e_1 we now need to have $\partial u/\partial x$ and $\partial v/\partial y$, so in fact we would like to have u_E, u_C, v_{ne} and v_{se} in order to make compact central differences around e_1. Checking the rest of the sides of V_1 makes it clear that we need: u_E, u_S, u_W, u_N, u_C and $v_{ne}, v_{nw}, v_{sw}, v_{se}$, see Figure 4.13.

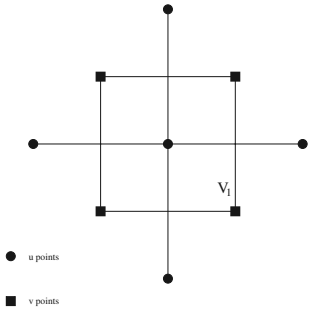

Figure 4.13: V_1-variables.

Exercise 4.4.1 *Derive the discretization in the* displacement *variables u and v for Equation* (4.4.4a) *in the V_1 volume.* □

When we apply FVM with volume V_2 to Equation (4.4.4b) we need $\partial u/\partial y$ and $\partial v/\partial x$ in e_2, so now we would like to have v_E, v_C, u_{ne} and u_{se}.

Exercise 4.4.2 *Derive the discretization in the* displacement *variables u and v for Equation* (4.4.4b) *in the V_2 volume.* □

So apparently we must choose a grid in such a way that both V_1 and V_2 can be accommodated and the natural way to do that is take u and v in different nodal points, like in Figure 4.14.

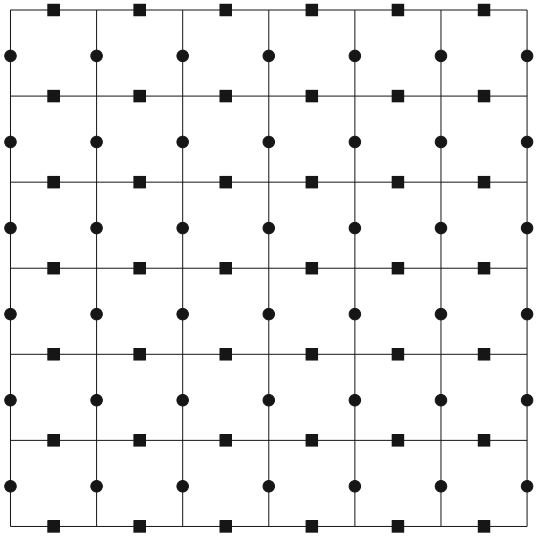

Figure 4.14: Staggered grid.

Such an arrangement of nodal point is called a *staggered grid*. This means that in general different problem variables reside in different nodes.

4.4.2 Boundary conditions

When discretizing a scalar equation you can often choose the grid in such a fashion that the boundary conditions can be easily implemented. With two or more components, especially on a staggered grid, this is no longer true.

Consider the *W*-boundary of our fixed plate in Figure 4.12. On this boundary we have the boundary conditions $u = 0$ and $v = 0$. A quick look at the staggered grid of Figure 4.14 shows a fly in the ointment. The u-points are on the boundary all right. Let us distinguish between equations derived from Equation (4.4.4a) (type 1) and those derived from Equation (4.4.4b) (type 2). In equations of type 1 you can easily implement the boundary conditions on the W-boundary. By the same token, you can easily implement the boundary condition on the N-boundary in type 2 equations. For equations of the "wrong" type you have to resort to a trick. The generic form of an equation of type 2 in

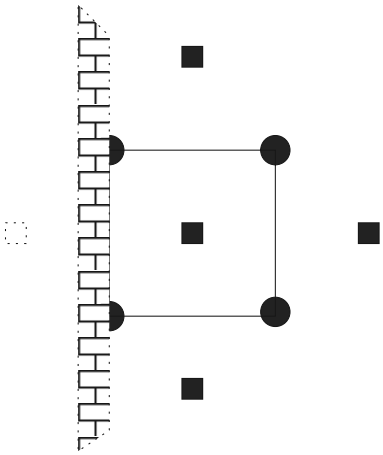

Figure 4.15: Ghost point.

the displacement variables is:

$$B_W v_W + B_{nw} u_{nw} + B_N v_N + B_{ne} u_{ne} + B_E v_E + B_{se} u_{se} + B_S v_S + B_{sw} u_{sw} + B_C v_C = h^2 b_C. \quad (4.4.5)$$

To implement the boundary condition on the W-side in equations of type 2, we assume a virtual ("ghost") grid point on the other side of the wall acting as W-point, see Figure 4.15

Now we eliminate v_W by linear interpolation: $(v_W + v_C)/2 = 0$, hence $v_W = -v_C$ and Equation (4.4.5) transforms into

$$B_{nw} u_{nw} + B_N v_N + B_{ne} u_{ne} + B_E v_E + B_{se} u_{se} + B_S v_S + B_{sw} u_{sw} + (B_C - B_W) v_C = h^2 b_C. \quad (4.4.6)$$

Exercise 4.4.3 *Explain how to implement the boundary condition on the N-boundary in equations of type 1.* □

The boundary conditions on the E- and S boundary are natural boundary conditions. When a boundary of a full volume coincides with such a boundary, there are no problems, the boundary condition can be substituted directly. That is equations of type 2 are easy at the E-boundary, equations of type 1 are easy at the S-boundary.

Exercise 4.4.4 *Derive the equation of type 1 at the S-boundary in the displacements and substitute the natural boundary condition.* □

What of the half volumes? Consider an equation of type 1 at the E-boundary, see Figure 4.16.

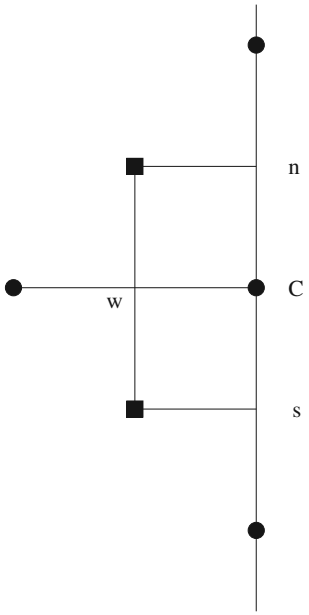

Figure 4.16: Half volume at natural boundary.

Let us integrate Equation (4.4.1a) over a half volume V_1 to obtain:

$$h(-\sigma_{xxw} + \sigma_{xxC}) + \frac{1}{2}h(\tau_{xyn} - \tau_{xys}) = -\frac{1}{2}h^2 b_{1C}. \tag{4.4.7}$$

Since by the natural boundary conditions $\sigma_{xx} = f_1$ and $\tau_{xy} = f_2$ are given quantities at the boundary this transforms into

$$h\sigma_{xxw} = hf_{1C} + \frac{1}{2}h(f_{2n} - f_{2s}) + \frac{1}{2}h^2 b_{1C}. \tag{4.4.8}$$

Again one-point integration of the right-hand side causes a perturbation of $\mathcal{O}(h^3)$, because it is not in the gravicenter of the volume, and also the integration along the n- and s-sides of the volume has an error of $\mathcal{O}(h^3)$.

Exercise 4.4.5 *Prove these last two assertions.* □

Since this perturbation is of the same order as a perturbation of $\mathcal{O}(h^2)$ in the stresses applied at the boundary, we may expect that this gives a perturbation of the same order in the displacements u and v.

4.5 Stokes equations for incompressible flow

A fairly simple and admittedly artificial model for stationary viscous incompressible flow is represented by the *Stokes equations*:

$$-\operatorname{div} \mu \operatorname{grad} u + \frac{\partial p}{\partial x} = 0 \tag{4.5.1a}$$

$$-\operatorname{div} \mu \operatorname{grad} v + \frac{\partial p}{\partial y} = 0 \tag{4.5.1b}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{4.5.1c}$$

In these equations the first two ones describe the equilibrium of the viscous stresses, the third equation is the incompressibility condition. The viscosity μ is a given material constant, but the velocities u and v and the pressure p have to be calculated. Let us consider this problem in a straight channel (see Figure 4.17).

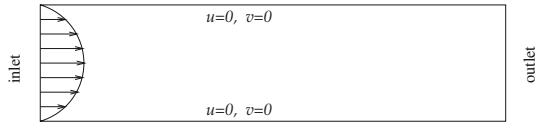

Figure 4.17: Channel for Stokes flow.

At the inlet the velocities are given: $u = u_0(y), v = v_0(y)$, the channel walls allow no slip, so $u = 0$ and $v = 0$ at both walls. At the outlet there is a reference pressure p_0 in the natural boundary conditions: $-\mu\frac{\partial u}{\partial x} + p = p_0$ and $\frac{\partial v}{\partial x} = 0$.

To solve the equations, we use a staggered approach, in which the unknowns are ordered as in Figure 4.18. For the horizontal component of the velocity u, the finite volume method gives

$$-\int_{\Omega_u} \nabla \cdot (\mu \nabla u) d\Omega + \int_{\Omega_u} \frac{\partial p}{\partial x} d\Omega = 0, \tag{4.5.2}$$

where Ω_u is a control volume with a u-node as the center. The divergence theorem yields

$$-\int_{\Gamma_u} \mu \frac{\partial u}{\partial n} d\Gamma + \int_{\Gamma_u} p n_x d\Gamma = 0. \tag{4.5.3}$$

This equation is discretized by similar procedures as the Laplace equation. Note that n_x represents the horizontal component of the unit outward normal vector. The equation for the vertical component of the velocity is worked out

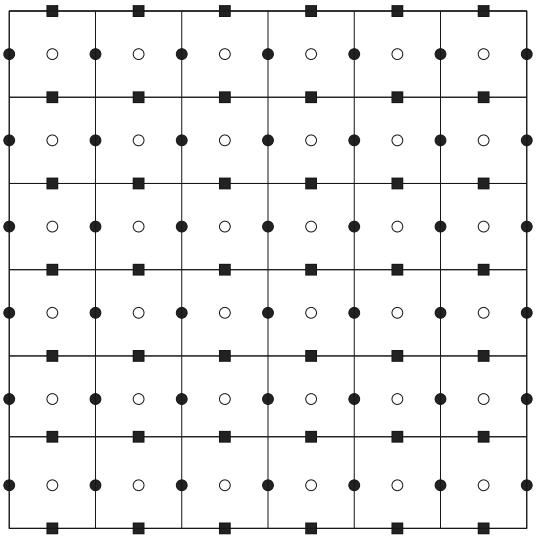

Figure 4.18: The ordering of the unknowns in a staggered approach for the Stokes equations. The solid circles and squares respectively correspond to u and v indicating the horizontal and vertical components of the fluid velocity. The open circles denote the pressure nodes.

similarly, to get

$$-\int_{\Gamma_v} \mu \frac{\partial v}{\partial n} d\Gamma + \int_{\Gamma_v} p n_y d\Gamma = 0. \tag{4.5.4}$$

Subsequently, we consider the continuity equation $\operatorname{div} \boldsymbol{u} = 0$. This equation is integrated over a control volume with a pressure node as the center:

$$\int_{\Omega_p} \operatorname{div} \boldsymbol{u} \, d\Omega = \int_{\Gamma_p} \boldsymbol{u} \cdot \boldsymbol{n} \, d\Gamma. \tag{4.5.5}$$

For the implementation of the outlet condition $-\mu \frac{\partial u}{\partial x} + p = p_0$, we use half a cell over a u-node, in which the integral over the right (east) boundary Γ_u^R is given by

$$\int_{\Gamma_u^R} \left(-\mu \frac{\partial u}{\partial x} + p n_x \right) d\Gamma = \int_{\Gamma_u^R} p_0 d\Gamma \approx p_0 h.$$

Exercise 4.5.1 *Derive discrete equations for all three volumes Ω_u, Ω_v and Ω_p. Note that the pressure and equation of continuity are coupled, that is, the continuity equation is integrated over a pressure cell.* □

Exercise 4.5.2 *Explain how the no-slip boundary conditions are implemented in the equations (*Hint: Use ghost points and averaging in the spirit of Section 4.4.2.*).* □

Exercise 4.5.3 *Explain how to implement the inlet boundary conditions.* □

Exercise 4.5.4 *Take care to end in a vertical line with u points at the outlet. Now explain how to implement the outlet boundary conditions. Argue why you ended up with as many equations as unknowns.* □

Exercise 4.5.5 *In the half Ω_u volume at the outlet boundary the one-point integrations over the horizontal edges cause an error of $\mathcal{O}(h^3)$. Show this and argue that this is equivalent to a perturbation of $\mathcal{O}(h^2)$ in the reference pressure p_0.* □

4.6 Summary of Chapter 4

We have learned a new way to discretize: the *Finite Volume Method*, especially suited to conservation laws. We have seen a one-dimensional and a two-dimensional example with non-uniform stepsizes and radiation boundary conditions. Despite the fact that at the boundary the accurate midpoint integration rule was replaced by less accurate one-point integration, we have shown or made plausible that that would not affect the overall accuracy of the solution. We concluded the chapter with extensive treatment of the Laplacian in curvilinear coordinates and an example of the two-component problem of plane stress. We have seen that for problems of that kind it is sometimes useful to take the variables in different node points: staggered grids.

Chapter 5

Non-linear equations

Objectives

The discretization of an elliptic PDE leads always to a system of algebraic equations. If the PDE is linear, the resulting system of algebraic equations will be linear too, and the corresponding matrices are generally large and sparse. The efficient numerical solution of these large and sparse linear systems is important, but beyond the scope of this book.

If the PDE is non-linear, the resulting system of algebraic equations will be non-linear too. These non-linear equations are usually solved by a series of linear problems with the same structure. Although many methods to solve non-linear algebraic systems are available in the mathematical literature, we will only treat two classical iterative processes: *Picard iteration* and *Newton iteration*. These two methods usually exhibit linear and quadratic convergence, respectively.

5.1 Picard iteration

First we consider a class of problems that are small perturbations of linear problems. For instance

$$-\text{div grad}\, u = f(u), \quad \text{on } \Omega, \tag{5.1.1}$$

and $u = 0$ on Γ. If you discretize this the standard way, you end up with a set of equations of the form

$$A\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{u}), \tag{5.1.2}$$

in which $f_i(\boldsymbol{u}) = f(u_i)$. To approximate the solution of the above equation, we generate a sequence $\boldsymbol{u}^k$ with the goal that $\boldsymbol{u}^k \to \boldsymbol{u}$ as $k \to \infty$. The estimates $\boldsymbol{u}^k$ are obtained by solving a *linear* system of equations. Since we are only able

to solve linear problems such as $A\boldsymbol{u} = \boldsymbol{b}$, a natural way to go about this is to start out with an initial estimate $\boldsymbol{u}^0$ and solve the following iteratively:

$$A\boldsymbol{u}^{k+1} = \boldsymbol{f}(\boldsymbol{u}^k). \tag{5.1.3}$$

Such an iterative process is known as *Picard iteration*.

Exercise 5.1.1 *Show that if $\boldsymbol{u}$ is the solution of* (5.1.2) *and $\boldsymbol{\epsilon}^k = \boldsymbol{u} - \boldsymbol{u}^k$, with $\boldsymbol{u}^k$ the solution of* (5.1.3)*, that*

$$A\boldsymbol{\epsilon}^{k+1} = D(\boldsymbol{u})\boldsymbol{\epsilon}^k + \mathcal{O}(\|\boldsymbol{\epsilon}^k\|^2), \tag{5.1.4}$$

in which D is a diagonal matrix with $d_{ii}(\boldsymbol{u}) = f'(u_i)$. Show that this process cannot converge if at least one eigenvalue of $A^{-1}D$ has a modulus larger than 1. □

Another example concerns the case of an elliptic equation in which the coefficients depend on the solution u. Let us consider the following equation

$$-\text{div } (D(u)\text{grad}\, u) = f(\boldsymbol{x}). \tag{5.1.5}$$

If $D(u)$ is not a constant, for instance $D(u) = u$, then the above equation is nonlinear. To solve the above equation, we generate a sequence of approximations u^k as in the previous example. Here the above equation is solved by iterating

$$-\text{div } \left(D(u^k)\text{grad}\, u^{k+1}\right) = f(\boldsymbol{x}). \tag{5.1.6}$$

After construction of an appropriate discretization, a linear system to obtain u^{k+1} has to be solved. In general if one wants to solve a nonlinear problem using Picard's method, convergence is not always guaranteed. One needs to use common-sense to solve the problem.

So a natural way to obtain an iterative process to a non-linear set of equations $\boldsymbol{F}(\boldsymbol{x}) = 0$ is to reform it to a *fixed point form* $\boldsymbol{x} = \boldsymbol{G}(\boldsymbol{x})$ with the same solution. On this fixed point form you graft an iterative process called *fixed point iteration* or *Picard iteration*:

$$\boldsymbol{x}^{k+1} = \boldsymbol{G}(\boldsymbol{x}^k). \tag{5.1.7}$$

There is a famous convergence result due to Banach on such processes.

Theorem 5.1.1 *Let $\mathcal{D}$ be a closed non-empty subset of $\mathbb{R}^n$ and let $\boldsymbol{G}$: $\mathcal{D} \to \mathbb{R}^n$ be a mapping such that*

(i) If $\boldsymbol{x} \in \mathcal{D}$ then $\boldsymbol{G}(\boldsymbol{x}) \in \mathcal{D}$;

(ii) There exists an $\alpha \in [0, 1)$ such that $\|\boldsymbol{G}(\boldsymbol{x}) - \boldsymbol{G}(\boldsymbol{y})\| \leq \alpha\|\boldsymbol{x} - \boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$.

Then $\mathcal{D}$ contains precisely one fixed point of $\boldsymbol{G}$.

Proof
Choose $\boldsymbol{x}^0 \in \mathcal{D}$. Then it follows from (i) that the sequence $(\boldsymbol{x}^k)_{k=0}^{\infty}$ defined by (5.1.7) lies in $\mathcal{D}$. To prove convergence of a sequence in a finite-dimensional space like $\mathbb{R}^n$, it is sufficient to show that this sequence is a so-called Cauchy sequence, that is,

$$\lim_{k,\ell\to\infty} \|\boldsymbol{x}^k - \boldsymbol{x}^\ell\| = 0. \tag{5.1.8}$$

As a first step, note that

$$\|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\| = \|\boldsymbol{G}(\boldsymbol{x}^j) - \boldsymbol{G}(\boldsymbol{x}^{j-1})\| \leq \alpha\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\| \leq \ldots \leq \alpha^j\|\boldsymbol{x}^1 - \boldsymbol{x}^0\|.$$

As a second step, one can repeatedly make use of the triangle inequality ($\|u + v\| \leq \|u\| + \|v\|$) to show that for all k, ℓ with $k \geq \ell \geq 0$ that

$$\|\boldsymbol{x}^k - \boldsymbol{x}^\ell\| \leq \sum_{j=\ell}^{k-1} \|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\|. \tag{5.1.9}$$

Combining both steps we obtain for all k, ℓ with $k \geq \ell \geq 0$ that

$$\|\boldsymbol{x}^k - \boldsymbol{x}^\ell\| \leq \sum_{j=\ell}^{k-1} \alpha^j\|\boldsymbol{x}^1 - \boldsymbol{x}^0\| \leq \frac{\alpha^\ell}{1-\alpha}\|\boldsymbol{x}^1 - \boldsymbol{x}^0\|,$$

from which the Cauchy property (5.1.8) immediately follows. We may conclude that the sequence converges to a limit $\boldsymbol{\xi} \in \mathbb{R}^n$, which must lie in $\mathcal{D}$ since $\mathcal{D}$ is closed. Note that it follows from $\|\boldsymbol{G}(\boldsymbol{x}^k) - \boldsymbol{G}(\boldsymbol{\xi})\| \leq \alpha\|\boldsymbol{x}^k - \boldsymbol{\xi}\|$ that $\boldsymbol{G}(\boldsymbol{x}^k)$ converges to $\boldsymbol{G}(\xi)$, so $\boldsymbol{\xi}$ is a fixed point of $\boldsymbol{G}$. Finally, there cannot be two different fixed points $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. If there were, then

$$\|\boldsymbol{\xi} - \boldsymbol{\eta}\| = \|\boldsymbol{G}(\boldsymbol{\xi}) - \boldsymbol{G}(\boldsymbol{\eta})\| \leq \alpha\|\boldsymbol{\xi} - \boldsymbol{\eta}\|,$$

which is clearly impossible since $\alpha < 1$. □

Exercise 5.1.2 *Prove (5.1.9) by repeatedly using the triangle inequality.* □

A mapping that satisfies the conditions of Theorem 5.1.1 is called a *contraction* or a *contractive mapping* on the set $\mathcal{D}$.

Exercise 5.1.3 *Let $\boldsymbol{G}$: $\mathbb{R}^n \to \mathbb{R}^n$ be a mapping with fixed point $\boldsymbol{\xi}$. Assume that $\boldsymbol{G}$ has continuous partial derivatives in a neighborhood $\mathcal{E}$ of $\boldsymbol{\xi}$. Further assume that $\|G'(\boldsymbol{x}))\| < 1$, $\boldsymbol{x} \in \mathcal{E}$, where G' is the matrix with elements*

$$g'_{ij} = \frac{\partial g_i}{\partial x_j}. \tag{5.1.10}$$

Show that $\mathcal{E}$ contains a closed neighborhood $\mathcal{D}$ of $\boldsymbol{\xi}$ on which $\boldsymbol{G}$ is a contraction. □

5.2 Newton's method in more dimensions

In order to find a faster converging solution process to the set of non-linear equations

$$\boldsymbol{F}(\boldsymbol{x}) = 0, \quad \boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n, \quad \boldsymbol{x} \in \mathbb{R}^n \tag{5.2.1}$$

we try to find an analogue to Newton's method for functions of one variable:

$$x^{k+1} = x^k - F(x^k)/F'(x^k). \tag{5.2.2}$$

In the neighborhood of the root ξ we have by Taylor's theorem:

$$0 = F(\xi) = F(x) + (\xi - x)F'(x) + \mathcal{O}((\xi - x)^2), \tag{5.2.3}$$

for functions of one variable. We arrive at Newton's formula by neglecting the second order term. We try something similar in n dimensions. In the neighborhood of the root $\boldsymbol{\xi}$ we have:

$$0 = f_1(\boldsymbol{\xi}) = f_1(\boldsymbol{x}) + \sum_{j=1}^{n} (\xi_j - x_j) \frac{\partial f_1}{\partial x_j}(\boldsymbol{x}) + \mathcal{O}(\|\boldsymbol{\xi} - \boldsymbol{x}\|^2), \tag{5.2.4a}$$

$$0 = f_2(\boldsymbol{\xi}) = f_2(\boldsymbol{x}) + \sum_{j=1}^{n} (\xi_j - x_j) \frac{\partial f_2}{\partial x_j}(\boldsymbol{x}) + \mathcal{O}(\|\boldsymbol{\xi} - \boldsymbol{x}\|^2), \tag{5.2.4b}$$

$$\vdots$$

$$0 = f_n(\boldsymbol{\xi}) = f_n(\boldsymbol{x}) + \sum_{j=1}^{n} (\xi_j - x_j) \frac{\partial f_n}{\partial x_j}(\boldsymbol{x}) + \mathcal{O}(\|\boldsymbol{\xi} - \boldsymbol{x}\|^2). \tag{5.2.4c}$$

Neglecting the second order terms in Equations (5.2.4) we arrive at an iteration process that is analogous to (5.2.2):

$$f_1(\boldsymbol{x}^k) + \sum_{j=1}^{n} (x_j^{k+1} - x_j^k) \frac{\partial f_1}{\partial x_j}(\boldsymbol{x}^k) = 0, \tag{5.2.5a}$$

$$f_2(\boldsymbol{x}^k) + \sum_{j=1}^{n} (x_j^{k+1} - x_j^k) \frac{\partial f_2}{\partial x_j}(\boldsymbol{x}^k) = 0, \tag{5.2.5b}$$

$$\vdots$$

$$f_n(\boldsymbol{x}^k) + \sum_{j=1}^{n} (x_j^{k+1} - x_j^k) \frac{\partial f_n}{\partial x_j}(\boldsymbol{x}^k) = 0. \tag{5.2.5c}$$

We can put this into vector notation:

$$F'(\boldsymbol{x}^k)(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) = -\boldsymbol{F}(\boldsymbol{x}^k), \tag{5.2.6}$$

where $F'(\boldsymbol{x})$ is the Jacobian matrix

$$F'(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix} (\boldsymbol{x}). \tag{5.2.7}$$

We now present the algorithmic form.

Newton's method for multivariate functions

1: Presets: $\boldsymbol{x}^0$ {initial estimate}, $\boldsymbol{r}^0 = \boldsymbol{F}(\boldsymbol{x}^0)$, $k = 0$
2: **while** $\|\boldsymbol{r}^k\| > \varepsilon$ **do**
3: Solve $F'(\boldsymbol{x}^k)\boldsymbol{c}^k = -\boldsymbol{r}^k$
4: $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \boldsymbol{c}^k$
5: $\boldsymbol{r}^{k+1} = \boldsymbol{F}(\boldsymbol{x}^{k+1})$
6: $k = k + 1$
7: **end while**

The calculation of the Jacobian matrix is often very time consuming and various schemes have been proposed to improve on that. For the solution of the linear system in line 3 we can use any type of solver. The Jacobian matrix often has the same sparsity pattern as the corresponding linearization of the PDE.

Example 5.2.1 *We consider the following differential equation in one spatial dimension, with homogeneous Dirichlet boundary conditions:*

$$u(1-u)\frac{d^2u}{dx^2} + x = 0, \qquad u(0) = u(1) = 0. \tag{5.2.8}$$

A finite difference discretization, with uniform grid-spacing h and n unknowns ($h = 1/(n+1)$), gives

$$F_i(\boldsymbol{u}) = u_i(1-u_i)\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + x_i = 0, \quad \textit{for } i \in \{1, \ldots, n\}. \tag{5.2.9}$$

Note that for $i = 1$ and $i = n$, the boundary conditions must be used. This system of n equations with n unknowns is seen as a system of non-linear equations. Using the Picard fixed point or Newton method requires an initial guess for the solution. This initial guess could be chosen by solving the linearized system or by choosing a vector that reflects the values at a Dirichlet boundary. In this particular case $\boldsymbol{u} = \boldsymbol{0}$ is not a good initial guess since the Jacobian matrix would be completely zero (why?). Let $\boldsymbol{u}^k$ represent the solution at the k-th iterate, then, one way of using the Picard fixed point method is the following:

$$u_i^k(1-u_i^k)\frac{u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}}{h^2} + x_i = 0, \quad \textit{for } i \in \{1, \ldots, n\}. \tag{5.2.10}$$

This requires the solution of a system of linear equations at each iteration.

If one prefers to use Newton's method, then the calculation of the Jacobian matrix is necessary. Considering the i-th row of the Jacobian matrix, all entries are zero, except the one on and the ones adjacent to the main diagonal, that is

$$\frac{\partial f_i}{\partial u_{i-1}}(\boldsymbol{u}^k) = \frac{u_i^k(1-u_i^k)}{h^2},$$

$$\frac{\partial f_i}{\partial u_i}(\boldsymbol{u}^k) = -\frac{2u_i^k(1-u_i^k)}{h^2} + (1-2u_i^k)\frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h^2}, \tag{5.2.11}$$

$$\frac{\partial f_i}{\partial u_{i+1}}(\boldsymbol{u}^k) = \frac{u_i^k(1-u_i^k)}{h^2}.$$

The rest of the procedure is straightforward. □

Exercise 5.2.1 *Consider on the square* $(0,1)\times(0,1)$ *the discretization of*

$$-\operatorname{div}\operatorname{grad} u = e^u. \tag{5.2.12}$$

Calculate $F'(\boldsymbol{u})$*. Compare the structure of the Jacobian matrix to that of the matrix generated by the discretization of the Laplacian.* □

Exercise 5.2.2 *Consider on the square* $(0,1)\times(0,1)$ *the discretization of*

$$\operatorname{div}\left(\frac{\operatorname{grad} u}{\sqrt{1+u_x^2+u_y^2}}\right) = 0, \tag{5.2.13}$$

by the finite volume method. What is the sparsity structure of $F'(\boldsymbol{u})$*?* □

5.2.1 Starting values

Although Newton's method converges quadratically in a neighborhood of the root, convergence is often very sensitive to good initial estimates. These are suggested sometimes by the technical context, but if obtaining an initial estimate appears to be a problem, the following trick, known as the *homotopy method*, may be applied.

Suppose the solution to some other problem, say $\boldsymbol{F}_0(\boldsymbol{x}) = 0$ is known (e.g. a linearization of the original). Consider the following set of problems:

$$(1-\lambda)\boldsymbol{F}_0(\boldsymbol{x}) + \lambda\boldsymbol{F}(\boldsymbol{x}) = 0, \quad \lambda\in[0,1]. \tag{5.2.14}$$

For $\lambda = 1$ we have our original problem, for $\lambda = 0$ we have our auxiliary problem. Now the idea is to proceed in small steps h from $\lambda_0 = 0, \lambda_1 = h, \lambda_2 = 2h$ to $\lambda_N = Nh = 1$, using Newton's method as solver and always taking the solution to the problem with λ_k as initial estimate to the problem with λ_{k+1}. This is an expensive method but somewhat more robust than simple Newton.

5.3 Summary of Chapter 5

In this chapter we have studied methods to solve non-linear sets of equations. We looked at *Picard iteration* and *Newton's method*. The *homotopy* method can be used to find a starting value if all other inspiration fails.

Chapter 6

The heat- or diffusion equation

Objectives

In this chapter several numerical methods to solve the heat equation are considered. Since this equation also describes diffusion, the equation is referred to as the diffusion equation. The equation describes very common processes in physics and engineering and we would like our numerical models to inherit certain properties of the physics. The most important aspect - and typical for diffusion equations - is the property that the solution generally tends to an equilibrium solution as time proceeds. If the coefficients in the heat equation and the boundary conditions do not depend on time, there exists exactly one equilibrium solution (unless the whole boundary is a Neumann boundary), and the solution of the heat equation tends to this equilibrium solution independent of the initial condition. If the whole boundary is a Neumann boundary then the situation is more complicated.

6.1 A fundamental inequality

The next theorem states this result more precisely.

Theorem 6.1.1 *Let Ω be a bounded domain in $\mathbb{R}^2$ with a boundary Γ consisting of 3 parts Γ_1, Γ_2 and Γ_3. One or more of these parts may be empty, but $\Gamma_2 \neq \Gamma$. Let Δ be given by*

$$\Delta = \text{div grad } = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \tag{6.1.1}$$

Let $u_E(\boldsymbol{x})$ be the solution of

$$\Delta u + f(\boldsymbol{x}) = 0, \tag{6.1.2}$$

with boundary conditions

$$u(\boldsymbol{x}) = g_1(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_1, \tag{6.1.3}$$

$$\frac{\partial u}{\partial n}(\boldsymbol{x}) = g_2(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_2, \tag{6.1.4}$$

$$(\sigma u)(\boldsymbol{x}) + \frac{\partial u}{\partial n}(\boldsymbol{x}) = g_3(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_3. \tag{6.1.5}$$

Further, let $u(\boldsymbol{x}, t)$ be the solution of the initial value problem

$$\frac{\partial u}{\partial t} = \Delta u + f(\boldsymbol{x}), \tag{6.1.6}$$

with initial condition $u(\boldsymbol{x}, t_0) = u_0(\boldsymbol{x})$ and boundary conditions (6.1.3)–(6.1.5). Let $R(t)$ be the quadratic residual, which is

$$R(t) = \int_\Omega \left(u(\boldsymbol{x}, t) - u_E(\boldsymbol{x})\right)^2 d\Omega. \tag{6.1.7}$$

Then there is a $\gamma > 0$ such that

$$R(t) \leq R(t_0) e^{-\gamma(t-t_0)}, \quad \forall t > t_0. \tag{6.1.8}$$

Proof
Note that u_E is a solution of (6.1.6) with $\partial u_E / \partial t = 0$. The difference $v = u_E - u$ therefore satisfies

$$\frac{\partial v}{\partial t} = \Delta v, \tag{6.1.9}$$

with initial condition $v(\boldsymbol{x}, t_0) = u_E(\boldsymbol{x}) - u_0(\boldsymbol{x})$ and boundary conditions

$$v(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Gamma_1, \tag{6.1.10}$$

$$\frac{\partial v}{\partial n}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Gamma_2, \tag{6.1.11}$$

$$(\sigma v)(\boldsymbol{x}) + \frac{\partial v}{\partial n}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Gamma_3. \tag{6.1.12}$$

Multiplication of Equation (6.1.9) by v and subsequent integration over Ω, gives

$$\int_\Omega v \frac{\partial v}{\partial t}\, d\Omega = \int_\Omega v \Delta v\, d\Omega,$$

which can be rewritten as

$$\int_\Omega \frac{1}{2} \frac{\partial v^2}{\partial t}\, d\Omega = -\int_\Omega \|\mathrm{grad}\, v\|^2\, d\Omega + \int_\Gamma v \frac{\partial v}{\partial n}\, d\Gamma.$$

Here the right-hand side follows from Green's theorem 1.3.3. We interchange the order of integration over Ω, differentiate with respect to time and apply the boundary conditions to get:

$$\frac{1}{2}\frac{dR}{dt} = -\int_{\Omega} \|\text{grad}\, v\|^2 \, d\Omega - \int_{\Gamma_3} \sigma v^2 \, d\Gamma. \tag{6.1.13}$$

According to Poincaré's Lemma [1], which is an extension of Theorem 1.5.1 to the case of more general boundary conditions, there exists a $\gamma_0 > 0$ such that

$$\int_{\Omega} \|\text{grad}\, v\|^2 \, d\Omega \geq \gamma_0 \int_{\Omega} v^2 \, d\Omega = \gamma_0 R. \tag{6.1.14}$$

Letting $\gamma = 2\gamma_0$ we obtain

$$\frac{dR}{dt} \leq -\gamma R, \tag{6.1.15}$$

that is,

$$\frac{dR}{dt} + \gamma R \leq 0. \tag{6.1.16}$$

This inequality holds for all $t > t_0$. We multiply this inequality by $e^{\gamma t}$ to get

$$e^{\gamma t}\left(\frac{dR}{dt} + \gamma R\right) = \frac{d\left(e^{\gamma t} R\right)}{dt} \leq 0. \tag{6.1.17}$$

After integration from t_0 to t this yields

$$e^{\gamma t} R(t) - e^{\gamma t_0} R(t_0) \leq 0, \tag{6.1.18}$$

so that

$$R(t) \leq e^{-\gamma(t-t_0)} R(t_0). \tag{6.1.19}$$

This proves the theorem. □

Remarks

1. The quadratic residual $R(t)$ tends to zero exponentially. Hence the time-dependent solution tends to the equilibrium solution exponentially.

2. If a *Neumann* boundary condition is given on the *entire* boundary, a compatibility condition (which?) has to be satisfied in order that a physical equilibrium is possible. For this particular case the conditions of the theorem have to be adapted and the physical equilibrium depends on the initial condition. If the compatibility condition condition is not satisfied, the solution of the time-dependent problem is unbounded. Depending on the sign of the net heat production, the temperature goes to $\pm\infty$.

3. This theorem, proved for the Laplace operator, also holds for the general elliptic operator

$$L = \sum_{\alpha}^{n} \sum_{\beta}^{n} \frac{\partial}{\partial x_\alpha} K_{\alpha\beta} \frac{\partial}{\partial x_\beta},$$

with K positive definite.

4. In a similar way, it is possible to establish *analytical stability* for this problem, i.e., one can demonstrate well-posedness in relation to the initial conditions: Given two solutions u and v with initial conditions u_0 and $u_0 + \epsilon_0$ respectively, then, for $\epsilon(\boldsymbol{x}, t) = (v - u)(\boldsymbol{x}, t)$, we have

$$\left(\int_\Omega \epsilon^2 \, d\Omega \right)(t) \leq e^{-\gamma(t-t_0)} \int_\Omega \epsilon_0^2 \, d\Omega. \tag{6.1.20}$$

Hence, for this problem, we have *absolute (or asymptotic) stability*, because the error tends to zero as $t \to \infty$.

□

Exercise 6.1.1 *Prove Theorem 6.1.1 for the general elliptic operator mentioned in Remark 3 above.*
Hint: For any real symmetric matrix K it follows from Theorem 1.4.2 that we have $(K\boldsymbol{x}, \boldsymbol{x}) \geq \lambda_1(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, *where* λ_1 *denotes the smallest eigenvalue of K.* □

Exercise 6.1.2 *Prove the stability estimate* (6.1.20) *mentioned in Remark 4 above.* □

6.2 Method of lines

A very general method to solve time-dependent problems is the *method of lines*. In this method we start with the *spatial* discretization of the problem

$$\frac{\partial u}{\partial t} = \Delta u + f. \tag{6.2.1}$$

This spatial discretization can be based on Finite Differences or Finite Volumes. It can also be based on Finite Elements (FEM), but we limit ourselves to FDM and FVM in this book. The spatial discretization results in a system of ordinary differential equations the size of which is determined by the number of parameters used to approximate u. Formally, this system can be written as

$$M \frac{d\boldsymbol{u}_h}{dt} = S\boldsymbol{u}_h + \boldsymbol{f}_h. \tag{6.2.2}$$

The quantities with index $_h$ represent the discrete approximations of the continuous quantities. Note the matrix M, the *mass matrix*, in the left-hand side. It

is the identity matrix in Finite Differences, and a (positive) diagonal matrix in Finite Volumes. In case of Finite Elements, which is beyond the scope of this book, the matrix M is generally not diagonal. The mass matrix M represents the scaling of the equations in the discretization. The matrix S is a (possibly scaled) discrete representation of the elliptic operator L. We will refer to S as the *stiffness matrix*, but would like to point out that in the literature this name is also often reserved for $-S$. We illustrate the method with a few examples.

6.2.1 One-dimensional examples

In this section we consider the following equation with one space coordinate:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \quad x \in [0,1], \tag{6.2.3}$$

with initial condition $u(x,t_0) = u_0(x)$. We look at two different discretization methods.

Example 6.2.1 FDM, Dirichlet
We use as boundary conditions: $u(0) = u(1) = 0$. Similarly as in Chapter 3, the interval $(0,1)$ is divided into sub-intervals of size h, such that $Nh = 1$. The second order derivative is discretized using the second divided difference in each internal grid node x_j, $j = 1,2,\ldots,N-1$. In each grid node x_j, $j = 0,\ldots,N$, there is a u_j, which, of course, also depends on time. From the boundary conditions it follows that $u_0 = 0 = u_N$, so the remaining unknowns are $u_1,\ldots,u_{N-1}$. After elimination of u_0 and u_N we obtain the following system of ordinary differential equations:

$$\frac{d\boldsymbol{u}_h}{dt} = S\boldsymbol{u}_h + \boldsymbol{f}_h, \tag{6.2.4}$$

with

$$S = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 & \ldots & \ldots & 0 \\ 1 & -2 & 1 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 1 & -2 & 1 \\ 0 & \ldots & \ldots & 0 & 1 & -2 \end{pmatrix}, \tag{6.2.5}$$

$$\boldsymbol{u}_h = \begin{pmatrix} u_1 \\ \vdots \\ u_{N-1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{f}_h = \begin{pmatrix} f_1 \\ \vdots \\ f_{N-1} \end{pmatrix}, \tag{6.2.6}$$

in which $\boldsymbol{u}_h$ and $\boldsymbol{f}_h$ both depend on t. □

Example 6.2.2 FVM, right-hand boundary point Neumann
We take as boundary conditions $u(0) = 0$, $u'(1) = 0$. Further, a non-equidistant grid is used with $N+1$ grid nodes, and $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, N$. As a control volume around the node x_i, $i = 1, 2, \ldots, N-1$, the interval $V_i = (x_{i-1/2}, x_{i+1/2})$ is used, with $x_{i-1/2} = \frac{x_i + x_{i-1}}{2}$ and $x_{i+1/2} = \frac{x_{i+1} + x_i}{2}$. Integration of the differential equation over the control volume gives

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \frac{\partial u}{\partial t} dx = \int_{x_{i-1/2}}^{x_{i+1/2}} (\frac{\partial^2 u}{\partial x^2} + f) dx, \tag{6.2.7}$$

and therefore

$$\frac{d}{dt} \int_{x_{i-1/2}}^{x_{i+1/2}} u dx = \frac{\partial u}{\partial x}\Big|_{x_{i+1/2}} - \frac{\partial u}{\partial x}\Big|_{x_{i-1/2}} + \int_{x_{i-1/2}}^{x_{i+1/2}} f dx. \tag{6.2.8}$$

For the integrals

$$\int_{x_{i-1/2}}^{x_{i+1/2}} u dx \quad \text{and} \quad \int_{x_{i-1/2}}^{x_{i+1/2}} f dx$$

the mid-point rule will be used. The Neumann boundary condition is treated by integrating the differential equation over the control volume $V_N = (x_{N-1/2}, x_N)$ and proceeding as in Section 4.1.1, with $\lambda = 1$ and $\alpha = 0$. □

Exercise 6.2.1 *Show that the discretization of this problem can be written as*

$$M \frac{d\boldsymbol{u}_h}{dt} = S\boldsymbol{u}_h + \boldsymbol{f}_h, \tag{6.2.9}$$

where the mass matrix M and the stiffness matrix S are given by

$$M = \begin{pmatrix} \frac{1}{2}(h_1 + h_2) & & & & \\ & \frac{1}{2}(h_2 + h_3) & & & \\ & & \ddots & & \\ & & & \frac{1}{2}(h_{N-1} + h_N) & \\ & & & & \frac{1}{2}h_N \end{pmatrix}, \tag{6.2.10}$$

$$S = \begin{pmatrix} -\frac{1}{h_1} - \frac{1}{h_2} & \frac{1}{h_2} & & & \\ \frac{1}{h_2} & -\frac{1}{h_2} - \frac{1}{h_3} & \frac{1}{h_3} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{h_{N-1}} & -\frac{1}{h_{N-1}} - \frac{1}{h_N} & \frac{1}{h_N} \\ & & & \frac{1}{h_N} & -\frac{1}{h_N} \end{pmatrix}. \tag{6.2.11}$$

□

6.2.2 Two-dimensional example

In this section we consider the 2D heat equation with source term,

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f(\boldsymbol{x}, t), \quad \boldsymbol{x} \in \Omega = [0,1] \times [0,1]. \tag{6.2.12}$$

We assume homogeneous Dirichlet boundary conditions $u|_\Gamma = 0$ and an initial condition $u(\boldsymbol{x}, t_0) = u_0(\boldsymbol{x})$.

Example 6.2.3 FDM, Dirichlet
Following Section 3.4 (with $M = N$) we divide Ω into N^2 small squares with sides $\Delta x = \Delta y = h$ with $Nh = 1$. At each grid node (x_i, y_j) (or (i, j) for short) there is an unknown $u_{i,j}$, $i, j = 0, 1, \ldots, N$, which, of course, also depends on time.

The Laplacian is discretized using second divided differences (see (3.4.2)), which leads to the following differential equation in all internal grid nodes (i, j) with $i, j = 1, 2, \ldots, N-1$:

$$\frac{du_{i,j}}{dt} = \frac{1}{h^2}\left[-4u_{i,j} + u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}\right] + f_{i,j}, \tag{6.2.13}$$

where $f_{i,j} = f(x_i, y_j)$.

From the boundary conditions it follows that $u_{i,j} = 0$ for all boundary nodes (i, j). For the remaining $(N-1)^2$ unknowns $u_{i,j}$ (corresponding to the internal nodes) we obtain a system of ordinary differential equations of the form

$$\frac{d\boldsymbol{u}}{dt} = S\boldsymbol{u} + \boldsymbol{f}. \tag{6.2.14}$$

The exact definition of $\boldsymbol{u}$, S and $\boldsymbol{f}$ depends on the chosen numbering scheme (cf. Section 3.4.1). □

Exercise 6.2.2 *Show that for horizontal (or vertical) numbering the matrix S in the above example is an $(N-1) \times (N-1)$ block matrix given by*

$$S = \frac{1}{h^2}\begin{pmatrix} T & I & 0 & \ldots & \ldots & 0 \\ I & T & I & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & I & T & I \\ 0 & \ldots & \ldots & 0 & I & T \end{pmatrix}, \tag{6.2.15}$$

where I is the $(N-1) \times (N-1)$ identity matrix and T is the $(N-1) \times (N-1)$ tridiagonal matrix with -4 on its main diagonal and 1 on its first sub- and super-diagonal. □

6.3 Consistency of the spatial discretization

In Chapter 3 we discussed consistency of a discretization of a differential operator. For the FVM discretization of the diffusion equation, it is necessary to include the scaling of the mass matrix M. This means that consistency of the discretization implies that $M^{-1}Sy$ tends to Ly as h tends to zero. In practical situations this can be hard to verify. In order to determine the order of consistency, it suffices to scale each equation from a FVM discretization by the area of the control volume.

We will demonstrate (in Theorem 6.3.1) that the truncation error of the spatial discretization (6.2.2) causes a global error of the same order. We start with substituting the *exact* solution of the heat equation into Equation (6.2.2) of the *discrete* approximation,

$$M\frac{d\boldsymbol{y}}{dt} = S\boldsymbol{y} + \boldsymbol{f} + M\boldsymbol{E}(t), \tag{6.3.1}$$

where $E_k(t) = \mathcal{O}(h^p)$ is the error of the kth equation, which, of course, depends on t. The generic discretization parameter (for instance the diameter of the largest element) is denoted by h and p represents the order of the consistency. In the remaining part of this chapter, the following properties of S and M will be used:

- M and S are real symmetric,
- M is positive definite, S is negative definite (i.e. $(\boldsymbol{x}, S\boldsymbol{x}) < 0$, for $\boldsymbol{x} \neq 0$).

These properties are sufficiently general to include not only FDM and FVM (where M is a diagonal matrix) but also FEM (where M is generally non-diagonal). One easily verifies that these properties imply (see Exercise 6.3.1) that there is a $\gamma_0 > 0$ such that

$$\frac{(S\boldsymbol{x}, \boldsymbol{x})}{(M\boldsymbol{x}, \boldsymbol{x})} \leq -\gamma_0 \text{ for all } \boldsymbol{x} \neq \boldsymbol{0}. \tag{6.3.2}$$

This inequality can be seen as a discrete analogue of Poincaré's inequality (Theorem 1.5.1) and its generalization (6.1.14) used in the proof of Theorem 6.1.1.

Now we will show that the difference between the exact solution of the heat equation and the solution of the system of ordinary differential equations can be bounded in terms of the error $\boldsymbol{E}(t)$. Since M is a positive definite matrix, the expression $\|\boldsymbol{x}\|_M$ defined by $\|\boldsymbol{x}\|_M = \sqrt{(M\boldsymbol{x}, \boldsymbol{x})}$ is a proper vector norm. We formulate our result in this norm.

Theorem 6.3.1 *The difference $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{u}$ between the exact solution, $\boldsymbol{y}$, of the heat equation and the solution, $\boldsymbol{u} = \boldsymbol{u}_h$, of the system of ordinary differential equations (6.2.2), satisfies the following estimate:*

$$\|\boldsymbol{\epsilon}\|_M \leq \frac{1}{\gamma_0} \sup_{t > t_0} \|\boldsymbol{E}(t)\|_M. \tag{6.3.3}$$

Proof
The proof is similar to the proof of the fundamental inequality (6.1.8) of Theorem 6.1.1. We subtract the solution of

$$M\frac{d\boldsymbol{u}}{dt} = S\boldsymbol{u} + \boldsymbol{f},$$

from (6.3.1), to obtain:

$$M\frac{d\boldsymbol{\epsilon}}{dt} = S\boldsymbol{\epsilon} + M\boldsymbol{E}.$$

Since $\boldsymbol{y}$ and $\boldsymbol{u}$ have the same initial condition, we have $\boldsymbol{\epsilon}(t_0) = \boldsymbol{0}$. Taking the inner product of the above equation with $\boldsymbol{\epsilon}$ we get:

$$\begin{aligned}\frac{1}{2}\frac{d(M\boldsymbol{\epsilon},\boldsymbol{\epsilon})}{dt} &= (S\boldsymbol{\epsilon},\boldsymbol{\epsilon}) + (M\boldsymbol{E},\boldsymbol{\epsilon}), \text{ or}\\ \|\boldsymbol{\epsilon}\|_M\frac{d\|\boldsymbol{\epsilon}\|_M}{dt} &= (S\boldsymbol{\epsilon},\boldsymbol{\epsilon}) + (M\boldsymbol{E},\boldsymbol{\epsilon}).\end{aligned}$$

With $(S\boldsymbol{\epsilon},\boldsymbol{\epsilon}) \leq -\gamma_0(M\boldsymbol{\epsilon},\boldsymbol{\epsilon})$ and the Cauchy-Schwarz inequality $(M\boldsymbol{E},\boldsymbol{\epsilon}) \leq \|\boldsymbol{E}\|_M\|\boldsymbol{\epsilon}\|_M$ this transforms into

$$\frac{d\|\boldsymbol{\epsilon}\|_M}{dt} \leq -\gamma_0\|\boldsymbol{\epsilon}\|_M + \|\boldsymbol{E}\|_M,$$

and hence

$$\frac{d}{dt}\left(e^{\gamma_0 t}\|\boldsymbol{\epsilon}\|_M\right) \leq e^{\gamma_0 t}\|\boldsymbol{E}\|_M.$$

We integrate this expression and use $\boldsymbol{\epsilon}_0 = \boldsymbol{0}$ to obtain

$$e^{\gamma_0 t}\|\boldsymbol{\epsilon}\|_M \leq \int_{t_0}^{t} e^{\gamma_0\tau}\|\boldsymbol{E}\|_M\, d\tau.$$

Hence

$$\|\boldsymbol{\epsilon}\|_M \leq \frac{1}{\gamma_0}(1 - e^{-\gamma_0(t-t_0)})\sup_{t>t_0}\|\boldsymbol{E}\|_M,$$

and the theorem follows. □

Exercise 6.3.1 *Prove inequality (6.3.2).*
Hint: Consider

$$\sup_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{(S\boldsymbol{x},\boldsymbol{x})}{(\boldsymbol{x},\boldsymbol{x})}\frac{(\boldsymbol{x},\boldsymbol{x})}{(M\boldsymbol{x},\boldsymbol{x})} \leq \sup_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{(S\boldsymbol{x},\boldsymbol{x})}{(\boldsymbol{x},\boldsymbol{x})}\inf_{\boldsymbol{y}\neq\boldsymbol{0}}\frac{(\boldsymbol{y},\boldsymbol{y})}{(M\boldsymbol{y},\boldsymbol{y})}$$

and apply Theorem 1.4.2 and Corollary 1.4.3. □

Exercise 6.3.2 *Prove the Cauchy-Schwarz inequality*

$$|(M\boldsymbol{x},\boldsymbol{y})| \leq \|\boldsymbol{x}\|_M \|\boldsymbol{y}\|_M \tag{6.3.4}$$

for the inner product $(M\boldsymbol{x},\boldsymbol{y})$ *and associated norm* $\|\boldsymbol{x}\|_M = \sqrt{(M\boldsymbol{x},\boldsymbol{x})}$.
Hint: similar to the proof of the Cauchy-Schwarz inequality (1.4.8) in Exercise 1.4.2. □

Exercise 6.3.3 *Formulate and prove a discrete equivalent of the fundamental inequality* (6.1.8) *of Theorem 6.1.1 for the solution of*

$$M\frac{d\boldsymbol{u}}{dt} = S\boldsymbol{u} + \boldsymbol{f}. \tag{6.3.5}$$

Do the same for the stability estimate (6.1.20). □

6.4 Time integration

The next step we have to take is to integrate in time our system of ordinary differential equations, that we obtained by the method of lines. To this end we use well-known methods for numerical integration of initial value problems, like Euler, improved Euler, Runge-Kutta or the trapezoidal rule.

Example 6.4.1 *Application of Euler's method gives:*

$$M\frac{\boldsymbol{u}^{n+1}}{\Delta t} = M\frac{\boldsymbol{u}^n}{\Delta t} + S\boldsymbol{u}^n + \boldsymbol{f}^n, \tag{6.4.1}$$

in which $\boldsymbol{u}^{n+1}$ *and* $\boldsymbol{u}^n$ *represent the solutions at times* t_{n+1} *and* t_n *respectively, with* $t_n = t_0 + n\Delta t$. □

Exercise 6.4.1 *Formulate the implicit (backward) method of Euler for the system of ordinary differential equations as obtained from the method of lines.* □

Exercise 6.4.2 *Formulate the improved Euler method for this system.* □

Example 6.4.2 *The method of Crank-Nicolson or* the trapezoidal rule *for our system of ordinary differential equations is given by:*

$$(\frac{M}{\Delta t} - \frac{1}{2}S)\boldsymbol{u}^{n+1} = (\frac{M}{\Delta t} + \frac{1}{2}S)\boldsymbol{u}^n + \frac{1}{2}\left(\boldsymbol{f}^n + \boldsymbol{f}^{n+1}\right). \tag{6.4.2}$$

□

Example 6.4.3 *Let* $\theta \in [0,1]$ *be given. The* θ-method *for the system of ordinary differential equations is defined by*

$$(\frac{M}{\Delta t} - \theta S)\boldsymbol{u}^{n+1} = (\frac{M}{\Delta t} + (1-\theta)S)\boldsymbol{u}^n + (1-\theta)\boldsymbol{f}^n + \theta\boldsymbol{f}^{n+1}. \tag{6.4.3}$$

Note that $\theta = 0$, $\theta = 1$ *and* $\theta = \frac{1}{2}$ *correspond to the Forward, Backward Euler and the Crank-Nicolson method respectively.* □

For the θ-method it can be shown that the global error in the time integration is of second order if $\theta = \frac{1}{2}$ and of first order if $\theta \neq \frac{1}{2}$.

6.5 Stability of the numerical integration

In Section 6.1 we demonstrated that the heat equation is *absolutely* stable with respect to the initial conditions (see inequality (6.1.20)). This means that if two solutions have different initial conditions, the difference between these two solutions vanishes as $t \to \infty$. This property also holds for the system of ordinary differential equations obtained by the method of lines (see Exercise 6.3.3). We want to make sure that the numerical time integration inherits this property, so that the numerical time integration is absolutely stable as well. Stability of numerical integration methods in time is treated more extensively in [4]. We state the most important results. The stability of the system of ordinary differential equations

$$\frac{d\boldsymbol{u}}{dt} = A\boldsymbol{u} + \boldsymbol{f}, \tag{6.5.1}$$

is determined by the 'error-equation'

$$\frac{d\boldsymbol{\epsilon}}{dt} = A\boldsymbol{\epsilon}. \tag{6.5.2}$$

1. The system is absolutely stable if and only if the real part of the eigenvalues λ_k of the matrix A is negative, i.e. $\mathrm{Re}(\lambda_k) < 0$.

2. Each numerical solution procedure has an *amplification matrix* $G(\Delta t A)$, given by the numerical solution of (6.5.2):

   $$\boldsymbol{\epsilon}^{n+1} = G(\Delta t A)\boldsymbol{\epsilon}^n. \tag{6.5.3}$$

   If the error equation is *scalar* (i.e. the system reduces to one equation only: $\epsilon' = \lambda\epsilon$), the matrix reduces to an *amplification factor* , which is denoted by $C(\Delta t\lambda)$.

3. A numerical solution method is absolutely stable if all eigenvalues μ_k of $G(\Delta t A)$ have the property $|\mu_k| < 1$.

4. The eigenvalues μ_k of $G(\Delta t A)$ can be obtained by substitution of the eigenvalues λ_k of the matrix A into the amplification factor:

   $$\mu_k = C(\Delta t\lambda_k). \tag{6.5.4}$$

   Hence, for stability we need $|C(\Delta t\lambda_k)| < 1$.

Exercise 6.5.1 *The amplification matrices for forward Euler, improved Euler, backward Euler, Crank-Nicolson and the θ-method are given by*

$$I + \Delta t A,$$
$$I + \Delta t A + \frac{1}{2}(\Delta t A)^2,$$
$$(I - \Delta t A)^{-1},$$
$$(I - \frac{1}{2}\Delta t A)^{-1}(I + \frac{1}{2}\Delta t A),$$
$$(I - \theta \Delta t A)^{-1}(I + (1-\theta)\Delta t A).$$

Show this. What are the corresponding amplification factors? □

We recall that our matrix A is of the form $A = M^{-1}S$, with M and S satisfying the conditions of Section 6.3. Hence, in order to investigate the stability of the numerical time integration, the eigenvalues of $M^{-1}S$ have to be estimated.

Lemma 6.5.1 *Let M and S be real symmetric matrices of the same size. If M is positive definite and S negative definite, then the matrix $A = M^{-1}S$ is diagonalizable with all eigenvalues real and negative.*

Proof: It follows from Theorem 1.4.1 and Corollary 1.4.3 that M has only positive, real eigenvalues, and can be written as M=$Q\Lambda Q^T$, where Λ is a diagonal matrix with the eigenvalues of M on the diagonal, and Q is a real orthogonal matrix. This implies that $M^{-1/2} = Q\Lambda^{-1/2}Q^T$ exists and is symmetric positive definite too. Hence $A = M^{-1}S$ is similar to $B = M^{1/2}AM^{-1/2} = M^{-1/2}SM^{-1/2}$. The matrix B is real symmetric and is therefore diagonalizable and has only real eigenvalues (see Theorem 1.4.1). Since A is similar to B, A is diagonalizable as well, and has the same eigenvalues as B (see Exercise 1.4.1). Furthermore, S is symmetric negative definite, i.e. $(S\boldsymbol{x}, \boldsymbol{x}) < 0$ for all $\boldsymbol{x} \neq \mathbf{0}$. Hence $(B\boldsymbol{x}, \boldsymbol{x}) = (M^{-1/2}SM^{-1/2}\boldsymbol{x}, \boldsymbol{x}) = (SM^{-1/2}\boldsymbol{x}, M^{-1/2}\boldsymbol{x}) < 0$ for all $\boldsymbol{x} \neq \mathbf{0}$. This implies that the eigenvalues of B (and therefore also of A) are negative. □

We note that the eigenvalues of $M^{-1}S$ are the same as the eigenvalues of the following so-called generalized eigenvalue problem:

$$\text{Determine } \lambda \text{ and } \boldsymbol{x} \neq \mathbf{0} \text{ such that } S\boldsymbol{x} = \lambda M\boldsymbol{x}. \tag{6.5.5}$$

All eigenvalues of the above generalized eigenvalue problem are therefore real-valued and negative (see also [13]). Hence all eigenvalues are contained in the real interval $[\lambda_{min}, 0)$, where $\lambda_{min} < 0$ denotes the minimal eigenvalue of A. In this case, the following criterion for stability holds:

$$\Delta t < \frac{c}{|\lambda_{min}|}, \tag{6.5.6}$$

with $c = 2$ for Euler and improved Euler and $c = 2.8$ for Runge-Kutta (see [4]). Hence we have to estimate the minimal eigenvalue of the generalized eigenvalue problem. This is treated in the next section.

6.5.1 Gershgorin's disk theorem

We recall that the mass matrix M is diagonal in case of Finite Differences (FDM) or Finite Volumes (FVM). We extend Gershgorin's theorem (see Theorem 1.4.5) for this case to estimate the location of the eigenvalues.

Theorem 6.5.2 (Gershgorin)
If M is diagonal, then, for each eigenvalue λ of $M^{-1}S$, there exists an index i with

$$|m_{ii}\lambda - s_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{N} |s_{ij}|. \tag{6.5.7}$$

Exercise 6.5.2 *Prove Theorem 6.5.2 using Theorem 1.4.5.* □

Example 6.5.1 *For the heat equation in one spatial dimension (see Example 6.2.1) the Finite Difference Method gives $M = I$ and hence*

$$|\lambda_{min}| \leq \frac{4}{h^2}. \tag{6.5.8}$$

From this we obtain a stability criterion for the Forward Euler method:

$$\Delta t < \frac{2h^2}{4} = \frac{1}{2}h^2. \tag{6.5.9}$$

For the 2D heat equation, the Finite Difference Method (see Example 6.4.1) gives in a similar way:

$$|\lambda_{min}| \leq \frac{4}{(\Delta x)^2} + \frac{4}{(\Delta y)^2}, \tag{6.5.10}$$

and the following stability criterion for Euler's method:

$$\Delta t < \frac{\beta^2}{2(1+\beta^2)}(\Delta x)^2, \tag{6.5.11}$$

where $\beta = \frac{\Delta y}{\Delta x}$. □

Example 6.5.2 *We consider the matrices M and S of Example 6.2.2.*

Exercise 6.5.3 *Let s_i denote the sum of the absolute values of the elements in the i-th row of $M^{-1}S$. Show that the following relations for s_i hold:*

$$\begin{aligned}
s_1 &= \frac{2}{h_1+h_2}\left(\frac{1}{h_1}+\frac{2}{h_2}\right) < \frac{2}{h_1+h_2}\left(\frac{2}{h_1}+\frac{2}{h_2}\right) = \frac{4}{h_1 h_2}, \\
s_i &= \frac{2}{h_i+h_{i+1}}\left(\frac{2}{h_i}+\frac{2}{h_{i+1}}\right) = \frac{4}{h_i h_{i+1}}, \quad i = 2,\ldots,N-1, \\
s_N &= \frac{2}{h_N}\frac{2}{h_N} = \frac{4}{h_N^2}.
\end{aligned}$$

□

Gershgorin's theorem results in the following estimate:

$$|\lambda_{min}| \leq \max\left(\frac{4}{h_N^2}, \max_{1\leq i\leq N-1}\frac{4}{h_i h_{i+1}}\right),$$

and a stability criterion for the Forward Euler method of the form

$$\Delta t < \frac{1}{2}\min\left(h_N^2, \min_{1\leq i\leq N-1} h_i h_{i+1}\right).$$

□

In all the examples the time step has to be smaller than the product of a factor times the square of the grid spacing. In practical situations, this could imply that the time step has to be very small. For that reason explicit time integration methods are not popular for the heat equation. Implicit methods such as the Crank-Nicolson method or the implicit (backward) Euler method are usually preferred. *This always implies the solution of a problem with the complexity of the Laplacian in each time step*. In one space dimension, this amounts to the solution of a tridiagonal system of equations in each time step, which is no big deal. Two and more space dimensions, however, lead to the same type of problems as the Laplacian. For iterative methods the solution on the previous time level is of course an excellent starting value.

For regions with simple geometries some special implicit methods for the heat equation are available. This will be addressed in Section 6.8.

Exercise 6.5.4 *Prove that Euler backward and Crank-Nicolson are absolutely stable for each value of the stepsize Δt if $Re(\lambda_k) < 0$.* □

Exercise 6.5.5 *Prove that the θ-method is absolutely stable for all $\Delta t > 0$ if $\theta \geq \frac{1}{2}$. Derive a condition for stability for the case that $\theta < \frac{1}{2}$.* □

As an illustration of the stability of the numerical solution to the heat problem we consider a Finite Difference solution in the square $\Omega = [0,1] \times [0,1]$, on which

$$\frac{\partial u}{\partial t} = 0.5\Delta u. \tag{6.5.12}$$

We take as initial condition and boundary condition on the whole boundary Γ:

$$\begin{aligned} u(x,y,0) &= \sin(x)\sin(y), \quad (x,y) \in \Omega, \\ u(x,y,t) &= \sin(x)\sin(y), \quad (x,y) \in \Gamma. \end{aligned} \tag{6.5.13}$$

Exercise 6.5.6 *Prove that the analytical solution to the above problem is given by*

$$u(x,y,t) = e^{-t}\sin(x)\sin(y). \tag{6.5.14}$$

□

In Figure 6.1 we show the numerical solution to the above problem as computed by the use of the Forward Euler method with $\Delta t = 0.1$. For this case the stability criterion is violated and hence the solution exhibits unphysical behavior. In Figure 6.2 we show the solution that has been obtained for the same data by the backward Euler method. Now the solution exhibits the expected physical behavior. The contour lines are nice and smooth and are similar to the ones of the analytical solution.

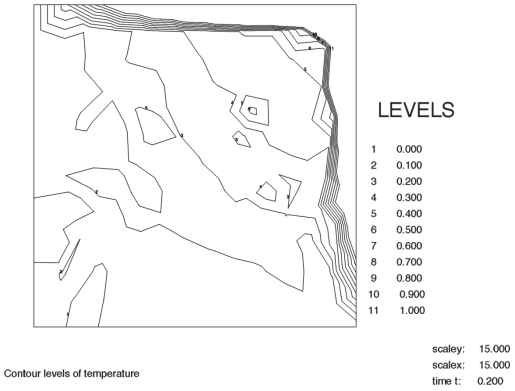

Figure 6.1: Contour lines of the numerical solution to the heat equation with $\Delta t = 0.1$ as obtained by the use of the *Forward* (explicit) Euler method (unstable solution).

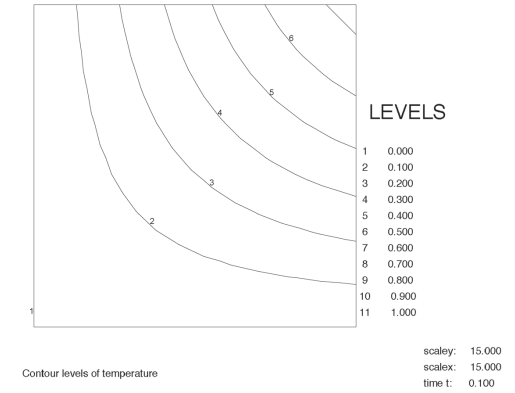

Figure 6.2: Contour lines of the numerical solution to the heat equation with $\Delta t = 0.1$ as obtained by the use of the *Backward* (implicit) Euler method.

6.5.2 Stability analysis of Von Neumann

As an alternative method for estimating the eigenvalues of the matrix $M^{-1}S$ we present a method due to the American mathematician John von Neumann. We recall that in the method-of-lines setting, stability of the (fully discrete) numerical scheme is determined by the eigenvalues $\mu_k = C(\Delta t \lambda_k)$ of the amplification matrix $G(\Delta t A)$, where $A = M^{-1}S$ is the spatial discretization matrix with eigenvalues λ_k and C is the amplification factor of the time stepping method.
The Von Neumann method is applicable to linear equations with *constant coefficients* such as the heat equation when using *equidistant grids*. By neglecting the boundary conditions the eigenvectors can be assumed of the form

$$v_k = e^{i\rho kh} \tag{6.5.15}$$

in one and

$$v_{kl} = e^{i(\rho k\Delta x + \sigma l \Delta y)} \tag{6.5.16}$$

in two space dimensions. The region must be rectangular in 2D and the numbers ρ and σ are considered *arbitrary* real numbers. In order to find an interval $[a, b]$ containing the eigenvalues of A it is sufficient to substitute these expressions in one single equation of the generalized eigenvalue problem.
The Von Neumann method provides a necessary condition for stability by requiring that

$$|C(\Delta t \lambda)| \leq 1 \text{ for all } \lambda \in [a, b] \tag{6.5.17}$$

Example 6.5.3 *As an example we consider the heat equation with an equidistant grid in one space dimension,*

$$\lambda e^{i\rho kh} = \frac{1}{h^2}(e^{i\rho(k-1)h} - 2e^{i\rho kh} + e^{i\rho(k+1)h}). \tag{6.5.18}$$

We divide the left and right-hand sides of this equation by $e^{i\rho kh}$ and obtain, using the relation $1/2(e^{i\phi} + e^{-i\phi}) = \cos\phi$:

$$\lambda = \frac{2(\cos(\rho h) - 1)}{h^2} = -4\frac{\sin^2 \rho h/2}{h^2}. \tag{6.5.19}$$

From this we find that the interval $[-\frac{4}{h^2}, 0]$ contains all eigenvalues of A.
The corresponding Von Neumann stability criterion is

$$\Delta t \leq \frac{1}{2}h^2 \tag{6.5.20}$$

for the forward Euler time-integration. □

Remark

In case of two space dimensions, the domain of computation, in which the Von Neumann analysis is applied, does not necessarily have to be rectangular. In

that case the analysis gives a rough upper bound for the eigenvalues, which in fact holds for the smallest rectangle that encloses the domain of computation. The coefficients in the PDE have to be constant. Furthermore the discretization has to be equidistant, otherwise the analysis is not valid. In contrast, Gershgorin's theorem can also be applied for non-constant coefficients and non-equidistant grids, but the mass matrix has to be diagonal in that case.

6.6 The accuracy of the time integration

When we use a numerical method for time integration we make an error at each time step. These errors accumulate in general, and you might ask if this accumulation could be disastrous. From [4] we know that in a bounded time interval $(t_0, T]$ a local truncation error of the order $\mathcal{O}(h^m)$ gives a global error of the same order. The forward and backward methods of Euler have $m = 1$, whereas the improved Euler method and the method of Crank-Nicolson have $m = 2$. Absolutely stable systems like the heat equation have even better properties. If the numerical integration is stable, the global error is uniformly bounded on the interval (t_0, ∞).

Theorem 6.6.1 *Let $\boldsymbol{y}(t)$ be the solution of the absolutely stable system*

$$\frac{d\boldsymbol{y}}{dt} = A\boldsymbol{y} + \boldsymbol{f}, \quad \boldsymbol{y}(t_0) = \boldsymbol{y}_0, \tag{6.6.1}$$

where $A = M^{-1}S$ with M symmetric positive definite and S symmetric negative definite. Further, let $\boldsymbol{u}^n$ be the solution of the numerical method

$$\boldsymbol{u}^{n+1} = G(\Delta t A)\boldsymbol{u}^n + I_n(\boldsymbol{f}), \quad \boldsymbol{u}^0 = \boldsymbol{y}_0, \tag{6.6.2}$$

where $I_n(\boldsymbol{f})$ represents an approximation of

$$\int_{t_n}^{t_{n+1}} e^{(t_{n+1}-t)A}\boldsymbol{f}(t)\,dt.$$

Assume that there exists a stepsize $\tau > 0$ such that

$$\lim_{n\to\infty} G(\Delta t A)^n = 0, \quad \textit{for all } \Delta t \le \tau,$$

and

$$\boldsymbol{y}(t_{n+1}) = G(\Delta t A)\boldsymbol{y}(t_n) + I_n(\boldsymbol{f}) + (\Delta t)^{m+1}\boldsymbol{p}^n, \tag{6.6.3}$$

where $\|\boldsymbol{p}^n\|$ is uniformly bounded for all n and $\Delta t \le \tau$.
Then it follows that

$$\|\boldsymbol{y}(t_n) - \boldsymbol{u}^n\| = \mathcal{O}((\Delta t)^m). \tag{6.6.4}$$

In other words: if the local truncation error in time is of order m (after division of Equation (6.6.3) by Δt), the global error is also of order m provided the integration is stable.

Proof
We define $\boldsymbol{\epsilon}^n = \boldsymbol{y}(t_n) - \boldsymbol{u}^n$ and subtract Equation (6.6.2) from Equation (6.6.3) to get:

$$\boldsymbol{\epsilon}^{n+1} = G(\Delta t A)\boldsymbol{\epsilon}^n + (\Delta t)^{m+1}\boldsymbol{p}^n. \tag{6.6.5}$$

Taking into account that $\boldsymbol{\epsilon}^0 = \mathbf{0}$, this recurrence relation can be solved to give

$$\boldsymbol{\epsilon}^n = (\Delta t)^{m+1} \sum_{k=0}^{n-1} G(\Delta t A)^{n-k-1}\boldsymbol{p}^k. \tag{6.6.6}$$

Since $\|\boldsymbol{p}^k\|$ is uniformly bounded, there exists a vector $\boldsymbol{p}_{max}$ with $\|\boldsymbol{p}^k\| \leq \|\boldsymbol{p}_{max}\|$ for all k and Δt. Putting this into (6.6.6) and using Exercise 1.4.4 we obtain

$$\|\boldsymbol{\epsilon}^n\| \leq (\Delta t)^{m+1} \sum_{k=0}^{n-1} \|G(\Delta t A)^{n-k-1}\| \|\boldsymbol{p}_{max}\|.$$

It follows from Lemma 6.5.1 that A is diagonalizable and has real negative eigenvalues λ_k only, so can be written as

$$A = V\Lambda V^{-1},$$

where Λ is a diagonal matrix containing the eigenvalues λ_k and V is the matrix whose columns are the corresponding eigenvectors. Hence we have

$$G(\Delta t A) = VDV^{-1},$$

where D is the diagonal matrix with the eigenvalues μ_k of $G(\Delta t A)$, which are given by $\mu_k = C(\Delta t \lambda_k)$ (see (6.5.4)). This yields $G(\Delta t A)^k = VD^kV^{-1}$, and using the submultiplicativity of the matrix norm (cf. Exercise 1.4.5) we conclude that

$$\|G(\Delta t A)^k\| \leq |\mu_1|^k \|V\| \|V^{-1}\|,$$

where μ_1 is the eigenvalue of $G(\Delta t A)$ with the largest modulus. This gives

$$\|\boldsymbol{\epsilon}^n\| \leq (\Delta t)^{m+1} \frac{1 - |\mu_1|^n}{1 - |\mu_1|} \|V^{-1}\| \, \|V\| \, \|\boldsymbol{p}_{max}\|.$$

Since $\mu_1 = C(\lambda_1 \Delta t) = 1 + \lambda_1 \Delta t + \mathcal{O}((\Delta t)^2)$, we have $1 - |\mu_1| = |\lambda_1| \Delta t + \mathcal{O}((\Delta t)^2)$ and we finally obtain

$$\|\boldsymbol{\epsilon}^n\| \leq K(\Delta t)^m,$$

which proves the theorem. □

6.7 Conclusions for the method of lines

We summarize the results of the methods of lines for the heat/diffusion equation.

- Using the method of lines, the PDE is written as a system of ordinary differential equations by the spatial discretization of the elliptic operator.
- The global error of the *analytic* solution of this system of ordinary differential equations (compared to the solution of the solution of the PDE) is of the same order as the order of consistency of the FDM and FVM.
- The *numerical* solution of this system has an additional error due to the numerical time integration. This global error is of the order of $K\Delta t^m$ if the local truncation error is of the order $\mathcal{O}(\Delta t^m)$. This constant does not depend on time t and this estimate holds on the *entire* time interval (t_0, ∞).
- Explicit (and some implicit) methods have a stability criterion of the form
$$\Delta t < c(\Delta x)^2 \tag{6.7.1}$$
and hence these methods are less suitable for the heat equation.

6.8 Special difference methods for the heat equation

The method of lines is a general method, which is applicable to one, two or three spatial dimensions. At each time step, the implicit methods give a problem to be solved with the same complexity as the Poisson problem. Therefore, one has searched for methods that are *stable* but have a simpler complexity than the Poisson problem. We present one example of such a method: The ADI method. This method can only be used with regular grids with a five-point molecule for the elliptic operator. First we sketch the principle of the ADI method and subsequently a formal description of the ADI method is given.

6.8.1 The principle of the ADI method

The abbreviation ADI means *Alternating Direction Implicit*. This is a fairly accurate description of the working of the method. Suppose that we have to solve the heat equation on a rectangle with length l_x and width l_y and we use a discretization with stepsize Δx and Δy respectively, such that $N_x \Delta x = l_x$ and $N_y \Delta y = l_y$. For convenience we apply Dirichlet boundary conditions at all the boundaries of the domain of computation, where we set $u = 0$. For the time integration from t_n to t_{n+1} the ADI method uses two steps. The idea is

as follows: first we use a half time step with an intermediate auxiliary quantity u^*. To compute u^* we use the implicit Euler time integration method for the derivative with respect to x and the explicit Euler time integration for the derivative with respect to y. In the next half time step, we reverse this process. Hence: The first step, a so-called half time step, computes an auxiliary-quantity u^*_{ij} according to:

$$\begin{aligned} u^*_{ij} = u^n_{ij} + \frac{\Delta t}{2(\Delta x)^2}(u^*_{i+1,j} - 2u^*_{ij} + u^*_{i-1,j}) + \\ \frac{\Delta t}{2(\Delta y)^2}(u^n_{i,j+1} - 2u^n_{ij} + u^n_{i,j-1}) + \frac{\Delta t}{2} f^*_{ij}, \\ i = 1, \ldots, N_x - 1, j = 1, \ldots, N_y - 1, \end{aligned} \tag{6.8.1}$$

where f^*_{ij} denotes $f(i\Delta x, j\Delta y, t_n + \frac{1}{2}\Delta t)$. Subsequently u^{n+1} is calculated according to:

$$\begin{aligned} u^{n+1}_{ij} = u^*_{ij} + \frac{\Delta t}{2(\Delta x)^2}(u^*_{i+1,j} - 2u^*_{ij} + u^*_{i-1,j}) + \\ \frac{\Delta t}{2(\Delta y)^2}(u^{n+1}_{i,j+1} - 2u^{n+1}_{ij} + u^{n+1}_{i,j-1}) + \frac{\Delta t}{2} f^*_{ij}, \\ i = 1, \ldots, N_x - 1, j = 1, \ldots, N_y - 1. \end{aligned} \tag{6.8.2}$$

Equation (6.8.1) requires that, for each fixed index j, a tridiagonal system of equations has to be solved for $\boldsymbol{u}^*_j$, with

$$\boldsymbol{u}^*_j = \begin{pmatrix} u^*_{1j} \\ u^*_{2j} \\ \vdots \\ u^*_{N_x-1,j} \end{pmatrix}. \tag{6.8.3}$$

In total there are $N_y - 1$ systems like this one to be solved in order to determine all the values of $\boldsymbol{u}^*_j$. Similarly, one has to solve in Equation (6.8.2) for a fixed index i a tridiagonal system of equations in $\boldsymbol{u}^{n+1}_i$, with

$$\boldsymbol{u}^{n+1}_i = \begin{pmatrix} u^{n+1}_{i1} \\ u^{n+1}_{i2} \\ \vdots \\ u^{n+1}_{i,N_y-1} \end{pmatrix}. \tag{6.8.4}$$

This is exactly in the other direction, which explains the name of the method. In total we are faced with $N_x - 1$ of such systems. Hence to integrate the heat equation from t_n up to t_{n+1} one has to

- solve $N_y - 1$ tridiagonal systems of size $N_x - 1$,
- solve $N_x - 1$ tridiagonal systems of size $N_y - 1$.

Exercise 6.8.1 *Verify that the amount of computational effort per time step for the ADI method is proportional to the total number of gridpoints. (Hint: How many operations does it take to solve an $N \times N$ tridiagonal system of equations?)* □

Indeed the computational complexity of the ADI method is better than that of the method of lines. However, the question remains whether this benefit is not at the expense of the accuracy or the stability of the method. To scrutinize this, a formal description of the ADI method is presented in the next section.

6.8.2 Formal description of the ADI method

The ADI method can be seen as a special way to integrate the system of ordinary differential equations

$$\frac{d\boldsymbol{u}}{dt} = (A_x + A_y)\boldsymbol{u} + \boldsymbol{f}, \tag{6.8.5}$$

which arises from a PDE using the method of lines. The ADI method of this system is given by:

$$\boldsymbol{u}^* = \boldsymbol{u}^n + \frac{1}{2}\Delta t(A_x\boldsymbol{u}^* + A_y\boldsymbol{u}^n + \boldsymbol{f}^*), \tag{6.8.6}$$

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^* + \frac{1}{2}\Delta t(A_x\boldsymbol{u}^* + A_y\boldsymbol{u}^{n+1} + \boldsymbol{f}^*). \tag{6.8.7}$$

From this the intermediate quantity $\boldsymbol{u}^*$ can be eliminated as follows. First rewrite the above two equations as

$$(I - \frac{1}{2}\Delta tA_x)\boldsymbol{u}^* = (I + \frac{1}{2}\Delta tA_y)\boldsymbol{u}^n + \frac{1}{2}\Delta t\boldsymbol{f}^*, \tag{6.8.8}$$

$$(I - \frac{1}{2}\Delta tA_y)\boldsymbol{u}^{n+1} = (I + \frac{1}{2}\Delta tA_x)\boldsymbol{u}^* + \frac{1}{2}\Delta t\boldsymbol{f}^*. \tag{6.8.9}$$

After multiplication of the first relation by $I + \frac{1}{2}\Delta tA_x$ and the second one by $I - \frac{1}{2}\Delta tA_x$, and noting that these matrices commute, one easily obtains:

$$(I - \frac{1}{2}\Delta tA_x)(I - \frac{1}{2}\Delta tA_y)\boldsymbol{u}^{n+1} = \\ (I + \frac{1}{2}\Delta tA_x)(I + \frac{1}{2}\Delta tA_y)\boldsymbol{u}^n + \Delta t\boldsymbol{f}^*. \tag{6.8.10}$$

Equation (6.8.10) is the basis of our investigations. First, we make a statement about the accuracy.

Theorem 6.8.1 *Equation (6.8.10) differs from Crank-Nicolson's method applied to (6.8.5) by a term of the order of* $\mathcal{O}(\Delta t^3)$.

Proof
Crank-Nicolson applied to (6.8.5) gives

$$(I - \frac{1}{2}\Delta t A_x - \frac{1}{2}\Delta t A_y)\boldsymbol{u}^{n+1} = (I + \frac{1}{2}\Delta t A_x + \frac{1}{2}\Delta t A_y)\boldsymbol{u}^n + \frac{1}{2}\Delta t(\boldsymbol{f}^n + \boldsymbol{f}^{n+1}).$$

Elaboration of (6.8.10) gives:

$$(I - \frac{1}{2}\Delta t A_x - \frac{1}{2}\Delta t A_y)\boldsymbol{u}^{n+1} = (I + \frac{1}{2}\Delta t A_x + \frac{1}{2}\Delta t A_y)\boldsymbol{u}^n + \frac{1}{4}(\Delta t)^2 A_x A_y(\boldsymbol{u}^n - \boldsymbol{u}^{n+1}) + \Delta t \boldsymbol{f}^*.$$

Now the theorem immediately follows by noting that $\boldsymbol{u}^n - \boldsymbol{u}^{n+1}$ is of order $\mathcal{O}(\Delta t)$ and that $\boldsymbol{f}^* = \frac{1}{2}(\boldsymbol{f}^n + \boldsymbol{f}^{n+1}) + \mathcal{O}((\Delta t)^2)$. □

It follows from the above theorem that the ADI method has the same accuracy as the method of Crank Nicolson, which is $\mathcal{O}((\Delta t)^2)$.

It is hard to investigate the stability of the ADI method theoretically. In practical situations, it turns out that the ADI method does not require a stringent stability criterion. In a special case, there is a theoretical justification for the unconditional stability of the ADI method:

Theorem 6.8.2 *If* A_x *and* A_y *are commuting matrices (i.e.* $A_x A_y = A_y A_x$*), then the ADI method is unconditionally stable.*

Proof
We have to calculate the eigenvalues of

$$(I - \frac{1}{2}\Delta t A_y)^{-1}(I - \frac{1}{2}\Delta t A_x)^{-1}(I + \frac{1}{2}\Delta t A_x)(I + \frac{1}{2}\Delta t A_y),$$

but under the conditions of the theorem all these matrices commute. Then, the eigenvalues of this matrix are given by products of the eigenvalues of the separate matrices

$$(I - \frac{1}{2}\Delta t A_x)^{-1}(I + \frac{1}{2}\Delta t A_x) \text{ and } (I - \frac{1}{2}\Delta t A_y)^{-1}(I + \frac{1}{2}\Delta t A_y).$$

These eigenvalues are

$$\frac{1 + \frac{1}{2}\Delta t \lambda_x}{1 - \frac{1}{2}\Delta t \lambda_x} \text{ and } \frac{1 + \frac{1}{2}\Delta t \lambda_y}{1 - \frac{1}{2}\Delta t \lambda_y}.$$

Since λ_x and λ_y are real-valued and negative, the modulus of all these eigenvalues is less than one. □

Exercise 6.8.2 *Show that the operators A_x and A_y commute for the problem of the rectangle with Dirichlet conditions.* □

Extension of the ADI method to three spatial dimensions is not straightforward. The most straightforward way (three steps, subsequently for the x-, y- and z-coordinate) is no longer unconditionally stable. Further, its global error is of the order $\mathcal{O}(\Delta t)$. There exist adequate ADI methods for three spatial coordinates, see [10].

6.9 Summary of Chapter 6

In this chapter we looked at the numerical solution of the *heat* or *diffusion* equation. We have shown that with one exception this equation has an equilibrium solution and that independent of the initial values the transient solution tends to this equilibrium solution exponentially fast.

We introduced the *method of lines* for the numerical solution which transforms the PDE into a set of ODEs by discretizing first the spatial differential operators. We estimated the effect of the truncation error of the spatial discretization on the solution of this system of ODEs. We proved that this effect is uniformly bounded.

We briefly looked at the stability of the explicit integration schemes for which we had to estimate the location of the eigenvalues of the system matrix. To this end we could use *Gershgorin's disk theorem* or *Von Neumann's stability analysis.*

Finally we considered the *ADI-method*, an unconditionally stable method of much lower complexity than Crank-Nicolson's method, but with the same accuracy.

Chapter 7

The wave equation

Objectives

In this chapter we shall look at various methods for the time integration of the wave equation. This equation is crucial in applications dealing with electromagnetic radiation, wave propagation, acoustics and seismics (used for oil finding for instance). Before we do this, a conservation principle for the solution of the wave equation is derived. The numerical solution should satisfy this principle as well. Stability in terms of decay and growth of the numerical solution as a function of time is investigated for several methods. Furthermore, the concepts *dispersion* and *dissipation* will be introduced and an illustration of these concepts will be given. Finally a procedure to derive the CFL-criterion, a criterion for the numerical solution to represent the exact solution, will be given by using the concept of (analytical and numerical) domain of dependence.

7.1 A fundamental equality

Consider the wave equation on a domain Ω:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) =: c^2 \Delta u. \tag{7.1.1}$$

In Equation (7.1.1) no *internal* energy source term is taken into account. Further, homogeneous boundary conditions are imposed on the boundaries Γ_1,

Γ_2 and Γ_3 of the domain Ω, i.e.

$$\begin{aligned} u &= 0, && (x,y) \in \Gamma_1, \\ \frac{\partial u}{\partial n} &= 0, && (x,y) \in \Gamma_2, \\ \sigma u + \frac{\partial u}{\partial n} &= 0, && (x,y) \in \Gamma_3. \end{aligned} \tag{7.1.2}$$

Hence there is no transport of energy through the boundaries. Therefore the PDE (7.1.1) with boundary conditions (7.1.2) is homogeneous. As initial conditions, we have that u and $\frac{\partial u}{\partial t}$ are given at $t = t_0$ at all points in the domain of computation. Now we will show that the 'energy' of this equation is preserved in time.

Theorem 7.1.1 *The homogeneous wave equation* (7.1.1) *with homogeneous boundary conditions* (7.1.2) *satisfies the following conservation principle:*

$$\frac{1}{2}\int_\Omega \left\{ \left(\frac{\partial u}{\partial t}\right)^2 + c^2 \|\text{grad}\, u\|^2 \right\} d\Omega + \frac{1}{2}\int_{\Gamma_3} \sigma c^2 u^2 d\Gamma = \textit{Constant}. \tag{7.1.3}$$

Proof: We multiply both sides of the equality of Equation (7.1.1) by $\frac{\partial u}{\partial t}$ and integrate the results over the domain Ω to obtain

$$\int_\Omega \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)^2 d\Omega = \int_\Omega c^2 \frac{\partial u}{\partial t}\Delta u \, d\Omega = \int_\Omega c^2 \frac{\partial u}{\partial t}\text{div grad}\, u \, d\Omega.$$

Assuming that all derivatives are continuous and using the product rule for differentiation (Theorem 1.3.2), the integrand of the right-hand side can be written as

$$\text{div}\left(\frac{\partial u}{\partial t}\,\text{grad}\, u\right) - \text{grad}\left(\frac{\partial u}{\partial t}\right)\cdot \text{grad}\, u.$$

This yields

$$\int_\Omega \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)^2 d\Omega = \int_\Omega c^2 \text{div}\left(\frac{\partial u}{\partial t}\text{grad}\, u\right) d\Omega - \int_\Omega c^2 \text{grad}\left(\frac{\partial u}{\partial t}\right)\cdot \text{grad}\, u \, d\Omega.$$

We apply the divergence theorem to the first term on the right-hand side and use the product rule for differentiation on the second term of the right-hand

side to get

$$\int_\Omega \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)^2 d\Omega = \int_{\Gamma_1\cup\Gamma_2\cup\Gamma_3} c^2 \frac{\partial u}{\partial t}\frac{\partial u}{\partial n}\, d\Gamma - \frac{1}{2}\int_\Omega c^2 \frac{\partial}{\partial t}\left(\operatorname{grad} u \cdot \operatorname{grad} u\right)\, d\Omega. \quad (7.1.4)$$

The integrand of the boundary integral on the right-hand side vanishes on Γ_1 and Γ_2 due to the boundary conditions. Application of the boundary condition on Γ_3 then transforms Equation (7.1.4) into

$$\int_\Omega \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)^2 d\Omega = -\int_{\Gamma_3} c^2\sigma u \frac{\partial u}{\partial t}\, d\Gamma - \frac{1}{2}\int_\Omega c^2 \frac{\partial}{\partial t}\left(\operatorname{grad} u \cdot \operatorname{grad} u\right)\, d\Omega.$$

Finally using a standard differentiation property we get

$$\int_\Omega \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)^2 d\Omega = -\int_{\Gamma_3} \frac{1}{2} c^2\sigma \frac{\partial u^2}{\partial t} d\Gamma - \frac{1}{2}\int_\Omega c^2 \frac{\partial}{\partial t}\left(\operatorname{grad} u \cdot \operatorname{grad} u\right)\, d\Omega.$$

Interchanging the differentiation and integration operations in the above expression and subsequent integration over time t proves the theorem. □

Remarks

1. Consider the wave equation with a source term,

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u + f(\boldsymbol{x}, t). \quad (7.1.5)$$

   The difference between two solutions of Equation (7.1.5) with the same source term f and the same boundary conditions satisfies the homogeneous wave equation (7.1.1) and homogeneous boundary conditions (7.1.2).

2. The first term in Equation (7.1.3) gives the kinetic energy of the vibrating medium, whereas the second and a third term involve the potential energy. Therefore, the left-hand side of Equation (7.1.3) is commonly referred to as (the square of) the energy norm.

3. The total amount of energy is entirely defined by the two initial conditions $u(x, y, t_0)$ and $\dfrac{\partial u}{\partial t}(x, y, t_0)$.

4. The difference in this 'energy-norm', between two solutions of (7.1.5) with the same boundary conditions and different initial conditions is constant at all stages.

Exercise 7.1.1 *Prove remarks 1 and 4.* □

Exercise 7.1.2 *The solution of the heat equation in the previous chapter tends to an equilibrium solution (i.e. a steady-state) as t tends to infinity. Does the solution of the wave equation tend to a steady state as t tends to infinity?* □

From remark 4 it follows that the solution of the wave equation is neutrally stable, that is, an error made in the initial conditions will neither decrease nor increase and hence it persists. This property must also hold for our numerical methods. Otherwise the numerical solution would not exhibit the same physical characteristics as the analytical solution.

7.2 The method of lines

In a similar way as we did for parabolic equations we may first discretize only the spatial part of the wave equation. The difference with the previous chapter is that we now have to deal with a second order system with respect to time. After the discretization of Equation (7.1.5), we obtain:

$$M\frac{d^2\boldsymbol{u}}{dt^2} = c^2 S\boldsymbol{u} + \boldsymbol{f}, \quad \boldsymbol{u}(t_0) = \boldsymbol{u}_0, \quad \frac{d\boldsymbol{u}}{dt}(t_0) = \boldsymbol{v}_0. \tag{7.2.1}$$

Here M and S are the *mass matrix* and *stiffness matrix*, respectively, just like in the previous chapter. Next, we establish that Equation (7.2.1) also conserves the energy if $\boldsymbol{f} = \boldsymbol{0}$.

Theorem 7.2.1 *If* $\boldsymbol{f} = \boldsymbol{0}$, *then*

$$\frac{1}{2}\left(M\frac{d\boldsymbol{u}}{dt}, \frac{d\boldsymbol{u}}{dt}\right) - \frac{1}{2}c^2(S\boldsymbol{u}, \boldsymbol{u}) = \textit{constant}. \tag{7.2.2}$$

Exercise 7.2.1 *Prove this theorem. Hint: take the inner product of* (7.2.1) *with* $d\boldsymbol{u}/dt$ *and use the symmetry of M and S.* □

7.2.1 The error in the solution of the system

Application of the method of lines generates a truncation error $\boldsymbol{E}$ in the spatial discretization. This may be defined by

$$M\frac{d^2\boldsymbol{y}}{dt^2} = c^2 S\boldsymbol{y} + \boldsymbol{f} + M\boldsymbol{E}, \tag{7.2.3}$$

where $\boldsymbol{y}$ denotes the exact solution to the wave equation. This truncation error causes an error in the solution of (7.2.1) of the form Ch^p, where h denotes a generic discretization parameter (such as the diameter of the largest element used in the discretization) and p represents the order of consistency. For the

heat equation it was possible to find a constant C, valid for the entire interval of integration (t_0, ∞). For the wave equation this is not possible. The constant C depends linearly on the length of the integration interval (t_0, T). A complete analysis of the error is beyond the scope of the book, but qualitatively the phenomenon is explained as follows: An *eigenvibration* of (7.1.1) is given by a function of the form of $e^{i\lambda ct}U(x, y)$, where U satisfies the *homogeneous* boundary conditions (note that the boundary conditions can be of several types). Substitution into Equation (7.1.1) yields

$$-\lambda^2 c^2 U = c^2 \Delta U. \tag{7.2.4}$$

This is just the eigenvalue problem for the Laplace operator, which has an infinite number of solutions in terms of eigenpairs λ_k and U_k. Here λ_k is the *eigenfrequency* of the vibration and U_k the *eigenfunction*. These quantities depend on the domain of computation Ω. Generally speaking the wavelength of the eigenfunction (which is inversely related to the number of peaks) decreases as the eigenfrequency increases.

Consider the discrete version of Equation (7.1.1), which is given by system (7.2.1). We obtain:

$$-\lambda_h^2 c^2 M\boldsymbol{U} = c^2 S\boldsymbol{U}. \tag{7.2.5}$$

The subscript h indicates that eigenvalues of the discretized problem are considered. The discretized system only has a finite number of eigenvalues, or to put it differently: the resolution is finite on the discrete grid. The shortest wave that can be represented on a grid has wavelength $\mathcal{O}(2h)$. For eigenfunctions that can be represented well on the grid we have

$$|\lambda - \lambda_h| = \mathcal{O}(h^p) \text{ and } \|\boldsymbol{U} - \boldsymbol{U}_h\| = \mathcal{O}(h^p). \tag{7.2.6}$$

Since the eigenfrequencies of numerical and exact solution differ, the difference between the numerical solution and the exact solution increases as the simulation proceeds. This results in a *phase-shift error*. Moreover, this phase-shift error differs for the different eigenvibrations. This phenomenon is called *dispersion*. Since each solution can be written as a linear combination of eigenvibrations, there will be dispersion in the solution of Equation (7.2.1) in relation to the solution of Equation (7.1.1). This dispersion even exists for the eigenfunctions, which are represented well on the grid (i.e. eigenfunctions with a large wavelength, i.e. a small frequency). Therefore, the difference between the solution of (7.2.1) and the exact solution of the wave equation (7.1.1) increases as the interval of the time integration increases. Since the error is of the form $C(T - t_0)h^p$, one has to use a more accurate spatial discretization as T increases if the same absolute accuracy is to be maintained for the final stages of the time interval as for the initial stages of the computation process.

As an example, we consider

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \text{ for } 0 < x < 1, \tag{7.2.7}$$

subject to boundary conditions $u(0,t) = 0 = u(1,t)$ and some initial condition. It was shown in Section 3.1.2 that the eigenvalues and eigenfunctions of the spatial differential operator $\partial^2/\partial x^2$ with the given boundary conditions are respectively given by

$$\lambda_k = k\pi \quad \text{and} \quad U_k = \sin k\pi x, \quad k = 1, 2, \ldots. \tag{7.2.8}$$

Note that λ_k are the eigenfrequencies of the vibrations, but $-\lambda_k^2$ are the actual eigenvalues of the spatial differential operator $\partial^2/\partial x^2$. Once a finite difference method with an equidistant grid for which $h = \frac{1}{N}$ (where h represents the stepsize) has been used, it follows (see again Section 3.1.2) that the eigenvalues and eigenvectors of the discretized problem are respectively given by

$$\lambda_{hk} = \frac{2}{h}\sin\left(\tfrac{1}{2}k\pi h\right) \quad \text{and} \quad \boldsymbol{U}_k = \begin{pmatrix} \sin(k\pi h) \\ \sin(2k\pi h) \\ \ldots \\ \sin((N-1)k\pi h) \end{pmatrix}. \tag{7.2.9}$$

Note that $-\lambda_{hk}^2$ are the actual eigenvalues of the spatial discretization matrix $M^{-1}S$ of the discretized problem (see (3.1.20)). Note that the eigenvectors are exact. It can be demonstrated that $|\lambda_1 - \lambda_{h1}| = \mathcal{O}(h^2)$ and that for $k = \frac{N}{2}$ the phase-shift error is already significant. In the following exercise, the claims that we made in this paragraph must be proved from scratch, without using the results from Section 3.1.2.

Exercise 7.2.2 *Consider the initial boundary value problem in Equation* (7.2.7).

- *Verify by substitution that the eigenfunctions U_k and eigenvalues $-\lambda_k^2$ of the spatial differential operator $\partial^2/\partial x^2$ are given by (7.2.8).*

- *Use the Finite Difference Method to create an equidistant discretization for which $h = \frac{1}{N}$, with h representing the stepsize.*

- *Verify by substitution that the eigenvectors $\boldsymbol{U}_k$ and eigenvalues $-\lambda_{hk}^2$ of the spatial discretization matrix are given by (7.2.9). Note that the eigenvectors are exact. Further, show that $|\lambda_1 - \lambda_{h1}| = \mathcal{O}(h^2)$ and that for $k = \dfrac{N}{2}$ the phase-shift error is already significant.*

□

7.3 Numerical time integration

One possibility to integrate Equation (7.2.1) numerically is to write it as a system of first order differential equations with respect to time:

$$\begin{aligned} \frac{d\boldsymbol{u}}{dt} &= \boldsymbol{v}, \\ M\frac{d\boldsymbol{v}}{dt} &= c^2 S\boldsymbol{u} + \boldsymbol{f}, \end{aligned} \tag{7.3.1}$$

with initial conditions $\boldsymbol{u}(t_0) = \boldsymbol{u}_0$ and $\boldsymbol{v}(t_0) = \boldsymbol{v}_0$. For this system the ordinary numerical methods for initial value problems can be used.

Example 7.3.1 *Forward Euler applied to System* (7.3.1) *gives*

$$\begin{aligned} \frac{\boldsymbol{u}^{n+1}}{\Delta t} &= \frac{\boldsymbol{u}^n}{\Delta t} + \boldsymbol{v}^n, \\ M\frac{\boldsymbol{v}^{n+1}}{\Delta t} &= M\frac{\boldsymbol{v}^n}{\Delta t} + c^2 S\boldsymbol{u}^n + \boldsymbol{f}^n. \end{aligned} \tag{7.3.2}$$

Exercise 7.3.1 *Give the equations for $\boldsymbol{u}$ and $\boldsymbol{v}$ when a Crank-Nicolson time integration of System* (7.3.1) *is applied.* □

7.4 Stability of the numerical integration

From the conservation of energy of the solutions of both the wave equation and the discretization based on the method of lines, it follows that asymptotic stability does not make much sense here. A perturbation of the initial conditions will never vanish. A *fundamental solution* of the form $\boldsymbol{u}(t) = e^{\lambda ct}\boldsymbol{u}, \boldsymbol{v}(t) = e^{\lambda ct}\boldsymbol{v}$ of system (7.3.1) with $\boldsymbol{f} = \boldsymbol{0}$ has a purely imaginary λ as is shown in the next theorem.

Theorem 7.4.1 *Consider system* (7.3.1) *and let λ be an eigenvalue of the generalized eigenvalue problem*

$$\begin{aligned} \lambda c\boldsymbol{u} &= \boldsymbol{v}, \\ \lambda cM\boldsymbol{v} &= c^2 S\boldsymbol{u}. \end{aligned} \tag{7.4.1}$$

If M is symmetric positive definite and S symmetric negative definite, then, the eigenvalues of this generalized eigenvalue problem are purely imaginary.

Proof: We use the upper equation to eliminate $\boldsymbol{v}$ from the lower equation,

$$\lambda^2 M\boldsymbol{u} = S\boldsymbol{u}, \tag{7.4.2}$$

which shows that λ^2 is an eigenvalue of the generalized eigenvalue problem for M and S. Hence λ^2 is an eigenvalue of the matrix $M^{-1}S$. It follows from

Lemma 6.5.1 that λ^2 is real and negative, implying that λ is purely imaginary. □

With the purely imaginary eigenvalues of the above generalized eigenvalue problem (7.4.1), it follows that the solution of system (7.3.1) is neutrally stable. An absolutely stable time integration method decays the error of the solution and also the solution itself as $t \to \infty$. An unstable time integration method blows up the error and the solution. This implies that with neither of these time integration methods, the wave equation can be integrated numerically up to any large time t. Hence we have to define an *end time* T and choose the time step Δt accordingly small. If $T = n\Delta t$ and $\lim_{\Delta t \to 0} |C(\lambda \Delta t)|^n = 1$ for a particular method, then the wave equation can be integrated up to this bounded time T. Note that $n \to \infty$ as $\Delta t \to 0$.

7.5 Total dissipation and dispersion

Since the eigenvalues of (7.4.1) are purely imaginary, the solution of (7.3.1) can be written as a linear combination of products of eigenvectors and undamped vibrations. Hence it is sufficient to consider a single differential equation of the form

$$\frac{dw}{dt} = i\mu w, \text{ subject to } w(t_0) = w_0. \tag{7.5.1}$$

The behavior of this differential equation qualitatively reflects the behavior of the total system (7.3.1). The exact solution is

$$w(t) = w_0 e^{i\mu(t-t_0)}. \tag{7.5.2}$$

For the solution at $t^{n+1} = t_0 + (n+1)\Delta t$ we note that

$$w(t^{n+1}) = w(t^n) e^{i\mu\Delta t}. \tag{7.5.3}$$

Hence the *amplification factor* of the exact solution is given by

$$C(i\mu\Delta t) = e^{i\mu\Delta t} \Rightarrow |C(i\mu\Delta t)| = 1 \text{ and } \arg(C(i\mu\Delta t)) = \mu\Delta t. \tag{7.5.4}$$

The argument of the amplification factor, $\arg(C(i\mu\Delta t))$, is referred to as the *phase shift*. Hence in each time step there is a phase shift in the exact solution, whereas the modulus of the exact solution does not change.

Exercise 7.5.1 *Show that the complex differential equation* (7.5.1) *is equivalent to the system*

$$\begin{aligned} \frac{du}{dt} &= -\mu v \\ \frac{dv}{dt} &= \mu u, \end{aligned} \tag{7.5.5}$$

where $u = Re\{w\}$ *and* $v = Im\{w\}$. *Show that* $|w(t)| = Constant$ *is equivalent to conservation of energy.* □

For the numerical method, the following relation holds

$$w^{n+1} = C(i\mu\Delta t)w^n. \tag{7.5.6}$$

If the modulus of the amplification factor is larger than one, the energy increases in each time step. This is called *amplification*. Conversely, if the amplification factor is smaller than one the energy decreases. This is called *dissipation*.

Example 7.5.1 *The modulus of the amplification factor of Euler's method is*

$$|C(i\mu\Delta t)| = \sqrt{1 + (\mu\Delta t)^2}. \tag{7.5.7}$$

Hence the amplification of the method is $\mathcal{O}(\mu^2(\Delta t)^2)$ accurate.

The phase shift per time step of a numerical method is defined by the argument of the amplification factor, i.e.

$$\Delta\Phi = \arg(C(i\mu\Delta t)) = \arctan\left(\frac{\mathrm{Im}\{C\}}{\mathrm{Re}\{C\}}\right). \tag{7.5.8}$$

Remark: the last equals sign is only true if the argument is between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, which is the case if Δt is small enough.

Example 7.5.2 *The phase shift of the improved Euler method is given by*

$$\Delta\Phi = \arctan\left(\frac{\mu\Delta t}{1 - \frac{1}{2}(\mu\Delta t)^2}\right). \tag{7.5.9}$$

The phase error or *dispersion* is the difference between the exact and numerical phase shifts. This is referred to as *dispersion* because the phase shifts differ for the different values of μ_k in Equation (7.5.1).

Exercise 7.5.2 *Show that the phase error of the improved Euler method per time step is $\mathcal{O}((\mu\Delta t)^3)$.* □

The *total dissipation*, $D_n(i\mu\Delta t)$, is the product of the dissipations of all the time steps from t_0 up to the end time T. The *total dispersion*, $\Delta\Phi_n(i\mu\Delta t)$, is the sum over the phase errors of all the time steps. Note that we have $n\Delta t = T - t_0$. The total dissipation and the total dispersion are measures of the error in the numerical solution. As $\Delta t \to 0$ the total dissipation should tend to 1 and the total dispersion should tend to 0.

Exercise 7.5.3 *Why do we need*

$$\lim_{\Delta t \to 0} D_n(i\mu\Delta t) = 1? \tag{7.5.10}$$

□

As an illustration we calculate the total dissipation and total dispersion for the forward Euler method:

$$D_n = |C(i\mu\Delta t)|^n = (1 + (\mu\Delta t)^2)^{\frac{T-t_0}{2\Delta t}}. \tag{7.5.11}$$

From a Taylor series of the exponential, we see that

$$1 \leq D_n \leq \left[\exp\left((\mu\Delta t)^2\right)\right]^{\frac{T-t_0}{2\Delta t}}. \tag{7.5.12}$$

Subsequently, from a linearization of the exponential, we get

$$\exp\left((\mu\Delta t)^2 \frac{T-t_0}{2\Delta t}\right) = 1 + \mathcal{O}(\mu^2\Delta t). \tag{7.5.13}$$

So the condition $\lim_{\Delta t\to 0} D_n(i\mu\Delta t) = 1$ is satisfied. For the total dispersion we have

$$\begin{aligned} &\Delta\Phi_n(i\mu\Delta t) = n(\mu\Delta t - \Delta\Phi) = n(\mu\Delta t - \arctan(\mu\Delta t)) = \\ &n(\mu\Delta t - (\mu\Delta t + \mathcal{O}((\mu\Delta t)^3))) = n\mathcal{O}((\mu\Delta t)^3) = \mathcal{O}(\mu^3(\Delta t)^2). \end{aligned} \tag{7.5.14}$$

Note that $n\Delta t = T - t_0$ and that the exact phase shift is $\mu\Delta t$. This has been used in this expression. It is clear from the expression that the total dispersion tends to zero as the time step tends to zero. In Figures 7.1 and 7.2 the total dissipation and dispersion are plotted as a function of the time step Δt.

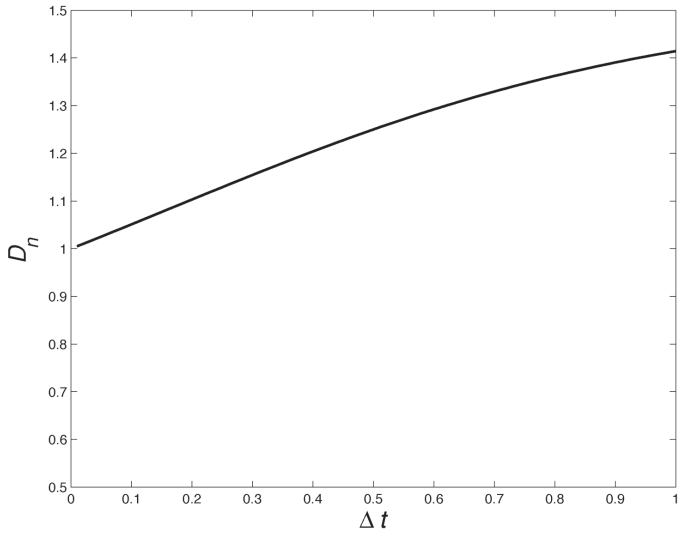

Figure 7.1: Total dissipation of the forward Euler method for $\mu = 1$, $T - t_0 = 1$.

The total dissipation D_n and total dispersion $\Delta\Phi_n$ can be investigated for other time integration methods as well. We leave this as an exercise to the reader.

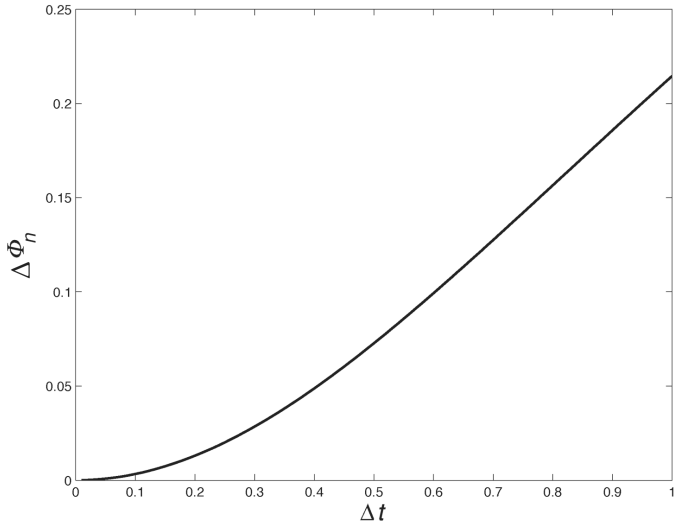

Figure 7.2: Total dispersion of the forward Euler method for $\mu = 1$, $T - t_0 = 1$.

7.6 Direct time integration of the second order system

In principle it is not necessary to write Equation (7.2.1) as a system (7.3.1) of two first order differential equations. A lot of methods are available to integrate a second order differential equation of the form

$$\frac{d^2\boldsymbol{y}}{dt^2} = f(\boldsymbol{y}, t) \tag{7.6.1}$$

directly. For a comprehensive survey of numerical methods to solve this system of second order differential equations we refer to [8]. In this course we will treat two example schemes applied to (7.2.1):

1. Explicit scheme:

$$M\boldsymbol{u}^{n+1} - 2M\boldsymbol{u}^n + M\boldsymbol{u}^{n-1} = (\Delta t)^2 \left(c^2 S\boldsymbol{u}^n + \boldsymbol{f}^n \right). \tag{7.6.2}$$

2. Implicit scheme:

$$M\boldsymbol{u}^{n+1} - 2M\boldsymbol{u}^n + M\boldsymbol{u}^{n-1} = \frac{(\Delta t)^2}{4} \left(c^2 (S\boldsymbol{u}^{n+1} + 2S\boldsymbol{u}^n + S\boldsymbol{u}^{n-1}) + \boldsymbol{f}^{n+1} + 2\boldsymbol{f}^n + \boldsymbol{f}^{n-1} \right). \tag{7.6.3}$$

Both methods are consistent of $\mathcal{O}((\Delta t)^2)$ in time. These methods are referred to as *three-level* schemes because they are defined as a recurrence relation involving three time levels. This implies that these schemes are not self-starting: one first has to do one step of a two-level method, such as Euler explicit:

$$\boldsymbol{u}_1 = \boldsymbol{u}_0 + \Delta t \boldsymbol{v}_0. \tag{7.6.4}$$

Using the explicit Euler method for the first step is satisfactory, since its error for the first step is $\mathcal{O}(\Delta t^2)$.

Equations (7.6.2) and (7.6.3) are special cases of the popular Newmark-(β, γ) scheme. This scheme is usually written in a form based on displacement $\boldsymbol{u}$, velocity $\boldsymbol{v}$ and acceleration $\boldsymbol{a}$. It uses a Taylor expansion, where the higher order terms are averaged.

The Newmark scheme reads:

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \Delta t \boldsymbol{v}^n + \frac{(\Delta t)^2}{2}((1-2\beta)\boldsymbol{a}^n + 2\beta \boldsymbol{a}^{n+1}), \tag{7.6.5}$$

$$\boldsymbol{v}^{n+1} = \boldsymbol{v}^n + \Delta t((1-\gamma)\boldsymbol{a}^n + \gamma \boldsymbol{a}^{n+1}), \tag{7.6.6}$$

$$M\boldsymbol{a}^{n+1} - c^2 S \boldsymbol{u}^{n+1} = \boldsymbol{f}^{n+1}. \tag{7.6.7}$$

At $t = t_0$ we solve $\boldsymbol{a}^0$ from the equation of motion (7.6.7). In the following steps we substitute (7.6.5) in (7.6.7) to get an equation for $\boldsymbol{a}^{n+1}$. Finally (7.6.5) and (7.6.6) are used to compute $\boldsymbol{u}^{n+1}$ and $\boldsymbol{v}^{n+1}$.
It is possible to rewrite Newmark as a three-level scheme for the displacements $\boldsymbol{u}$:

$$\begin{aligned}(M - \beta c^2 (\Delta t)^2 S)\boldsymbol{u}^{n+1} - (2M + (\frac{1}{2} + \gamma - 2\beta)c^2(\Delta t)^2 S)\boldsymbol{u}^n + \\ (M - (\frac{1}{2} - \gamma + \beta)c^2(\Delta t)^2 S)\boldsymbol{u}^{n-1} = (\Delta t)^2 \boldsymbol{F}^n,\end{aligned} \tag{7.6.8}$$

with

$$\boldsymbol{F}^n = (\frac{1}{2} - \gamma + \beta)\boldsymbol{f}^{n-1} + (\frac{1}{2} + \gamma - 2\beta)\boldsymbol{f}^n + \beta \boldsymbol{f}^{n+1}. \tag{7.6.9}$$

Remark
At $t = t_0$, (7.6.7) can not be used to compute $\boldsymbol{a}^0$ at boundaries with prescribed displacements. Why not? In practice one often takes $\boldsymbol{a}^0 = \boldsymbol{0}$ in that case.
An alternative is to use a Taylor series expansion at $t = t_0 + \Delta t$ and to express $\boldsymbol{a}^0$ in $\boldsymbol{u}^0$, $\boldsymbol{v}^0$, and $\boldsymbol{u}^1$ at that boundary.

Exercise 7.6.1 *Prove that (7.6.8), (7.6.9) follows from (7.6.5)-(7.6.7).*
Hint: Eliminate $\boldsymbol{v}^n$ from (7.6.5) *by using* (7.6.6)*, and replace in the resulting equation the index n by $n-1$:*

$$\boldsymbol{u}^{n-1} = \boldsymbol{u}^n - \Delta t \boldsymbol{v}^n + \frac{\Delta t^2}{2}\left[(1 - 2(\gamma - \beta))\boldsymbol{a}^{n-1} + 2(\gamma - \beta)\boldsymbol{a}^n\right]. \tag{7.6.10}$$

Add (7.6.5) and (7.6.10) to get a relation between $\boldsymbol{u}^{n-1}$, $\boldsymbol{u}^n$, $\boldsymbol{u}^{n+1}$, $\boldsymbol{a}^{n-1}$, $\boldsymbol{a}^n$, $\boldsymbol{a}^{n+1}$. *Then use the equation of motion* (7.6.7) *to eliminate* $\boldsymbol{a}^{n-1}$, $\boldsymbol{a}^n$, $\boldsymbol{a}^{n+1}$. □

Exercise 7.6.2 *Show that the Newmark scheme reduces to the explicit central difference scheme* (7.6.2) *if* $\beta = 0$ *and* $\gamma = \frac{1}{2}$. □

Exercise 7.6.3 *Show that the Newmark scheme reduces to the implicit central difference scheme* (7.6.3) *if* $\beta = \frac{1}{4}$ *and* $\gamma = \frac{1}{2}$. □

Exercise 7.6.4 *Show that the three-level implicit scheme (7.6.3) is identical to Crank-Nicolson's method for* (7.3.1). *(Hint: write out the steps for n and* $n+1$ *and eliminate all the* $\boldsymbol{v}$*'s.) Note that the first step of the three-level method should be taken with Crank-Nicolson's method instead of the previously mentioned Euler explicit method.* □

7.7 The CFL criterion

From the section about the numerical time integration, it is clear that the time step plays an important role in the numerical integration. In general the time step Δt and stepsize Δx cannot be chosen independently. This was already observed for Euler's method. In 1928 Courant, Friedrichs and Lewy formulated a condition for the time step for the numerical solution to be a representation of the exact solution. Their condition was obtained by using a physical argument. Commonly one refers to it as the CFL criterion. Often this CFL condition is used in relation with stability of a numerical method. Strictly, this is not true since the CFL criterion represents a condition for convergence. In the following text an intuitive justification of the CFL criterion will be given. It is possible though to derive the CFL criterion in full mathematical rigor.

The solution of the wave equation can be represented by a superposition of linear waves, which all have a velocity c. Consider the solution at any node x_i at time t^j, then, within a time interval Δt, this *point source* influences the solution within the distance $c\Delta t$ from position x_i. Within a time interval Δt, the solution at locations with distance larger than $c\Delta t$ from x_i is not influenced by the solution at x_i on t^j. In this way we obtain the (analytical) *forward cone of influence* of $u(x_i, t^j)$. Vice versa, $u(x_i, t^{j+1})$ is determined by the *point sources* of $u(x, t^{j+1} - \tau)$, where $\tau > 0$ and $|x - x_i| < c\tau$. This leads to the (analytical) *backward cone of influence* of $u(x_i, t^{j+1})$ usually referred to as the (analytical) *domain of dependence* of $u(x_i, t^{j+1})$ and indicated by the grey region in Figure 7.3. For the explicit time integration of the wave equation, the spatial discretization is done at time t^j. For the finite differences solution with one spatial coordinate at x_i on t^j, one uses $u(x_i, t^j)$, $u(x_{i-1}, t^j)$ and $u(x_{i+1}, t^j)$, i.e.

$$\left.\frac{d^2u}{dx^2}\right|_{t=t^j} = \frac{u(x_{i-1}, t^j) - 2u(x_i, t^j) + u(x_{i+1}, t^j)}{(\Delta x)^2}. \tag{7.7.1}$$

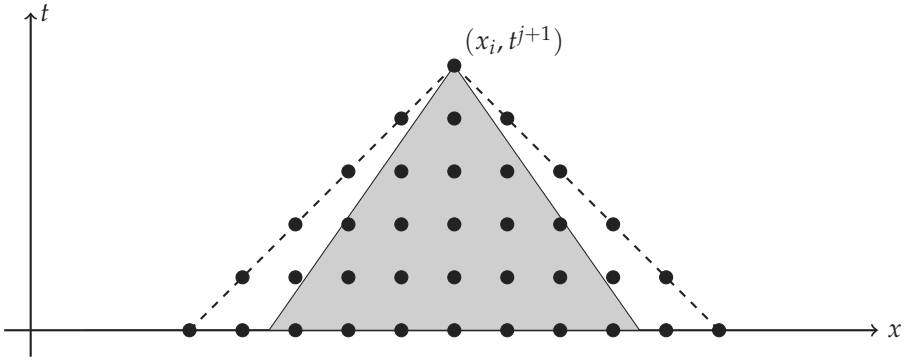

Figure 7.3: In this situation the CFL condition is satisfied because the analytical domain of dependence (the grey cone) is contained in the numerical domain of dependence (the cone bounded by the dotted lines).

If we use explicit time integration, the recursive computation of the finite difference approximation of $u(x_i, t^{j+1})$ involves many grid points. The cone corresponding to these grid points is the *numerical backward cone of influence* of $u(x_i, t^{j+1})$, and is usually referred to as the *numerical domain of dependence* of $u(x_i, t^{j+1})$.

The CFL criterion of an explicit scheme for the wave equation is as follows: *The numerical domain of dependence must contain the analytical domain of dependence.*

The CFL criterion guarantees that the numerical solution has access to all the point sources that physically have an influence on this solution. In the case of Figure 7.3, it turns out that the numerical domain of dependence is a wider cone than the analytical domain of dependence (the grey region) and hence for this Δt the CFL criterion is satisfied and convergence of the numerical solution is to be expected. An example of a time step *not* satisfying the CFL criterion is shown in Figure 7.4. Before we present an example showing the derivation of the CFL criterion for an actual method, we make the following remarks.

Remarks

1. For an implicit scheme the CFL criterion is satisfied for all $\Delta t > 0$ since the numerical solution at a point (x_i, t^{n+1}) depends on all numerical solution values at the previous time level t^n.

2. In the literature the analytical and numerical domain of dependence are actually not exactly defined as a cone, but as the intersection of the cone with the line segments on which the initial and boundary conditions are defined. This slight difference is not essential, however, and does not alter our conclusions.

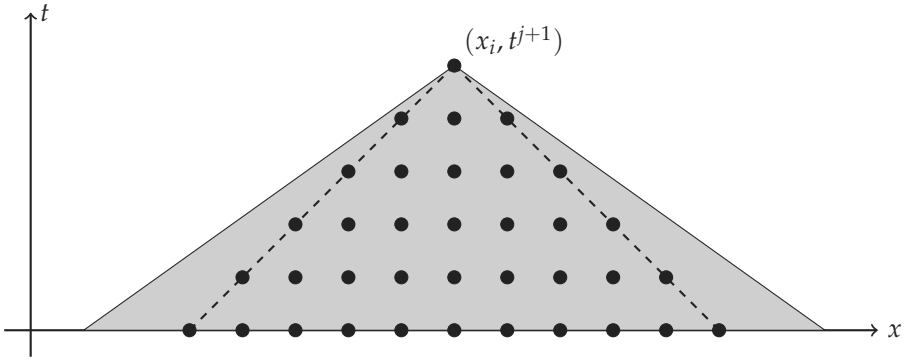

Figure 7.4: In this situation the CFL condition is violated because the analytical domain of dependence (the grey cone) is not contained in the numerical domain of dependence (the cone bounded by the dotted lines).

We now present an example of the derivation of the CFL criterion for an actual method:

Example 7.7.1 *Consider the explicit time integration of the wave equation in one dimension with equidistant nodes:*

$$u_i^{n+1} - 2u_i^n + u_i^{n-1} = \left(\frac{c\Delta t}{\Delta x}\right)^2 (u_{i+1}^n - 2u_i^n + u_{i-1}^n). \tag{7.7.2}$$

The analytical domain of dependence of u_i^{n+1} is the cone with apex (x_i, t^{n+1}) and slopes $\pm 1/c$, which consists of all points (x,t) with $|x - x_i| \leq c(t^{n+1} - t)$. The numerical domain of dependence is the cone with the same apex but with slopes $\pm\frac{\Delta t}{\Delta x}$. Hence, the CFL criterion for this case is given by $\frac{\Delta t}{\Delta x} \leq \frac{1}{c}$, which is equivalent to

$$\frac{c\Delta t}{\Delta x} \leq 1. \tag{7.7.3}$$

Exercise 7.7.1 *Show that for the wave equation with one spatial coordinate, the Euler forward method defined by*

$$\begin{aligned} u_i^{n+1} &= u_i^n + \Delta t v_i^n \\ v_i^{n+1} &= v_i^n + \frac{c^2\Delta t}{(\Delta x)^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n), \end{aligned} \tag{7.7.4}$$

can be written in the form

$$u_i^{n+2} - 2u_i^{n+1} + u_i^n = \left(\frac{c\Delta t}{\Delta x}\right)^2 \left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right). \tag{7.7.5}$$

Use this last formula to determine the numerical domain of dependence, and show that the CFL criterion is given by

$$\frac{c\Delta t}{\Delta x} \leq \frac{1}{2}. \tag{7.7.6}$$

□

Exercise 7.7.2 *Show that if the first equation in* (7.7.4) *in Exercise 7.7.1 is replaced by*

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{4}(v_{i-1}^n + 2v_i^n + v_{i+1}^n), \tag{7.7.7}$$

the system can be written in the form:

$$u_i^{n+2} - 2u_i^{n+1} + u_i^n = \left(\frac{c\Delta t}{2\Delta x}\right)^2 \left(u_{i+2}^n - 2u_i^n + u_{i-2}^n\right). \tag{7.7.8}$$

Use this last formula to determine the numerical domain of dependence, and show that the CFL criterion is given by

$$\frac{c\Delta t}{\Delta x} \leq 1. \tag{7.7.9}$$

□

7.8 Summary of Chapter 7

This chapter has dealt with numerical methods for the solution of the (hyperbolic) wave equation. The hyperbolic nature of the wave equation is important for the nature of the numerical solutions. To solve the PDE the method of lines has been used. It first deals with the spatial derivatives and considers time integration of the resulting system of ODEs as a separate problem.

A direct time integration scheme for the second time derivative has also been presented. The numerical amplification factor, which determines dissipation and phase shift of the numerical solution, has been defined and analyzed. Finally, the derivation of the CFL-criterion, using the concept of the (analytical and numerical) domain of dependence in the x, t plane, has been given. This CFL criterion is necessary for the numerical solution to be a representation of the exact solution.

Bibliography

[1] R.A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[2] Robert A. Adams. *Calculus, a complete course. Fifth Edition.* Addison Wesley Longman, Toronto, 2003.

[3] R. Aris. *Vectors, Tensors and the Basic Equations of Fluid Mechanics*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962. Reprinted, Dover, New York, 1989.

[4] R.L. Burden and J.D. Faires. *Numerical analysis*. Brooks/Cole, Pacific Grove, 2001.

[5] R. Courant and D. Hilbert. *Methods of Mathematical Physics, Vol. 2. Partial Differential Equations*. Interscience, New York, 1989.

[6] C. Cuvelier, A. Segal, and A.A. van Steenhoven. *Finite Element Methods and Navier-Stokes Equations*. Reidel Publishing Company, Dordrecht, Holland, 1986.

[7] L.C. Evans. *Partial Differential Equations*. American Mathematical Society, 2010.

[8] J.D. Lambert. *Numerical methods in ordinary differential equations*. John Wiley, Englewood Cliffs, 1991.

[9] David C. Lay. *Linear Algebra and its applications*. Addison Wesley, New York, 1993.

[10] A.R. Mitchell and D.F. Griffiths. *The Finite Difference Method in Partial Differential Equations*. Wiley, Chichester, 1994.

[11] M.H. Protter and H.F. Weinberger. *Maximum Principles in Differential Equations*. Prentice-Hall, Englewood Cliffs, 1967.

[12] James Stewart. *Calculus. Fifth Edition*. Brooks/Cole, New York, 2002.

[13] G. Strang. *Linear Algebra and its Applications, (third edition)*. Harcourt Brace Jovanovich, San Diego, 1988.

[14] R. Temam. *Navier-Stokes Equations*. North-Holland, Amsterdam, 1985.

[15] J. van Kan, A. Segal, and F. Vermolen. *Numerical Methods in Scientific Computing, Second Edition*. Delft Academic Press, Delft, 2014.

Index

Classical Numerical Methods in Scientific Computing

Jos van Kan, Guus Segal, Fred Vermolen

Partial differential equations are paramount in mathematical modelling with applications in engineering and science. The book starts with a crash course on partial differential equations in order to familiarize the reader with fundamental properties such as existence, uniqueness and possibly existing maximum principles. The main topic of the book entails the description of classical numerical methods that are used to approximate the solution of partial differential equations. The focus is on discretization methods such as the finite difference, finite volume and finite element method. The manuscript also makes a short excursion to the solution of large sets of (non)linear algebraic equations that result after application of discretization method to partial differential equations. The book treats the construction of such discretization methods, as well as some error analysis, where it is noted that the error analysis for the finite element method is merely descriptive, rather than rigorous from a mathematical point of view. The last chapters focus on time integration issues for classical time-dependent partial differential equations. After reading the book, the reader should be able to derive finite element methods, to implement the methods and to judge whether the obtained approximations are consistent with the solution to the partial differential equations. The reader will also obtain these skills for the other classical discretization methods. Acquiring such fundamental knowledge will allow the reader to continue studying more advanced methods like meshfree methods, discontinuous Galerkin methods and spectral methods for the approximation of solutions to partial differential equations.



Jos van Kan, Retired professor Delft University of Technology, Delft Institute of Applied Mathematics

Guus Segal, Retired professor Delft University of Technology, Delft Institute of Applied Mathematics

Fred Vermolen, University of Hasselt, Department of Mathematics and Statistics, Computational Mathematics Group

TU Delft

ISBN 978-94-6366-732-6
DOI https://doi.org/10.59490/t.2023.007

textbooks.open.tudelft.nl