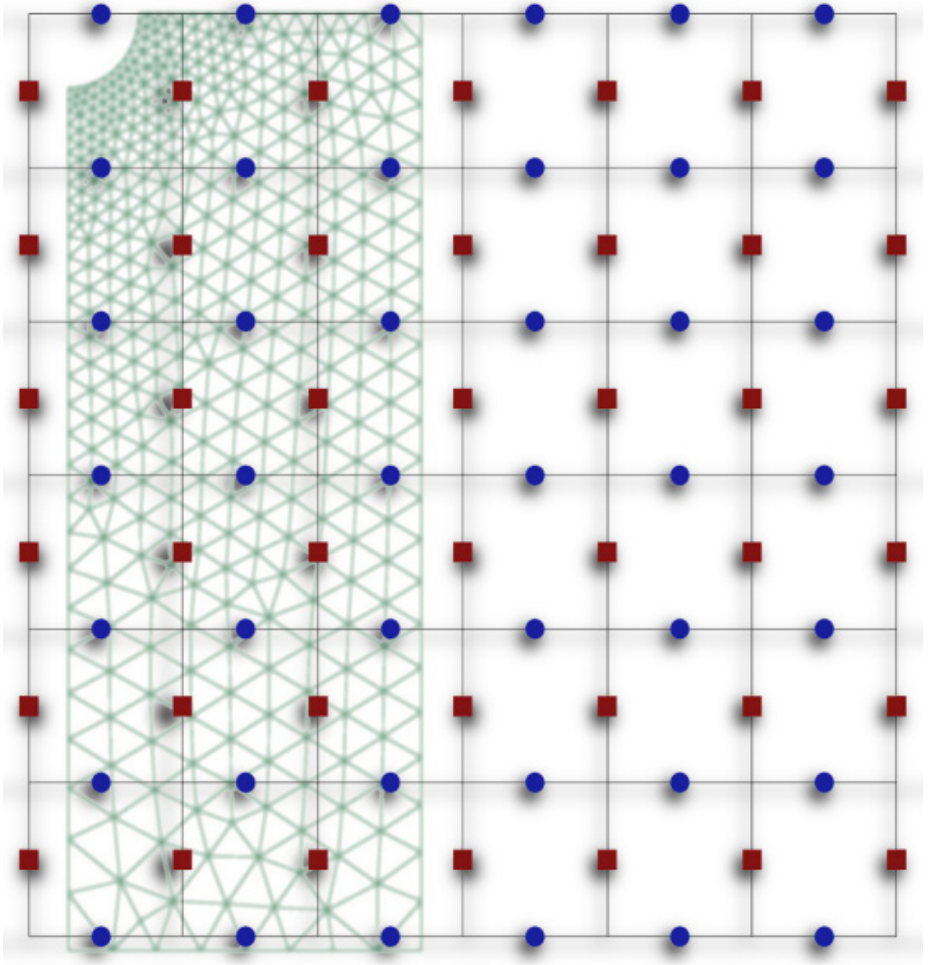
Numerical Methods in Scientific Computing

Jos van Kan, Guus Segal, Fred Vermolen



TU Delft OPEN

Numerical Methods in Scientific Computing

Jos van Kan (1944) graduated in 1968 from Delft University of Technology, Delft, Netherlands, in Numerical Analysis and has been assistant professor at the Department of Mathematics of that institute ever since. He wrote several articles on Numerical Fuid Mechanics (pressure correction methods) and has written a multigrid pressure solver for the Delft software package to solve the Navier Stokes equations. He has been teaching classes in Numerical Analysis since 1971 and wrote several books on the subject.

Guus Segal (1948) graduated in 1971 from Delft University of Technology, Delft, Netherlands, in Numerical Analysis and has been part time assistant professor at the Department of Mathematics of that institute ever since. He is also working in the consultancy and numerical software company SEPRA in Den Haag, The Netherlands. He wrote a number of articles on Finite Element Methods and several articles on curvilinear Finite Volume Methods and Numerical Fluid Mechanics. He has written a book on Finite Element Methods and Navier-Stokes equations. He is the main developer of the finite element package SEPRAN. He has been teaching classes in Numerical Analysis since 1973.

Fred Vermolen (1969) graduated in 1993 from Delft University of Technology, Delft, Netherlands. He wrote his PhD-thesis supervised by the promotores prof Pieter Wesseling (Numerical Analysis) and prof Sybrand van der Zwaag (Materials Science). He wrote several articles on Stefan problems and transport in porous media. His present interest is in mathematical issues in medicine. He has been teaching courses in Numerical Analysis since 2002.

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 edition 2005, second edition 2014
ISBN print version: 97890-6562-3638
ISBN electronic version: 978-90-6562-3645

ISBN paperback: 978-94-6366-738-8
ISBN Ebook: 978-94-6366-740-1
DOI: https://doi.org/10.59490/t.2023.009

NUR 919

Key words: numerical mathematics

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. Since this treatment had to be on the superficial side we have provided further reference to the literature where necessary.

Delft, June 2005

Jos van Kan
Guus Segal
Fred Vermolen

Note to the first edition improvements

In this improved first edition exercises and theory are more separately presented. Furthermore, some parts, such as the parts on boundary fitted coordinates, on coordinate transformation, the treatment of essential boundary conditions for FEM and the solution of non-linear systems of equations, have been rewritten to make them easier to understand.
Newmark-type solvers for the wave equation have been added.

Delft, April 2008

Jos van Kan
Guus Segal
Fred Vermolen

Note to the second edition improvements

In this improved second edition the treatment of boundary conditions for all types of discretization methods has been extended. Periodical boundary conditions have been included. Furthermore,the description of the FEM has been simplified.

Delft, August 2014

Guus Segal
Fred Vermolen

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 back bone 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

We first take a look at second order partial differential equations and their relation with various physical problems. Then we look at numerical methods for those equations. First we look at finite difference methods, of respectable age but still very much in use. Subsequently we take on finite volume methods, a typical engineers option, constructed for conservation laws. Finally we turn to finite element methods (FEM) which have gained tremendous popularity over the last decades. Before we can move to FEM, however, we have to delve a bit into minimization problems to provide a proper background. We shall show, that FEM may be considered as a special case of Ritz's method, a particular way of obtaining an approximate solution to a minimization problem. We shall establish a relation between minimization problems and partial differential equations. But not all PDEs can be formulated as a minimization problem and we shall consider a generalization that will enable us to apply the FEM also to those problems.

These methods generally leave us with a large set of linear or non-linear equations and we consider ways of how to solve them. In particular we shall pay some attention to efficient methods that are relatively young, like preconditioned Krylov space methods and multi-grid methods. The treatment can be only cursory but further references will be provided.

We also pay some attention to special methods for specific problems like heat and wave equations. Finally we consider transport equations. They do not fall within the previous context, being only first order, yet they are very important and deserve a chapter of their own. The last chapter will be dedicated to miscellaneous problems that fall outside the classification so far.

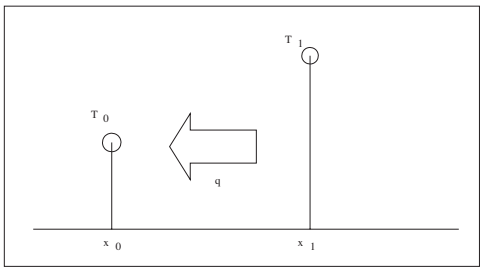


Figure 1.1: 1-dimensional heat flow.

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

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

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. The simplest one dimensional model to reflect this is the following linear model. Let q be the generated flow, directly proportional to the temperature difference ΔT and inversely proportional to the distance Δx. This leads to:

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

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 continuous 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.

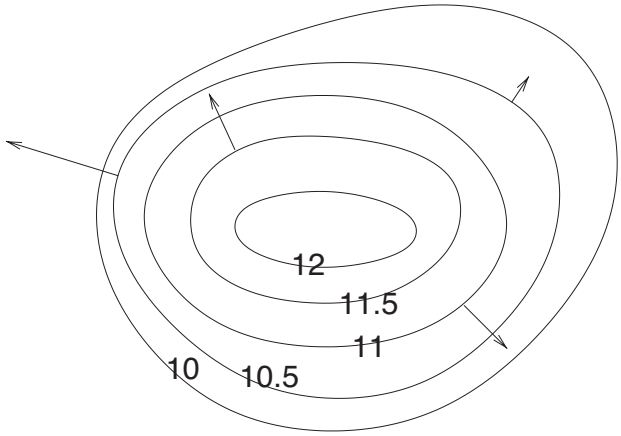


Figure 1.2: Isotherms.

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 $\mathbf{v}$ and $\mathbf{w}$ are orthogonal if their inner product $(\mathbf{v}, \mathbf{w})$ 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 equation 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:

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

If we compare Equation (1.3.4) with (1.3.5) we see that these can only be satisfied if

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

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 *(Darcy's Law). In ground water flow the velocities are very small, a few centimeters per day. This makes ground water flow basically a* hydrostatic *problem, in which the flow is driven by differences in hydrostatic pressure. This hydrostatic pressure depends linearly on the height of the ground water level h. So how does the flow* $\mathbf{q}$ *depend on h?* □

Exercise 1.3.2 *(Fick's Law) In diffusion the flow of matter,* $\mathbf{q}$*, is driven by differences in concentration c. Express* $\mathbf{q}$ *in c.* □

Scalar fields like T, h and c that drive a gradient flow field, $\mathbf{q}$, are called *potentials*. Not all flow fields are generated by the gradient of a potential. But those that are, are called *solenoidal* or *irrotational.*

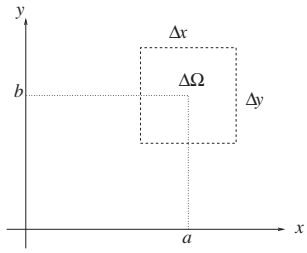


Figure 1.3: Square volume in river.

Exercise 1.3.3 *Let C be a closed contour in the x-y-plane and* $\mathbf{q}$ *a solenoidal vector field. Show that* $\int_C \mathbf{q} \cdot d\mathbf{s} = 0$. □

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 $\frac{dT}{ds}$ depends on the parameterization. A natural parameterization is the *arc length* of the curve.
Note, that in that case $(\frac{dx}{ds})^2 + (\frac{dy}{ds})^2 = 1$. This forms the basis of the following definition:

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

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

Exercise 1.3.4 *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 1.3.5 *For what value of* $\mathbf{n}$ *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 $\mathbf{v}$ is defined by:

$$\text{div}\, \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y}. \tag{1.3.7}$$

For $\mathbb{R}^3$ you have the obvious generalization and there is also a nabla notation: $\text{div}\, \mathbf{v} = \nabla \cdot \mathbf{v}$. 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 (see Figure 1.3).

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

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

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

$$\lim_{(\Delta x,\Delta y)\to(0.0)} \frac{\Delta\phi(a,b)}{\Delta x\Delta y} = \operatorname{div}\mathbf{v}(a,b). \tag{1.3.9}$$

From this formula, we see that $\operatorname{div}\mathbf{v}(a,b)$ is the outflow density (outflow per unit area) at point (a,b). Integration of the outflow density over an entire domain 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 domain 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.

Exercise 1.3.6 *Explain that for an* incompressible *flow field,* $\mathbf{u}$*, we must have* $\operatorname{div}\mathbf{u} = 0$. □

Exercise 1.3.7 *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.

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* $\mathbf{n}$ *be the outward normal and* $\mathbf{v}$ *a continuously differentiable vector field. Then*

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

□

Remark

1. The expression $\mathbf{v} \cdot \mathbf{n}$ is the outward normal component of the vector-field, $\mathbf{v}$, 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 [35]). 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 $\Delta x, \Delta y \to 0$ (contraction) in the Riemann sum.

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

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

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

Exercise 1.3.8 *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 a sufficiently smooth c, $\mathbf{u}$, we have

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

Exercise 1.3.9 *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 1.3.10 *Prove Theorem 1.3.4.*
Hint: choose an appropriate vector field, $\mathbf{u}$, in the previous exercise. □

1.3.5 Conservation laws

Let us consider some flow field, $\mathbf{u}$, 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:

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

The term $f(t,\mathbf{x})$ 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, $\mathbf{u}$, is also called the *flux vector* of the model. S just like f has the dimension of a *density.* Since Equation (1.3.15) has to hold for every conceivable volume in the flow field we may formulate a *point wise* conservation law as follows. First we apply Gauss' Theorem 1.3.10 to Equation (1.3.15) to obtain

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

Subsequently we invoke the mean-value theorem of integral calculus for each integral separately, assuming all integrands are continuous:

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

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

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

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

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

in which h is the heat density, $\mathbf{q}$ the heat flux vector and f the production density. Remember, that all quantities in such a point wise conservation law are densities. The heat density, h, stored in a material can be related to the materials (absolute) temperature T:

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

in which ρ is the 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, $\mathbf{q}$, is driven by the temperature gradient: $\mathbf{q} = -\lambda\,\nabla T$. This enables us to formulate everything in terms of temperature. Substituting this all we get:

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

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

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

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

1.4 Minimization

Another way of deriving models is by looking at the potential energy. This is most often used in mechanical problems, but can also be used in different contexts. An equilibrium state can be found by minimizing that potential energy. We also meet minimization problems in optics (optical length) and economics (cost).

1.4.1 Elastic string

As an example consider an elastic string fixed in $(0,0)$ and $(0,1)$, see Figure 1.4.

Without load, the string is undeformed: $u(x) = 0$. When we apply a load f the string deforms. What is the potential energy of the deformed string? First of all, there is an elastic energy proportional to the increase in length: $\Delta P_e = k\Delta L$. Over a small interval Δx this increase amounts to

$$\Delta L = \sqrt{\Delta x^2 + \Delta u^2} - \Delta x. \tag{1.4.1}$$

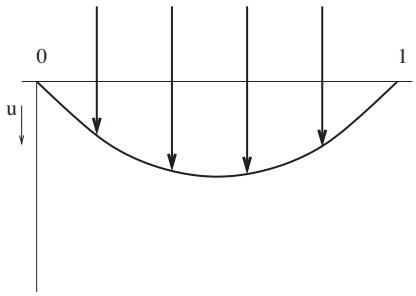


Figure 1.4: Deformed elastic string.

When the inclination $\Delta u/\Delta x$ is small (this is true in a realistic problem), this is approximately equal to

$$\Delta L = \Delta x(1 + \tfrac{1}{2}(\frac{\Delta u}{\Delta x})^2) - \Delta x, \tag{1.4.2}$$

$$= \tfrac{1}{2}\left(\frac{\Delta u}{\Delta x}\right)^2 \Delta x. \tag{1.4.3}$$

The work done by the load f per fragment Δx equals $\Delta W = uf\Delta x$, assuming we take the positive u–axis pointing down. The potential energy per fragment Δx then is given by $\Delta P_e - \Delta W$ and the potential energy over the whole string is obtained by integrating over the whole interval $(0,1)$:

$$P = P_e - W = \int_0^1 (\tfrac{1}{2}k\left(\frac{du}{dx}\right)^2 - uf)\,dx. \tag{1.4.4}$$

So any (sufficiently smooth) function u satisfying $u(0) = 0$ and $u(1) = 0$ yields a potential energy. The solution to the mechanical problem is that function u for which the potential energy P is minimal. In Chapter 5 we shall see how to deal with this.

Exercise 1.4.1 *Show by Taylor's theorem that* $\sqrt{1+x} = 1 + \frac{1}{2}x + O(x^2)$. □

1.5 Preliminaries from linear algebra

Let $\mathbf{x}, \mathbf{y}$ be vectors in $\mathbb{C}^n$. In this chapter we use the inner product $(\mathbf{x}, \mathbf{y})$ defined by

$$(\mathbf{x}, \mathbf{y}) = \sum_j x_j \bar{y}_j = \mathbf{x}^T \bar{\mathbf{y}} = \overline{(\mathbf{y}, \mathbf{y})}, \tag{1.5.1}$$

where $\bar{\mathbf{y}}$ is the conjugate complex of $\mathbf{y}$. Further $||x|| = \sqrt{(x, x)}$.

Definition 1.5.1 *Let A be a $n \times n$ matrix. Let λ be a complex number and $\mathbf{v}$ a complex vector such that*

$$A\,\mathbf{v} \;=\; \lambda\,\mathbf{v}, \qquad \mathbf{v} \neq 0, \tag{1.5.2}$$

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

Theorem 1.5.1 *All eigenvalues of a real symmetrical matrix are real, and eigenvectors corresponding to different eigenvalues are orthogonal.*

Proof Multiplication of Equation (1.5.2) by the vector $\bar{\mathbf{v}}^T$:

$$\bar{\mathbf{v}}^T A\mathbf{v} \;=\; \lambda \bar{\mathbf{v}}^T \mathbf{v}. \tag{1.5.3}$$

$\bar{\mathbf{v}}^T A\mathbf{v}$ is real, since

$$\overline{(\bar{\mathbf{v}}^T A\mathbf{v})}^T \;=\; \bar{\mathbf{v}}^T \bar{A}^T \mathbf{v} \;=\; \bar{\mathbf{v}}^T A\mathbf{v} \qquad \text{A symmetrical real.} \tag{1.5.4}$$

In the same way $\bar{\mathbf{v}}^T \mathbf{v}$ is real, hence λ is real.
Further $(\mathbf{v}_i, A\mathbf{v}_j) = (A\mathbf{v}_i, \mathbf{v}_i)$, where $\mathbf{v}_i$ and $\mathbf{v}_j$ are eigenvectors associated with eigenvalues λ_i, λ_j $(\lambda_i \neq \lambda_j)$. This implies $\lambda_j(\mathbf{v}_i, \mathbf{v}_j) = \lambda_i(\mathbf{v}, \mathbf{v}_j)$. Since $\lambda_i \neq \lambda_j$ it immediately follows that $(\mathbf{v}_i, \mathbf{v}_j) = 0$. □

Definition 1.5.2 *A matrix A is called skewed symmetric if $A^T = -A$.*

Theorem 1.5.2 *All eigenvalues of a real skewed symmetrical matrix are purely imaginary.*

Exercise 1.5.1 *Prove Theorem (1.5.2) analogously to the proof of Theorem (1.5.1).* □

Definition 1.5.3 *The Rayleigh quotient, $R(A, \mathbf{x})$, of a symmetrical matrix A is given by:*

$$R(A, \mathbf{x}) = \frac{(\mathbf{x}, \mathbf{A}\mathbf{x})}{(\mathbf{x}, \mathbf{x})}. \tag{1.5.5}$$

Theorem 1.5.3 *If $\mathbf{x}$ is an eigenvector of A, then R is equal to the corresponding eigenvalue.*

Proof
Let λ_i be an eigenvalue of A, then

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{1.5.6}$$

$$R(A, \mathbf{v}_i) \;=\; \frac{(\mathbf{v}, A\mathbf{v})}{(\mathbf{v}, \mathbf{v})} \;=\; \lambda_i. \tag{1.5.7}$$

□

Theorem 1.5.4 *For the Rayleigh quotient, $R(A, \mathbf{x})$, of a symmetrical matrix A we have*

$$\lambda_1 \leq R(A, \mathbf{x}) \leq \lambda_n \quad \forall \mathbf{x}, \tag{1.5.8}$$

with λ_1 the smallest and λ_n the largest eigenvalue of A.

Exercise 1.5.2 *Prove Theorem (1.5.4).*
Hint: use the fact that the eigenvectors of a symmetric matrix form an orthonormal basis of the space $\mathbb{R}^n$ *and expand the vector* $\mathbf{x}$ *as a linear combination of these eigenvectors.* □

Definition 1.5.4 *A matrix* A *is called positive if* $(\mathbf{x}, A\mathbf{x}) \geq 0, \ \forall \mathbf{x} \in \mathbb{R}^n$.

Definition 1.5.5 *A matrix* A *is called positive definite if* $\exists\ \alpha > 0$ *such that* $(\mathbf{x}, A\mathbf{x}) \geq \alpha ||\mathbf{x}||^2, \ \forall \mathbf{x} \in \mathbb{R}^n$.

Theorem 1.5.5

- *If* A *is positive, then its eigenvalues are non-negative.*
- *If* A *is positive definite, then its eigenvalues are positive.*

Theorem 1.5.6 *Let* A *be symmetric.*

- *If the eigenvalues of* A *are non-negative, then* A *is positive.*
- *If the eigenvalues of* A *are positive, then* A *is positive definite.*

Proof of Theorem 1.5.5

- Let $(\lambda, \mathbf{v})$ be an eigenpair of A, then by definition $(\mathbf{v}, A\mathbf{v}) = \lambda(\mathbf{v}, \mathbf{v}) \geq 0$. Since $(\mathbf{v}, \mathbf{v}) > 0$, we have $\lambda \geq 0$.
- Let $(\lambda, \mathbf{v})$ be an eigenpair of A, and let A be positive definite, then $\exists\ \alpha > 0$ such that $(\mathbf{v}, A\mathbf{v}) = \lambda(\mathbf{v}, \mathbf{v}) \geq \alpha(\mathbf{v}, \mathbf{v}) > 0$.
  Since $(\mathbf{v}, \mathbf{v}) > 0$, this immediately implies $\lambda > 0$.

□

Proof of Theorem 1.5.6

- We expand any vector $\mathbf{x}$ as a linear combination of the eigenvectors of A. Symmetry of A enables this procedure. Then

$$\mathbf{x} = \sum_j c_j \mathbf{v}_j. \tag{1.5.9}$$

  This implies with orthogonality of the eigenvectors

$$||\mathbf{x}||^2 = (\mathbf{x}, \mathbf{x}) = \sum_j c_j^2. \tag{1.5.10}$$

  Hence, we get similarly

$$(\mathbf{x}, A\mathbf{x}) = (\sum_j c_j \mathbf{v}_j, \sum_k \lambda_k c_k \mathbf{v}_k) = \sum_j c_j^2 \lambda_j. \tag{1.5.11}$$

  If $\lambda_j \geq 0, \ \forall j$, then this implies

$$(\mathbf{x}, A\mathbf{x}) \geq 0, \tag{1.5.12}$$

  which proves the first assertion.

- The second assertion follows from

$$(\mathbf{x}, A\mathbf{x}) = \sum_j c_j^2 \lambda_j \geq \sum_j c_j^2 \lambda_{min}, \tag{1.5.13}$$

  where $0 < \lambda_{min} = \min_j \lambda_j$.
  Since $(\mathbf{x}, \mathbf{x}) = \sum_j c_j^2$, we get

$$(\mathbf{x}, A\mathbf{x}) \geq \lambda_{min}(\mathbf{x}, \mathbf{x}), \quad \lambda_{min} > 0. \tag{1.5.14}$$

  Hence A is positive definite.

□

As a consequence the Rayleigh quotient of a positive matrix is non-negative, whereas the Rayleigh quotient of a positive definite matrix is positive.

The following theorem can be of great help in estimating bounds for eigenvalues of matrices. This is for example useful in stability analysis.

Theorem 1.5.7 (Gershgorin)
For all eigenvalues λ of the matrix A holds:

$$|\lambda - a_{kk}| \leq \sum_{i=1_{i\neq k}}^{N} |a_{ki}|. \tag{1.5.15}$$

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 are located within a circle in the complex plane and that is the reason why the theorem is also often referred to as Gershgorin's *circle* theorem. But for symmetric A, the eigenvalues of A are real-valued.

Proof
Let λ be an eigenvalue of the eigenvalue problem with corresponding eigenvector, $\mathbf{v}$, then, $A\mathbf{v} = \lambda\mathbf{v}$, and for each row, p, this gives

$$\sum_i a_{pi} v_i = \lambda v_p, \quad p = 1, \ldots, N. \tag{1.5.16}$$

Let v_k be the component of $\mathbf{v}$ with the largest modulus. For this index k we have

$$\lambda - a_{kk} = \sum_{i \ i\neq k} a_{ki} \frac{v_i}{v_k}, \tag{1.5.17}$$

and because $|v_i/v_k| \leq 1$, we get

$$|\lambda - a_{kk}| \leq \sum_{i \ i\neq k} |a_{ki}|. \tag{1.5.18}$$

This proves the theorem. □

Definition 1.5.6 *A matrix,* **A**, *is called a band-matrix if all elements, a_{ij}, 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 $b_1 + b_2 + 1$.*

1.6 Some theorems used in the mathematical theory

In some of the proofs used in this book we shall use the following theorems.
Let $L^2(\Omega) := \{u : \Omega \to \mathbb{R} : \int\limits_{\Omega} |u^2| d\Omega < \infty \}$.

Theorem 1.6.1 *Inequality of Poincaré (Friedrichs)*
Let $\Omega \subset \mathbb{R}^m$, $u \in H^1(\Omega) = \{u \in L^2(\Omega) | \frac{\partial u}{\partial x_1}, ..., \frac{\partial u}{\partial x_m} \in L^2(\Omega)\}$ and $u|_\Gamma = 0$, then $\exists \ K > 0$ such that

$$\int\limits_{\Omega} \sum_{i=1}^{m} (\frac{\partial u}{\partial x_i})^2 \, d\Omega \geq K \int\limits_{\Omega} u^2 \, d\Omega. \tag{1.6.1}$$

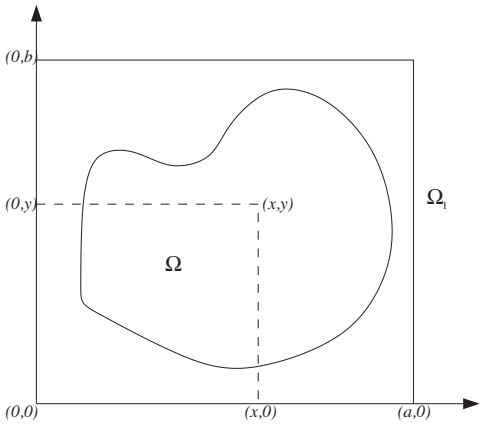


Figure 1.5: 2-dimensional region.

Proof We shall prove the theorem for $m = 2$.

By shifting coordinates we may assume that $(x, y) \in \Omega$ implies $x > 0$ and $y > 0$. The region Ω is enclosed by a rectangle Ω_1, given by $(a, 0) \times (0, b)$ as in Figure (1.5). Since $u(x, y) \in C^1(\Omega)$ and $u(x, y) = 0$ on Γ, we may extend $u(x, y)$ continuously to the whole domain Ω_1 by defining

$$u(x, y) = 0, \quad (x, y) \in \Omega_1 \backslash \Omega. \tag{1.6.2}$$

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

$$u(x_1, y_1) - u(0, y_1) = \int_0^{x_1} \frac{\partial u(x, y_1)}{\partial x}\, dx, \tag{1.6.3}$$

$$u(0, y_1) = 0 \quad \text{follows from Figure (1.5).} \tag{1.6.4}$$

According to Cauchy-Schwartz we have:

$$\{\int_\Omega uv\, d\Omega\}^2 < \int_\Omega u^2\, d\Omega \int_\Omega v^2\, d\Omega. \tag{1.6.5}$$

Hence

$$u^2(x_1, y_1) = \{\int_0^{x_1} \frac{\partial u(x, y_1)}{\partial x}\, dx\}^2 \leq x_1 \int_0^{x_1} \{\frac{\partial u(x, y_1)}{\partial x}\}^2\, dx \tag{1.6.6}$$

$$\leq a \int_0^{a} (\frac{\partial u(x, y_1)}{\partial x})^2\, dx. \tag{1.6.7}$$

Integration of Equation (1.6.6) over Ω_1 gives

$$\int_{\Omega_1} u^2(x, y)\, d\Omega \leq a^2 \int_{\Omega_1} (\frac{\partial u}{\partial x})^2\, d\Omega \leq a^2 \int_{\Omega_1} (\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2\, d\Omega \tag{1.6.8}$$

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

Exercise 1.6.1 *Prove Theorem (1.6.1) with $K = 1/b^2$.* □

From the proof of Theorem (1.6.1) and Exercise (1.6.1) if follows that K is overestimated by $K = \max(1/a^2, 1/b^2)$.
This theorem is used to prove that the Laplace operator is positive definite, when using Dirichlet or Robin boundary conditions. We will specify these boundary conditions in Chapter 2.

Definition 1.6.1 *A Banach space is a complete vector space defined over the real or complex numbers provided with a norm.*

Definition 1.6.2 *A Hilbert space is a Banach space provided with an inner product which defines the norm of the space.*

Definition 1.6.3 *A bilinear form $a(u, v)$ in V has the following properties*

- $a(u, v + w) = a(u, v) + a(u, w), \forall u, v, w, \in V,$
- $a(\lambda u, v) = \lambda a(u, v), \forall \lambda \in \mathbb{R}, \forall u, v \in V.$

Definition 1.6.4 *Let $a(.,.)$ be a bilinear form in V, then*

- *$a(.,.)$ is bounded if $\exists\, C > 0$ such that $|a(u, v)| \leq C||u||_V||v||_V, \forall u, v \in V.$*
- *$a(.,.)$ is coercive if $\exists\, C > 0$ such that $a(u, u) \geq C||u||_V^2, \forall u, v \in V.$*

The next theorem is used in existence and uniqueness proofs.

Theorem 1.6.2 *Lax-Milgram*
Let V be a Hilbert space and let $a(.,.)$ be a coercive and bounded, bilinear form on V. Further let $f \in V'$, where V' denotes the set (space) of linear functionals on V, then there is a unique solution $u \in V$, such that

$$a(u, v) = f(v), \forall v \in V. \tag{1.6.9}$$

This solution satisfies

$$||u|| \leq \frac{1}{c}||f||_{V'}. \tag{1.6.10}$$

For a proof of this theorem see for example [22].

1.7 Summary of Chapter 1

In this chapter we have seen the importance of conservation 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, diffusion and ground water flow. We concluded the chapter with an example of minimization as an instrument to derive a physical model. Besides that some standard mathematical theorems have been reviewed.

Chapter 2

A crash course in PDE's

Objectives

In the previous chapter we looked at PDE's 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 PDE's are a class of their own: the *transport* equations. We shall consider them in Chapter 11. In this chapter we shall take a look at second order PDE's 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 PDE's. 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 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 symmetric, 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. The columns of Q are the normalized (with length one) eigenvectors of A. Note that $Q^T = Q^{-1}$

due to symmetry 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 $a_{12}^2 - a_{11}a_{22} < 0$, $a_{12}^2 - a_{11}a_{22} = 0$ and $a_{12}^2 - a_{11}a_{22} > 0$, respectively correspond to $\alpha_{11}\alpha_{22} > 0$, $\alpha_{11}\alpha_{22} = 0$ and $\alpha_{11}\alpha_{22} < 0$ (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

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

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 hyperbolic, parabolic or elliptic 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. Parabolic and hyperbolic equations are time-dependent: they describe the evolution in time 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 type of equation may 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:

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

$a_{ij} = a_{ji}$ 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: 3D Laplacian
$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0. \tag{2.1.10}$$
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: 3D Wave equation
$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}. \tag{2.1.11}$$
3. Exactly one α_{ii} vanishes, say α_{11}. Then ξ_1 is a time variable and the equation is called *parabolic*. Example: 3D Heat equation
$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}. \tag{2.1.12}$$

Exercise 2.1.3 *If A is a symmetric $n \times n$ matrix there exists a real unitary matrix C such that $C^T AC = \Lambda$. Λ is a diagonal matrix containing the eigenvalues of A on the diagonal. Show that the substitution $\xi = C^T\mathbf{x}$ 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 in time-dependent problems we need initial conditions too. We will just consider here second order PDE's because the considerations for first order PDE's are very different and will be considered in Chapter 11.

2.2.1 Boundary conditions

Consider the bounded region in $\mathbb{R}^2$, Ω with boundary Γ in Figure 2.1. 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 $k > 0$ $\forall \mathbf{x} \in \overline{\Omega}$, the following boundary conditions guarantee a unique solution:

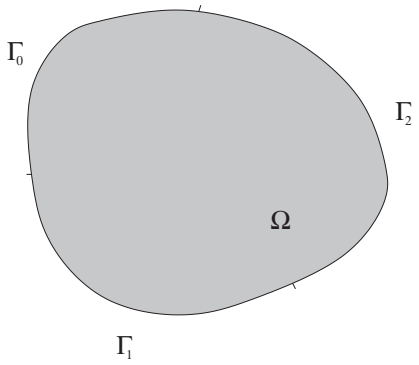


Figure 2.1: The bounded region Ω.

1.
$$u = g_0(\mathbf{x}), \quad \mathbf{x} \in \Gamma_0, \tag{2.2.2}$$
the *Dirichlet boundary condition*.

2.
$$k\frac{\partial u}{\partial n} = g_1(\mathbf{x}), \quad \mathbf{x} \in \Gamma_1, \tag{2.2.3}$$
the *Neumann boundary condition*.

3.
$$k\frac{\partial u}{\partial n} + \sigma u = g_2(\mathbf{x}), \quad \sigma \geq 0, \, \mathbf{x} \in \Gamma_2, \tag{2.2.4}$$
the *Robin*, *radiation*, *kinetic* or *mixed* boundary condition.

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 on 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 is needed if first and zeroth order terms are added to Equation (2.2.1).

2. On each part of the boundary *precisely one* boundary condition applies. (For second order PDE's)

3. Boundary conditions involving the flux vector (Neumann, Robin) are also called *natural boundary conditions.* (For second order PDE's) This term will be explained in Chapter 5.

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 the 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 in addition $\frac{\partial u}{\partial t}$ has to be given on all of Ω at $t = t_0$.

Exercise 2.2.2 *Consider the transversal vibrations of 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 ODE's a practical way to go about this is try and find one. In PDE's 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 Laplacian operator

The Laplacian 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.
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 interior of such a domain. We note that a continuous function $u(\mathbf{x})$ has an isolated maximum in some point $\mathbf{x}_0 \in \Omega$ if there exists a $\delta > 0$ such that $u(\mathbf{x}_0) > u(\mathbf{x})$ for $||\mathbf{x} - \mathbf{x}_0|| < \delta$.

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

$$H(\mathbf{x_0}) = \begin{pmatrix} \frac{\partial^2 u}{\partial x^2}(\mathbf{x_0}) & \frac{\partial^2 u}{\partial x \partial y}(\mathbf{x_0}) \\ \frac{\partial^2 u}{\partial y \partial x}(\mathbf{x_0}) & \frac{\partial^2 u}{\partial y^2}(\mathbf{x_0}) \end{pmatrix}. \tag{2.3.2}$$

Theorem 2.3.1 *If the function* $u(\mathbf{x})$ *in* $C^2(\Omega)$*, i.e. the second order derivatives are continuous, the Hessian matrix in an isolated maximum must be negative definite.*

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

$$u(\mathbf{x}) = u(\mathbf{x}_0) + \nabla u(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0, H(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)) + O(||\mathbf{x} - \mathbf{x}_0||^3). \tag{2.3.3}$$

Since u has a maximum for $\mathbf{x}_0$, the smoothness of u implies $\nabla u(x_0) = \mathbf{0}$. When $\mathbf{x}$ approaches $\mathbf{x}_0$ the third order error term becomes arbitrarily small. This implies, with $u(\mathbf{x}) - u(\mathbf{x}_0) < 0$, that there exists a $\delta > 0$ such that
$(\mathbf{x} - \mathbf{x}_0, H(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)) < 0$ for $||\mathbf{x} - \mathbf{x}_0|| < \delta$ and hence $H(\mathbf{x}_0)$ is negative definite. □

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

Exercise 2.3.2 *Show that if* H *is positive definite both diagonal elements must be positive. Hint: Make special choices for* $\mathbf{u}$ *in* $(\mathbf{u}, H\mathbf{u})$. □

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.

From Exercise 2.3.2 it is clear that $u_{xx}(\mathbf{x_0})$ and $u_{yy}(\mathbf{x_0})$ are negative if $u(\mathbf{x})$ has an isolated maximum in $\mathbf{x}_0$. This suggests the following theorem

Theorem 2.3.2 *(Strong maximum principle) Let* Ω *be an open bounded domain with boundary* Γ *and closure* $\overline{\Omega}$*, that is* $\overline{\Omega} = \Omega \cup \Gamma$*. Suppose* $u \in C^2(\Omega) \cap C(\overline{\Omega})$ *is harmonic within* Ω*.*
Then

(i) u takes its maximum at the boundary Γ*, hence* $\max_{\mathbf{x} \in \overline{\Omega}} u = \max_{\mathbf{x} \in \Gamma} u$*.*

(ii) Furthermore, if Ω *is connected and if there is an internal point* $\mathbf{x}_0$*, where u reaches its maximum* $(u(\mathbf{x}_0) = \max_{\mathbf{x} \in \overline{\Omega}} u)$*, then u is constant on* $\overline{\Omega}$*,*
that is $u(\mathbf{x}) = u(\mathbf{x}_0)$ *on* $\overline{\Omega}$*.*

This theorem is formulated and proved in Evans [16] among others. To prove the maximum principle, we shall use the arguments given in Protter and Weinberger [30]. Theorem 2.3.2 says that the maximum of a harmonic function is always found on the boundary Γ unless the function is constant. By replacing u by $-u$, we recover similar assertions as in Theorem 2.3.2 with *min* replacing *max*. Before we prove the theorem we give several consequences of the assertion.

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 2.3.3 *Prove Theorem 2.3.3.* □

Theorem 2.3.4 *(uniqueness) Let Ω be a bounded region in $\mathbb{R}^2$ with boundary Γ, and suppose that $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfies*

$$-\Delta u = f(x,y), \qquad (x,y) \in \Omega, \tag{2.3.4}$$

$$u = g(x,y), \qquad (x,y) \in \Gamma. \tag{2.3.5}$$

Then there exists at most one solution u.

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

Next we prove Theorem 2.3.2.

Proof of Theorem 2.3.2.
We prove the theorem for $\Omega \in \mathbb{R}^2$. 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

$$v(x,y) = u(x,y) + \epsilon(x^2 + y^2), \text{ with } \epsilon > 0 \text{ arbitrarily.} \tag{2.3.6}$$

Since u is harmonic, this implies

$$\Delta v = 4\epsilon > 0, \quad \text{in } \Omega. \tag{2.3.7}$$

Suppose that v has a maximum in the open domain Ω, then $\Delta v \leq 0$. This contradicts with the strict inequality (2.3.7), and hence v cannot have a maximum in Ω. Since Ω is a bounded domain in $\mathbb{R}^2$, there exists a radius R such that

$$R = \max_{\mathbf{x} \in \mathbf{\Gamma}} ||\mathbf{x}|| = \max_{\mathbf{x} \in \mathbf{\Gamma}} \sqrt{\mathbf{x^2} + \mathbf{y^2}}. \tag{2.3.8}$$

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

$$u(\mathbf{x}) \leq v(\mathbf{x}) \leq u_m + \epsilon R^2 \text{ in } \overline{\Omega} (= \Omega \cup \Gamma). \tag{2.3.9}$$

Since $\epsilon > 0$ can be taken arbitrarily small, we get $u \leq u_m$ in $\overline{\Omega}$. Since u_m is attained on Γ, it follows that a maximum can only be assumed on boundary Γ unless u is constant on $\overline{\Omega}$. □
Uniqueness for the solution to the Poisson equation with Robin conditions can also be proved easily.

Theorem 2.3.5 *Let Ω be a bounded domain in $\mathbb{R}^2$ with boundary Γ, and let $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ satisfy*

$$-\Delta u = f(x,y)), \qquad (x,y) \in \Omega, \tag{2.3.10}$$

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

with $\sigma > 0$. Then there exists at most one solution u.

Exercise 2.3.5 *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*

$$-\Delta u \geq 0, \qquad \text{in } \Omega, \tag{2.3.12}$$

$$u = 0, \qquad \text{on } \Gamma, \tag{2.3.13}$$

where Ω is a bounded domain with boundary Γ, then $u \geq 0$ in Ω

Exercise 2.3.6 *Prove Theorem 2.3.6.*
Reason by contradiction and use the Completeness Principle which is: if $u \in C(\overline{\Omega})$ where $\overline{\Omega}$ is a closed bounded set, then u must have a global maximum and minimum on $\overline{\Omega}$. □

Exercise 2.3.7 *Show that the elliptic operator $au_{xx} + 2bu_{xy} + cu_{yy}$, a,b,c constant, $ac - b^2 > 0$ satisfies the same maximum principle as the Laplacian operator.*
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 [12]. As we shall see in Chapter 7, there is an alternative way to obtain a *generalized* solution to these problems. The existence proof of such a solution is somewhat easier.

2.4 Examples

In this section we give a few examples of PDE's 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

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

For uniqueness c must be a monotone function of u and for stability it must be non-decreasing. In ordinary heat transfer, ground water flow 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 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 $u(\mathbf{x}, t)$, it is necessary that u is prescribed at $t = t_0$: $u(\mathbf{x}, t_0) = u_0(\mathbf{x}), \forall \mathbf{x} \in \Omega$.

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:

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

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, $\mathbf{u}$. The equation is

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

Comparison of Equation (2.4.3) with (2.4.1) shows that a *convection* term $\mathbf{u} \cdot \text{grad}\, c$ 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, $\mathbf{u}$, 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.

2.4.3 Navier-Stokes equations

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

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

$$\rho(\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y}) = \text{div}\, \mathbf{s}_y + \rho b_y. \tag{2.4.4b}$$

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, $\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}$, *moving along with the flow*. Thus, a particle $P \in V$ with coordinates $\mathbf{x}$ at time $t$has at time $t + \Delta t$, with $\Delta t \to 0$, coordinates $\mathbf{x} + \mathbf{u}\Delta t$. Therefore the change in velocity of a *moving* particle is described by

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

We recall Taylors theorem in three variables:

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

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 quantity

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

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):

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

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

$$\operatorname{div}\mathbf{u} = 0, \tag{2.4.12}$$

the *incompressibility condition*. 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 $(\mathbf{t}, \Sigma \cdot \mathbf{n})$ for given unit tangent vector, $\mathbf{t}$, and unit normal vector, $\mathbf{n}$. 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 see [39] and [15].

2.4.4 Plane stress

Consider the flat plate in Figure 2.2.

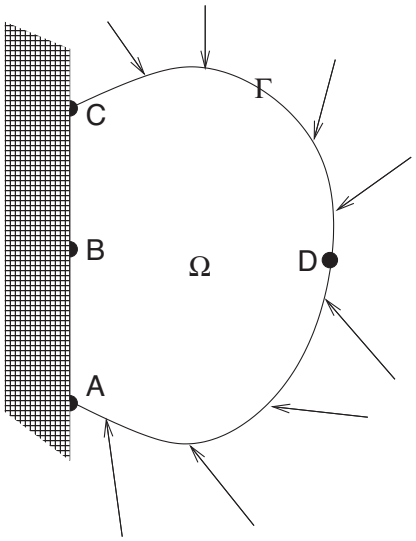


Figure 2.2: Fixed plate with forces applied along the boundary.

The plate is fixed along side ABC but forces are applied along the free boundary ADB 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* $\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}$. The differential equations for the stresses (compare also (2.4.4)) are 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 $\mathbf{b}$ 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 PDE's 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 constant, are material constants. Furthermore, for infinitesimal strains, there is a relation between strain and displacement:

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

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

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

This leads to the following set of PDE's in the displacements $\mathbf{u}$:

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

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

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:* $\mathbf{f}$. *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*

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

2. *Freely supported boundary*

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

3. *Free boundary*

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

$\frac{\partial}{\partial n}$ and $\frac{\partial}{\partial t}$ stand for the *normal* and *tangential* derivative respectively. Further ν is Poisson's constant, 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 PDE's into *hyperbolic*, *parabolic* and *elliptic* equations. We formulated appropriate initial and boundary conditions to guarantee a unique solution. We obtained a maximum principle for harmonic 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)

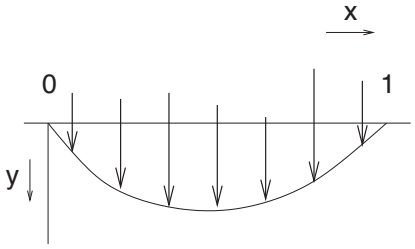


Figure 3.1: Loaded cable.

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 here also *one* boundary condition is necessary for the whole boundary, which just consists of two points.

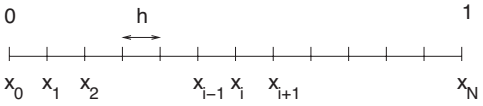


Figure 3.2: Subdivision of the interval $(0,1)$.

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 $f_i = f(x_i)$.
In the node point 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 Taylors formula for sufficiently smooth y:

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

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

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

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

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

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

The $O(h^2)$ 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*

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

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

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

□

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*

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

□

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 $O(h^2)$ 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

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

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

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

Or in matrix-vector notation:

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

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

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

Exercise 3.1.3 *Show that in case of non-homogeneous boundary conditions, $y(0) = a$, and $y(1) = b$, the matrix A is given by (3.1.13) and that the first and last element of the right-hand side $\mathbf{f}$ 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[24]) 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.

Exercise 3.1.4 *Show that matrix A is positive.*
Hint use Theorem (1.5.7). □

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

Exercise 3.1.5 *Show that*

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

Derive from this result that A is positive definite. □

Another method is to estimate the eigenvalues of the matrix. This can be done by the *von Neumann method*. This approach has the advantage that the smallest eigenvalue can be estimated more accurately than with the bounds that follow from Gershgorin's theorem. Later on it will be used to get a global error estimate. The von Neumann method is based on the fact that the solution of Equation (3.1.3) can be written as

$$y(x) = \sum_{\alpha=1}^{\infty} \rho_\alpha e^{-\pi\alpha x i}, \tag{3.1.15}$$

where i is the imaginary unit ($i^2 = -1$).
In the discrete case we expand the k-th component of u in a similar way ($x_k = kh$)

$$u_k = \sum_{\alpha=1}^{N-1} \rho_\alpha e^{-\pi\alpha k h i}. \tag{3.1.16}$$

The eigenvalue problem $A\mathbf{v} = \lambda\mathbf{v}$ results in

$$\frac{1}{h^2}(-u_{k-1} + 2u_k - u_{k+1}) = \lambda u_k. \tag{3.1.17}$$

Substitution of (3.1.16) in (3.1.17) gives

$$\frac{1}{h^2}\sum_{\alpha=1}^{N-1} \rho_\alpha(-e^{-\pi\alpha(k-1)hi} + 2e^{-\pi\alpha khi} - e^{-\pi\alpha(k+1)hi}) = \lambda \sum_{\alpha=1}^{N-1} \rho_\alpha e^{-\pi\alpha khi}. \tag{3.1.18}$$

This must be true for arbitrary ρ_α, hence each factor following ρ_α in the sum should be zero. Subdivision by $e^{-\pi\alpha khi}$ results in

$$\frac{1}{h^2}(2 - e^{\pi\alpha hi} + e^{-\pi\alpha hi}) = \lambda, \tag{3.1.19}$$

and since

$$cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2}, \tag{3.1.20}$$

we get $N-1$ eigenvalues λ_α:

$$\lambda_\alpha = \frac{2(1 - cos(\pi\alpha h))}{h^2}. \tag{3.1.21}$$

Exercise 3.1.6 *Use the Taylor expansion of the cosine 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 that A is positive definite.

Remarks

- The eigenvalue problem corresponding to the Laplace Equation, which is given by

$$-\frac{d^2\varphi}{dx^2} = \mu^2\varphi, \qquad \varphi(0) = \varphi(1) = 0, \tag{3.1.22}$$

  is a special case of the set of Sturm-Liouville problems. The eigenvalues of Equation (3.1.22) form an infinite set given by $\mu^2 = k^2\pi^2$ with k any positive integer. Hence the smallest eigenvalue is exactly π^2.

- The von Neumann 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 an error of $O(h^2)$, provided that y is sufficiently smooth. Suppose that the 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 O. Now let $\Delta y_k = y_k - u_k$, where y_k is the exact solution and u_k our numerical approximation. Then

$$A\mathbf{y} = \mathbf{f} + h^2\mathbf{p}, \tag{3.1.23}$$

and

$$A\mathbf{u} = \mathbf{f}. \tag{3.1.24}$$

We subtract (3.1.24) from (3.1.23) to obtain a set of equations for the error

$$A\Delta\mathbf{y} = h^2\mathbf{p}. \tag{3.1.25}$$

We shall show the global error Δy is of order $O(h^2)$. This is formulated in the following theorem:

Theorem 3.1.1 *The discretization of the Laplace Equation (3.1.1) with boundary conditions (3.1.2) by Equation (3.1.6) gives a global error of $O(h^2)$ in L_2-norm.* □

Proof
From Equation (3.1.25) we get

$$||\Delta\mathbf{y}||_2 \leq ||A^{-1}||_2 h^2 ||\mathbf{p}||_2, \tag{3.1.26}$$

and since the L_2-norm of the inverse of a positive definite matrix is equal to the inverse of the smallest eigenvalue, λ_1, we get

$$||\Delta\mathbf{y}||_2 \leq \frac{h^2}{\lambda_1}||\mathbf{p}||_2 \approx \frac{h^2}{\pi^2}||\mathbf{p}||_2. \tag{3.1.27}$$

□

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 3.1.2 *Let the discretization give a truncation error of $e_k = h^2 p_k$ in the k-th equation, then*

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

□

The above theorem will be proved in Exercises 3.1.7 and 3.1.10.

Exercise 3.1.7 *Let $\mathbf{d}$ be a vector with components $d_k = 1, k = 1, 2, \ldots, N-1$. Show by substitution that the solution $\mathbf{e}$ of the set of equations $A\mathbf{e} = \mathbf{d}$ has components $e_k = \frac{1}{2}h^2(N-k)k, k = 1, 2, \ldots, N-1$. Show from this result, that $\|\mathbf{e}\|_\infty \leq 1/8$. (Hint: $Nh = 1$, and by definition $\|\mathbf{p}\|_\infty = \max_k |p_k|$.)* □

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:

Theorem 3.1.3 *(Discrete Maximum Principle) The vector inequality* $\mathbf{y} \geq \mathbf{x}$ *means that the inequality is valid for* every *component. Let* A *be the discretization matrix as in* (3.1.13)*, then* $A\mathbf{u} \geq 0$ *implies* $\mathbf{u} \geq 0$. □

Exercise 3.1.8 *Prove Theorem 3.1.3. Reason by contradiction and assume that* $\mathbf{u}$ *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 3.1.4 *(Existence and uniqueness)*

1. *Let* A *be given as in equation* (3.1.13)*, then* $A\mathbf{u} = 0$ *implies* $\mathbf{u} = 0$.
2. *From this, it follows that the set of equations* $A\mathbf{u} = \mathbf{f}$ *has a solution for every* $\mathbf{f}$ *and that this solution is unique.*

□

Exercise 3.1.9 *Prove Theorem 3.1.4. Use the result from Theorem 3.1.3.* □

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

$$-h^2\|p\|_\infty \mathbf{e} \leq \Delta\mathbf{y} \leq h^2\|p\|_\infty \mathbf{e}. \tag{3.1.28}$$

Show that therefore

$$\|\Delta\mathbf{y}\|_\infty \leq \frac{h^2}{8}\|\mathbf{p}\|_\infty. \tag{3.1.29}$$

Hint: use Theorem (3.1.3). □

This concludes the proof of Theorem 3.1.2.

3.2 Some simple extensions of the cable equation

The Laplace 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 $O(h^2)$. 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.

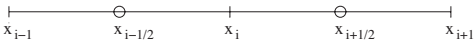


Figure 3.3: Position of discretization points.

One could use Taylor expansion to derive a 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 Equation (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) to get

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

Substitution of (3.2.4), (3.2.5) into (3.2.1) gives

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

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*

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

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

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

Hint: Use Equation (3.2.3). □

This discretization 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 O(h^2) accuracy, requires a thorough discretization.
Consider the Laplace 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) 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 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 $O(h^2)$, 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

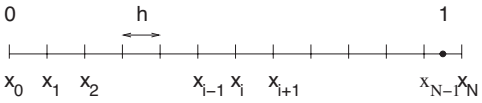


Figure 3.4: Subdivision with virtual point.

error, Equation (3.1.6) in $i = N$ becomes

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

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

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

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

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

Remark
A more simple way to get a symmetrical matrix would be to use the original matrix and to subdivide 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 $O(h^2)$ approximations are used, still the local truncation error of Equation (3.2.13) is $O(h)$, see Exercise (3.2.2)

Exercise 3.2.2 *Show that the Taylor series expansion around x_N of the left-hand side of Equation (3.2.13) can be written as*

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

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

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

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 O(h). □

It is rather disappointing that the local truncation error is $O(h)$, despite the fact that we used $O(h^2)$ approximations in each step. Fortunately it is possible to prove that the global error is still $O(h^2)$. For that purpose we write the truncation error for

the complete system as $h^2\mathbf{p} + h\mathbf{q}$, where $\mathbf{p}$ is defined as in (3.1.23) and $\mathbf{q}$ is a vector that is completely zero except for the last component which is equal to q_N, so

$$\mathbf{q} = (0, 0, ..., 0, q_N)^T. \tag{3.2.16}$$

The global error Δy can be split into $\Delta y = \Delta y_1 + \Delta y_2$, with

$$A\Delta y_1 = h^2\mathbf{p}, \tag{3.2.17}$$

and

$$A\Delta y_2 = h\mathbf{q}. \tag{3.2.18}$$

From Theorems (3.1.1) and (3.1.2) it follows that $||\Delta y_1|| = O(h^2)$. The exact solution of (3.2.18) is $(\Delta y_2)_i = h^2 x_i$, hence the global error $||\Delta y||$ is also O(h^2).

Exercise 3.2.3 *Show that $\varphi(x) = hq_n x$ is the solution of*
$-\frac{d^2\varphi}{dx^2} = 0, \quad \varphi(0) = 0, \frac{d\varphi}{dx}(1) = hq_N$.
Deduce from this result that $(\Delta y_2)_i = h^2 q_n x_i$, and hence $||\Delta y_2|| = |q_n|h^2$. □

Periodical boundary conditions require a slightly different approach. Such boundary conditions are for example used in case the solution repeats itself endlessly. Consider for example the Poisson equation

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

where $u(x)$ and $f(x)$ are periodical 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 3.2.4 *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 (3.2.22). Hence

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

Exercise 3.2.5 *Why is it sufficient to apply (3.2.23) only for $x = 1$?* □

Exercise 3.2.6 *Show that the discretization of (3.2.19) using (3.2.23) gives the following system of equations*

$$h^{-2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 & -1 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -1 & 2 & -1 \\ -1 & 0 & \dots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix} \tag{3.2.24}$$

□

Note that for a fast solution of systems of equations of the shape (3.2.24) an adapted solution method is required. See Chapter (9) for the details.

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:

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

that 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:

$$c(x) = \frac{e^{\frac{vx}{\varepsilon}} - 1}{e^{\frac{v}{\varepsilon}} - 1}. \tag{3.3.2}$$

For $v/\varepsilon = 40$ the solution has been plotted in Figure 3.5.

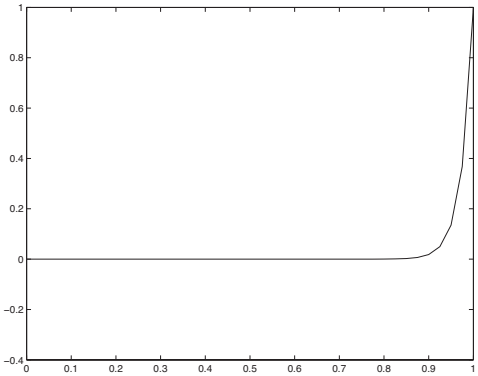


Figure 3.5: Analytic solution.

The quantity vL/ε that occurs regularly in convection diffusion problems is called the *Péclet number Pe*. The quantity L represents a characteristic length. It is a measure for by how much the convection dominates the diffusion. Note that there is a boundary layer at $x = 1$: 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 $O(h^2)$ consistent scheme. This gives us a set of equations

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

in which A is given by

$$A = h^{-2}\begin{pmatrix} 2 & -1+p_h & 0 & \dots & \dots & 0 \\ -1-p_h & 2 & -1+p_h & 0 & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -1-p_h & 2 & -1+p_h \\ 0 & \dots & \dots & 0 & -1-p_h & 2 \end{pmatrix} \tag{3.3.4}$$

and $\mathbf{f}$ by

$$\mathbf{f} = \frac{1}{h^2}\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1-p_h \end{pmatrix}, \tag{3.3.5}$$

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

Exercise 3.3.1 *Derive matrix* (3.3.4) *and vector* (3.3.5) □

In Figures 3.6 and 3.7 you see the numerical solution for $Pe = 40$ and $h = 0.1$ and $h = 0.025$ respectively.

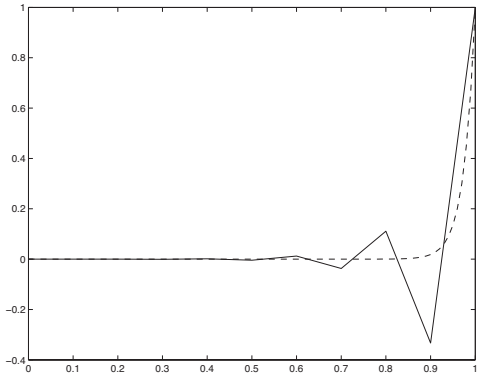


Figure 3.6: Solution (solid) and exact (dotted), coarse grid.

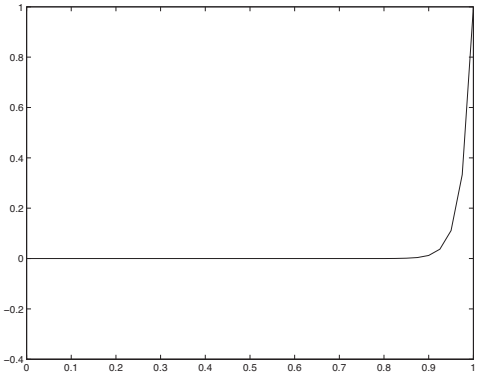


Figure 3.7: Solution, fine grid.

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.

3.3.2.1 Explanation

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

$$bu_{k-1} - (b+a)u_k + au_{k+1} = 0, \quad u_0 = 0, u_n = 1. \tag{3.3.6}$$

This system can be solved by substituting $u = r^k$. From Equation (3.3.6) it follows, that

$$b - (b+a)r + ar^2 = 0, \tag{3.3.7}$$

with solutions $r = 1$ and $r = b/a$. The general solution of (3.3.6) can now be written as

$$u_k = A + B\left(\frac{b}{a}\right)^k. \tag{3.3.8}$$

After application of the boundary conditions we find

$$u_k = \frac{\left(\frac{b}{a}\right)^k - 1}{\left(\frac{b}{a}\right)^n - 1}. \tag{3.3.9}$$

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 $h \leq \frac{2}{Pe}$. 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:*

$$(-1-2p_h)u_{k-1}+(2+2p_h)u_k-u_{k+1}=0. \tag{3.3.10}$$

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.

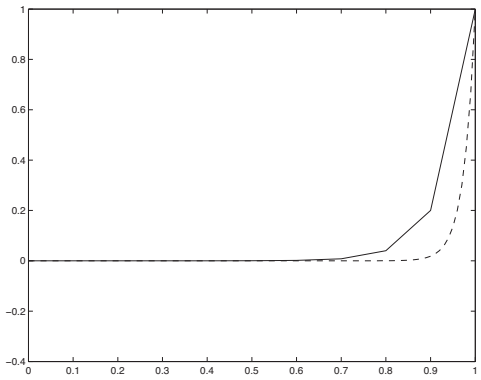


Figure 3.8: Upwind (solid) and exact (dotted) solution.

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*

$$\frac{c_k-c_{k-1}}{h}=c_k'-\frac{h}{2}c_k''+O(h^2). \tag{3.3.11}$$

Show that this approximation reduces the Péclet number to

$$\widehat{Pe}=\frac{Pe}{1+p_h}. \tag{3.3.12}$$

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

$$-(\varepsilon+\frac{vh}{2})\frac{d^2c}{dx^2}+v\frac{dc}{dx}=0. \tag{3.3.13}$$

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.

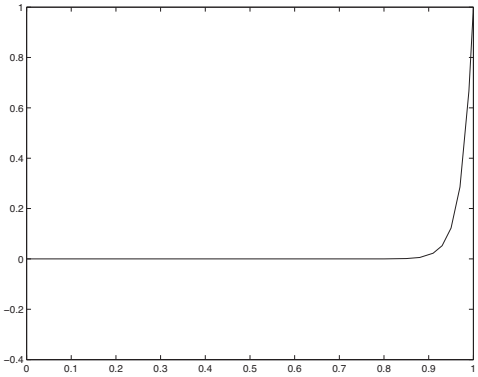


Figure 3.9: Non equidistant node points.

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*

$$-y'' + vy' = \pi^2 \sin \pi x + v\pi \cos \pi x, \quad y(0) = y(1) = 0. \tag{3.3.14}$$

Take $v = 40$ *and* $h = 0.1$. □

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

3.4 The Laplacian 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.
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*

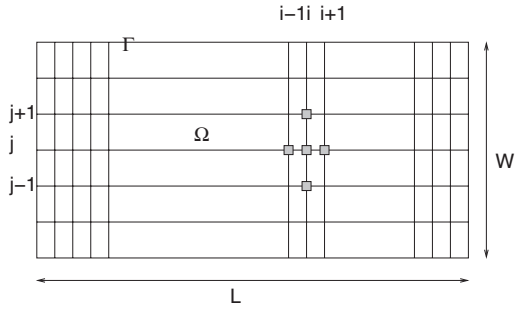


Figure 3.10: Rectangular grid with 5 point molecule.

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 node point (i, j), $i = 1, 2, \ldots, M-1$, $j = 1, 2, \ldots, N-1$ 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 $A\mathbf{u} = \mathbf{f}$. 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 simple 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. This can be done in a number of ways, that are most easily represented in a picture.

3.4.1.1 Horizontal numbering

The nodes are numbered sequentially (see Figure 3.11).

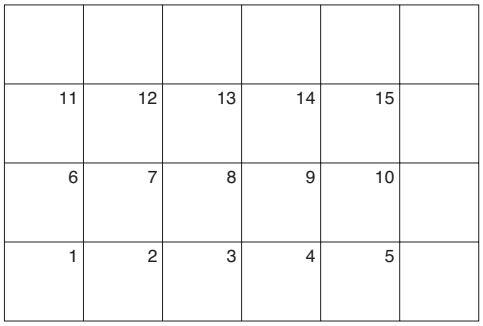


Figure 3.11: Horizontal numbering.

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 band width 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).

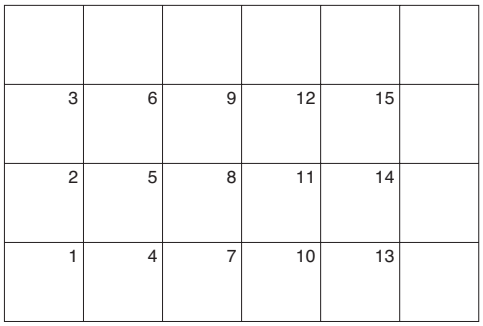


Figure 3.12: Vertical numbering.

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 band with of A?* □

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

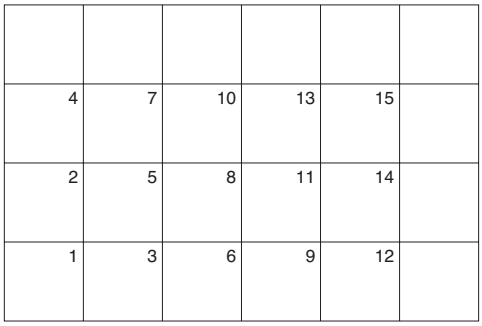


Figure 3.13: Oblique numbering.

3.4.1.3 Oblique numbering

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

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

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

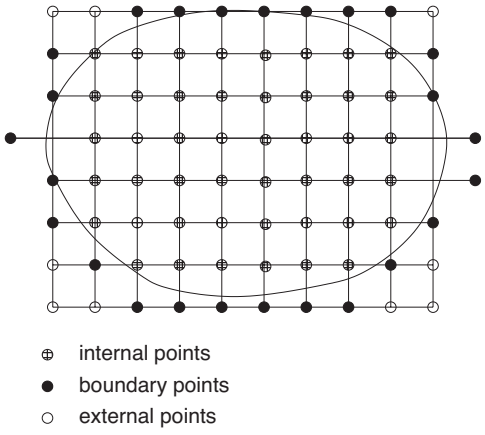


Figure 3.14: Grid on non rectangular region.

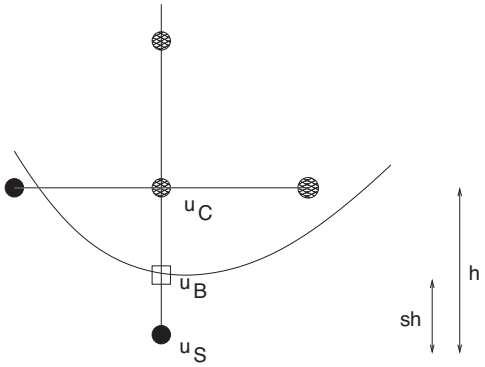


Figure 3.15: Boundary molecule.

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 and the formulae change correspondingly.

Here we have

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

Remark
The method treated here is quite old fashioned. It is better to use either a coordinate transformation (Section 3.7) or alternatively the Finite Element Method (Chapter 6).

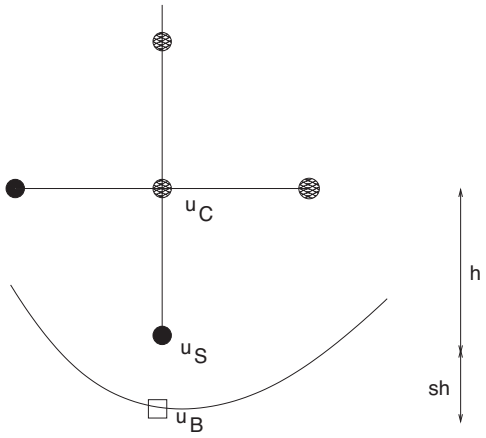


Figure 3.16: Boundary molecule, interior boundary point.

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 L_2 error estimate can be derived in the same way as in Theorem 3.1.1. Here we shall concentrate ourselves to point wise 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 A discrete maximum principle

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

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

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

We call matrices with property (3.6.1a) a Z-matrix and matrices with property (3.6.1b) *diagonally dominant*. If inequality (3.6.1b) holds strictly for at least one k the matrix is called *strictly* diagonally dominant. (There are complications if the system can be split into independent subsystems, but that won't bother us right now.)

We use the notation $\mathbf{y} \geq 0$ for $y_k \geq 0, k = 1, \ldots, N$. We formulate a very important theorem for the solution of systems of strictly diagonally dominant Z-matrices.

Theorem 3.6.1 *(Discrete Maximum Principle) Let A be a strictly diagonally dominant Z-matrix and let $A\mathbf{u} \geq 0$. Then*

- *(i)* $\mathbf{u} \geq 0$.

- *(ii)* $\mathbf{u} = \mathbf{0}$ *if and only if* $\mathbf{Au} = \mathbf{0}$.

Proof First we prove (i). Suppose $u_k < 0$ for some k then $\mathbf{u}$ has a negative minimum $-M$ for some index K. Hence $\mathbf{u} \geq -M$. Now consider the inequality with number K:

$$\sum_{j=1}^{N} a_{Kj} u_j \geq 0, \tag{3.6.2}$$

hence

$$a_{KK} u_K \geq -\sum_{\substack{j=1 \\ j \neq K}}^{N} a_{Kj} u_j, \tag{3.6.3}$$

and since $u_K = -M < 0$ by hypothesis, we have

$$a_{KK} M \leq \sum_{\substack{j=1 \\ j \neq K}}^{N} a_{Kj} u_j. \tag{3.6.4}$$

Because $a_{Kj} \leq 0, j \neq K$ the right hand side of the inequality is majorized by taking instead of u_j the negative minimum $-M$. We observe that $-a_{Kj}M = |a_{Kj}|M, j \neq K$ which lead us to:

$$a_{KK} M \leq \sum_{\substack{j=1 \\ j \neq K}}^{N} |a_{Kj}| M. \tag{3.6.5}$$

Since $M > 0$ we can divide both sides by M and arrive at a contradiction unless

$$a_{KK} = \sum_{\substack{j=1 \\ j \neq K}}^{N} |a_{Kj}|. \tag{3.6.6}$$

So (3.6.5) holds only for the equal case, and since $-M$ is the minimum (3.6.4) implies $u_j = -M$ for $a_{Kj} < 0$. This means that u_j is constant for all elements j in the molecule. We can repeat this argument for all molecules that have at least one element j in common with a previously considered molecule. Finally we arrive at a molecule where (3.6.1b) holds strictly. For that molecule we have a contradiction unless $M = 0$. This proves part (i) of the theorem.
The if part of (ii) is trivial. The *only if* part is proven in Exercise (3.6.1). □

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

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 3.6.2 it is shown that also the numerical solution is not unique.

Exercise 3.6.2 *Use the proof of Theorem 3.6.1 and Equation* (3.6.5) *to show that if equality holds in Equation* (3.6.1b) *for all k the system* $A\mathbf{u} = 0$ *(A being a Z-matrix) has a nontrivial solution. Determine that solution.* □

Exercise 3.6.3 *Use Theorem 3.6.1 to prove that if A is a strictly diagonally dominant Z-matrix and* $A\mathbf{u} = \mathbf{f}$ *and* $A\mathbf{w} = |\mathbf{f}|$*, with* $\mathbf{f} \neq 0$ *then* $|\mathbf{u}| \leq \mathbf{w}$.
Hint: also consider $A(-\mathbf{u})$. □

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 strong 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 real 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 a strictly diagonally dominant Z-matrix, and application of linear interpolation does not alter that. The inequality (3.6.1b) only holds for those molecules 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 have to be 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 3.6.4 *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:

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

in which the vector $\mathbf{p}$ is uniformly bounded as $h \to 0$. Suppose we had a solution $\mathbf{q}$ to the equation $A\mathbf{q} = \mathbf{p}$; we would then have an error estimate $\varepsilon = h^2\mathbf{q}$. Usually this is asking too much. But if we are able to *bound* the vector $\mathbf{p}$ by a vector $\mathbf{r} \geq \mathbf{p}$ then the solution $\mathbf{s}$ to $A\mathbf{s} = \mathbf{r}$ bounds $\mathbf{q}$ by the discrete maximum principle: $\mathbf{q} \leq \mathbf{s}$. This gives us an error estimate as well: $\varepsilon \leq h^2\mathbf{s}$. Such a *super solution* $\mathbf{s}$ 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 $\mathbf{r}$ would be the constant vector $h^2\|\mathbf{p}\|_\infty$. We will show that to get the solution $\mathbf{s}$, 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 circle 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 $A\mathbf{u} = \mathbf{e}$ in which $\mathbf{e}$ contains all ones and the single index vector $\mathbf{u}$ 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 3.6.5 *Show that* $\|\mathbf{u}\|_\infty = \frac{1}{4}$. □

Exercise 3.6.6 *Give the solution of* $-\Delta u = 1$ *with homogeneous Dirichlet boundary conditions on a circle C 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 C. 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 $A\mathbf{u} = \mathbf{e}$ but rather of a perturbed system $A\tilde{\mathbf{u}} = \mathbf{e} + \mathbf{e}_b$. The vector e_b contains the interpolation error of $O(h^2)$ at the boundary.

Exercise 3.6.7 *Consider the discretization of* $-\Delta u = 1$ *with homogeneous Dirichlet boundary conditions on the circle 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* $O(h^2)$. □

Exercise 3.6.8 *Show by using Theorem 3.6.2 and the result of Exercise 3.6.7 that* $\tilde{\mathbf{u}} - \mathbf{u} = O(h^2)$. □

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

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

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

Exercise 3.6.9 *Explain why the midpoint of the circle does not play a role in Theorem 3.6.3. Is it true that we can take the smallest circle that wholly contains G?* □

Exercise 3.6.10 *Show that if $A\mathbf{w} = \|\mathbf{p}\|_\infty\mathbf{e}$, then $|\varepsilon| < h^2\mathbf{w}$.* □

Exercise 3.6.11 *Prove Theorem 3.6.3.* □

3.7 Boundary fitted coordinates

In Section 3.5 we payed attention to 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 co-ordinate transformation such that the grid spacing is uniform in the transformed co-ordinate. Let this co-ordinate 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 co-ordinate transformation and we require that Γ is a *bijection* (that is, Γ is *one-to-one*). 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 equation (3.7.1), but it can be solved on an equidistant grid. After

the equation is solved, the solution is mapped onto the gridnodes on the x-number line. In practice, one often does not know the function $\Gamma(\xi)$ in an explicit form, then 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 $x = \Gamma(\xi) = \xi^2(3-2\xi)$. 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 circle.

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

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

$$x = rcos\theta, \qquad y = rsin\theta. \tag{3.7.7}$$

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*

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

□

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*

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

□

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 of Poisson's equation on a circle of radius 1 in the (r,θ)-plane. Use homogeneous Dirichlet boundary conditions on the circle. 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 point-wise 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 $(0,1)$:

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

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.

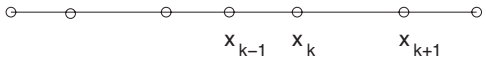


Figure 4.1: Non equidistant grid.

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

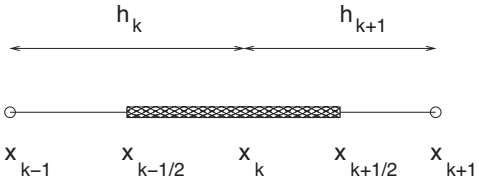


Figure 4.2: Control volume.

$x_{k+1} - x_k$ 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

$$\int_{x_{k-1/2}}^{x_{k+1/2}} -\frac{d}{dx}\left(\lambda \frac{dT}{dx}\right)\, dx = \int_{x_{k-1/2}}^{x_{k+1/2}} f\, dx, \tag{4.1.2}$$

$$-\lambda \frac{dT}{dx}\Big|_{x_{k+1/2}} + \lambda \frac{dT}{dx}\Big|_{x_{k-1/2}} = \int_{x_{k-1/2}}^{x_{k+1/2}} f\, dx. \tag{4.1.3}$$

Equation (4.1.3) represents the physical conservation law: the difference between the influx and outflux 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:

$$\lambda_{k-1/2}\frac{T_k - T_{k-1}}{h_k} - \lambda_{k+1/2}\frac{T_{k+1} - T_k}{h_{k+1}} = \frac{1}{2}(h_k + h_{k+1})f_k + E_T, \tag{4.1.4}$$

which after rearrangement becomes:

$$-\frac{\lambda_{k-1/2}}{h_k}T_{k-1} + \left(\frac{\lambda_{k-1/2}}{h_k} + \frac{\lambda_{k+1/2}}{h_{k+1}}\right)T_k - \frac{\lambda_{k+1/2}}{h_{k+1}}T_{k+1} = \frac{1}{2}(h_k + h_{k+1})f_k + E_T. \tag{4.1.5}$$

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 (4.1.5) without the error term is identical to the finite difference discretization of (4.1.1) multiplied by the length h.* □

The error E_T in Equation (4.1.5) 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 exercise, it is shown that $E_1 = O(h_k^2 - h_{k+1}^2)$ and $E_2 = O(h_k^2 - h_{k+1}^2)$. Further, if the grid spacing is determined by $h_{k+1} = h_k(1 + O(h))$, then it can be shown that $E_1 = O(h^3)$ and $E_2 = O(h^3)$. The global error is one order lower, that is $O(h^2)$, 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* $E_1 = O(h_k^2 - h_{k+1}^2)$.
Hint: Assume that $f(x)$ *is the derivative of* $F(x)$. *Express the integral in F and use Taylor series expansion.* □

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

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

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:

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

The left-hand boundary condition can be substituted directly:

$$-\lambda \frac{dT}{dx}\bigg|_{x_{1/2}} = \int_{x_0}^{x_{1/2}} f \, dx. \tag{4.1.7}$$

Application of central differences and one point integration gives:

$$\frac{\lambda_{1/2}}{h_1} T_0 - \frac{\lambda_{1/2}}{h_1} T_1 = \frac{1}{2} h_1 f_0 + E_T. \tag{4.1.8}$$

The truncation error E_T is $O(h_1^2)$ in this equation.

Exercise 4.1.5 *Show that* E_T *is* $O(h_1^2)$ *in the above equation.* □

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

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

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

$$\lambda \frac{dT}{dx}\bigg|_{x_{N-1/2}} + \alpha T_N = \int_{x_{N-1/2}}^{x_N} f \, dx + \alpha T_R. \tag{4.1.10}$$

Application of central differences and one point integration gives:

$$-\frac{\lambda_{N-1/2}}{h_N} T_{N-1} + \left(\frac{\lambda_{N-1/2}}{h_N} + \alpha \right) T_N = \frac{1}{2} h_N f_N + \alpha T_R + E_T. \tag{4.1.11}$$

Remark
If we have for example a Dirichlet boundary condition $T = T_0$, at the left-hand side 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 *fluxes* by applying Fick's (Darcy's, Ohm's, Fourier's) law for each finite volume (x_L, x_R) the equation looks like:

$$q_R - q_L = \int_{x_L}^{x_R} f \, dx, \tag{4.1.12}$$

or in words: what flows out minus what flows in equals the local production. 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 inflow flux plus the maximum error in the production provided the flux itself is not discretized.

Exercise 4.1.6 *Show, that if the equation*

$$-\left(\lambda y'\right)' = 0 \tag{4.1.13}$$

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*

$$-\left(\lambda y'\right)' = 1 \tag{4.1.14}$$

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*.* □

We call the *error in the fluxes* $\mathbf{dq}$ and we shall calculate the various contributions to it in the following exercises.

Exercise 4.1.8 Propagation of production error
Let $dq_k - dq_{k-1} = h_k E_k$*, where* $\sum_k h_k = 1$*. Show that* $|dq_k| < |dq_0| + \sup_{j \le k} |E_j|$*.* □

Exercise 4.1.9 Propagation of boundary error
Let $dq_k - dq_{k-1} = 0$*. Show that* $dq_k = dq_0, k = 1, \ldots, N$*.* □

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 $O(h^2)$ in the fluxes, q_k, for smoothly varying step sizes. By the same reasoning as in Exercise 4.1.8 we may now show, that the error in the temperatures *remains* $O(h^2)$, because if

$$-\lambda \tilde{T}'(x_{k+1/2}) = q_{k+1/2} + O(h^2), \tag{4.1.15}$$

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

$$\lambda_{k+1/2} \frac{dT_k - dT_{k+1}}{h_{k+1}} = E_{k+1}, \tag{4.1.16}$$

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

$$|dT_k| < |dT_N| + \sup_{j \geq k} |E_j| / \lambda_{j-1/2}. \tag{4.1.17}$$

However, by the right-hand-boundary condition we know that $q_N = \alpha(T_N - T_R)$ and that the numerical approximation to q_N has an error of $O(h^2)$. Therefore $dT_N = O(h^2)$ and backsubstitution into inequality (4.1.17) 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:

$$-\text{div}\, \lambda \,\text{grad}\, u = f, \quad \mathbf{x} \in \Omega, \tag{4.2.1a}$$

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

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

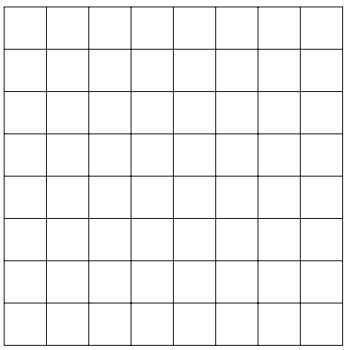


Figure 4.3: Subdivision of rectangular region into cells.

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 in 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 with edges half way between two nodes, like in Figure 4.4. We integrate the equation

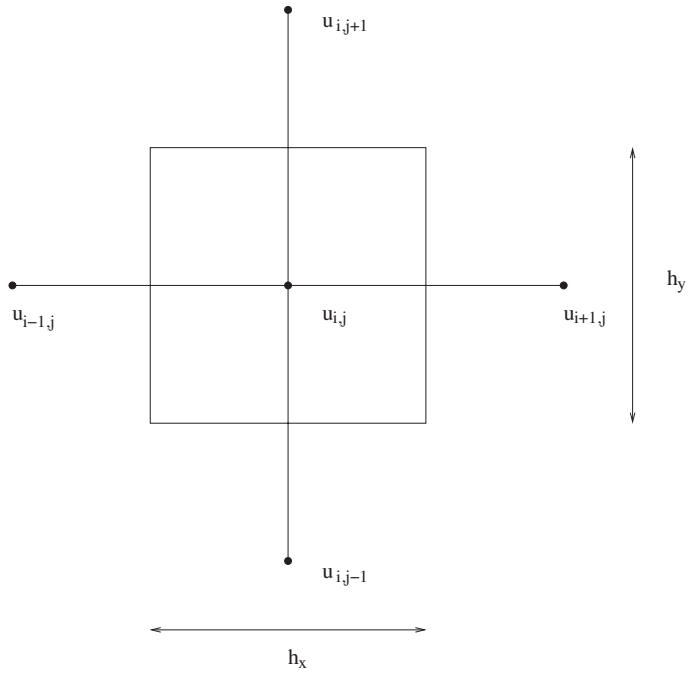


Figure 4.4: Control volume for the Diffusion equation.

over the control volume to obtain:

$$\int_V -\text{div}\,\lambda\,\text{grad}\,u\;dV = \int_V f\;dV, \tag{4.2.2a}$$

$$\oint_\Gamma -\lambda\frac{\partial u}{\partial n}\;d\Gamma = \int_V f\;dV. \tag{4.2.2b}$$

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:

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

Note that Equation (4.2.3) is identical to the finite difference Equation (3.4.2).

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

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

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

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

4.2.1 Boundary conditions

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 Laplacian equation (4.2.1a) over the

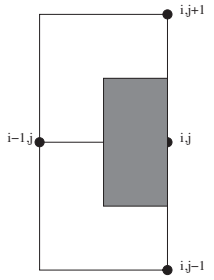


Figure 4.5: Control volume for the Robin boundary condition.

control volume gives Equation (4.2.2b). 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, however their length is multiplied by $\frac{1}{2}$. On the east edge boundary condition (4.2.1b) is applied to get

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

Discretization of (4.2.5) gives

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

So the complete discretization for point $u_{i,j}$ at the boundary becomes

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

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 that we have periodical boundary conditions at the boundaries $x = 0$ and $x = 1$. Furthermore boundary condition (4.2.1b) holds for the other two boundaries.*

- *Formulate the periodical boundary conditions at $x = 0$ and $x = 1$. Motivate why the number of boundary conditions is correct.*
- *Derive the finite volume discretization of the equation at the periodical 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 a 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 most easy 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, it is necessary to introduce a virtual point $i, j-1$ like in Figure 3.15. The value of $u_{i,j-1}$ can be expressed in $u_{i,j}$ and the boundary value $u_{i,j-1/2}$ using linear extrapolation. Substitution in the 5-point molecule results in a 4-point stencil.

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

$$\int_{\Gamma} \alpha(u - u_0)\, d\Gamma. \tag{4.2.9}$$

u is unknown, and not present on the boundary either. In order to keep the second order accuracy, the best way is express u using linear extrapolation from two internal points. Consider for example the south boundary in Figure 4.6. We can express

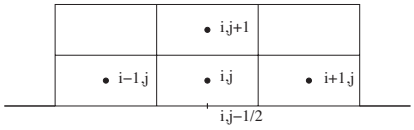


Figure 4.6: Control volume for the cell-centered Robin boundary condition.

u_B in u_C and u_N using linear extrapolation, resulting again in a 4-point molecule.

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

4.2.3 Boundary cells in case of a skewed boundary

The best way to treat a skewed boundary is to make sure, that control volume vertices fall *on* the boundary. This leads to triangle shaped grid-cells at the boundary, see Figure 4.7. Integration over the triangle and substitution of central differences give with the notations of Figure 4.7:

$$-\beta_W u_W - \beta_S u_S + (\beta_W + \beta_S) u_C + \int_{hyp} -\lambda \frac{\partial u}{\partial n}\, d\Gamma = \frac{1}{2} h_x h_y f_C, \tag{4.2.10}$$

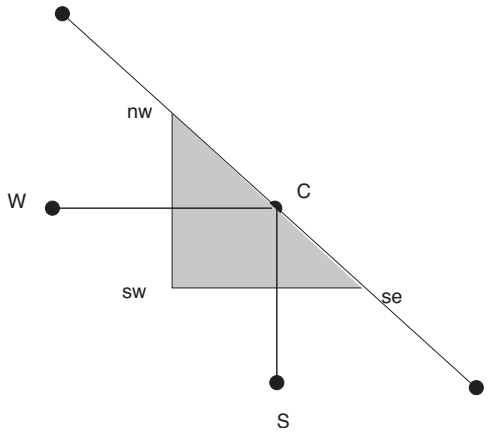


Figure 4.7: Boundary cell.

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 we get:

$$-\beta_W u_W - \beta_S u_S + (\beta_W + \beta_S + \alpha h_h) u_C = \alpha h_h u_{0C} + \frac{1}{2} h_x h_y f_C. \tag{4.2.11}$$

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 a diagonally dominant Z-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 of equations* (4.2.1). *If $f \geq 0$, for $\mathbf{x} \in \Omega$ and $u_0 \geq 0$, for $\mathbf{x} \in \Gamma$, then the solution of the discrete problem is positive.*

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 of equations* (4.2.1). *The solution of the discrete problem with $f = 0$ has a maximum and minimum at 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*:

$$\Phi_w + \Phi_n + \Phi_e + \Phi_s = \int_V f \, dV, \tag{4.2.12}$$

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. $O(h^2)$ accurate for smoothly changing step sizes, otherwise $O(h)$.
2. Approximate derivatives by central differences. $O(h^2)$ accurate for smoothly changing step sizes otherwise $O(h)$.
3. Approximate the right-hand side by one point integration. $O(h^2)$ accurate for smoothly changing step sizes otherwise $O(h)$.

It gets monotonous. From finite difference approximations we already know, that *equidistant* step sizes lead to overall $O(h^2)$ accuracy. So what are smoothly varying step sizes? Roughly speaking it says, that between neighboring step sizes there may be a factor $1 + O(h)$. This still gives 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 $O(h)$, because the integration point has to be the gravicenter for order $O(h^2)$ accuracy, whereas the integration point is always on the edge. (Note that in fact the absolute magnitude of the error is $O(h^3)$, but that is because the volume of integration is itself $O(h^2)$.)

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 $\mathbf{u}$ of the discrete equations with $f = 0$, we have

$$\|\mathbf{u}\|_\infty \leq \sup_{\mathbf{x}\in\Gamma} |u_0|. \tag{4.2.13}$$

Exercise 4.2.12 *Prove Inequality* (4.2.13)*. Use the results of Exercises 4.2.9 sseq.* □

Exercise 4.2.13 *Prove that if $\tilde{u}_0 = u_0 + \varepsilon_0$ 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* (4.2.13)*)* □

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

4.3 Laplacian in general coordinates

4.3.1 Discrete 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

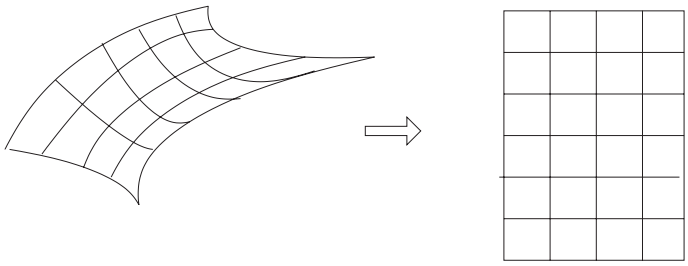


Figure 4.8: General region transformation.

$x(\xi,\eta)$ 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 $\mathbf{x}(\xi_0,\eta)$ and $\mathbf{x}(\xi,\eta_0)$ respectively. Such a transformation is called regular if it has an inverse, otherwise it is singular. Sufficient conditions for regularity is, that the *Jacobian* exists and is non-singular. The quantities needed to calculate the transformations are the derivatives of the transformation matrices:

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

and its inverse

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

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 in Figure 4.9, central

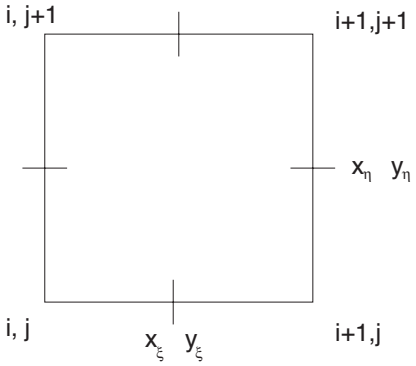


Figure 4.9: One cell with natural place of coordinate derivatives.

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 *Explain how to express $x_\xi, x_\eta, y_\xi, y_\eta$ in the cell center in the $\xi\eta-$plane (see Figure 4.9) in the cell coordinates in the $xy-$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 that is defined in the (x, y)-space, and suppose that Ω is mapped onto $\overline{\Omega}$ in the (ξ, η)-space, then, from Calculus, it follows

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

where the Jacobian is defined by

$$|\frac{\partial(x,y)}{\partial(\xi,\eta)}| = |det(J)|,$$

which is expressed in the co-ordinate framework (ξ, η). We use the notation $d\Omega_{xy}$ and $d\Omega_{\xi\eta}$ to emphasize that the integral is in the (x, y) and (ξ, η) framework respectively. For the derivation of this procedure, we refer to a textbook on Calculus, like Steward, Adam or Almering. 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 co-ordinate system.

4.3.2 An example of finite volume integration in polar co-ordinates

We will 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 co-ordinates by

$$\Omega_{r\theta} = \{(r,\theta) \in R^2 : 1 < r < 3,\ 0 < \theta < \pi/4\}.$$

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

$$-\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}\right) = f(rcos\theta, rsin\theta). \tag{4.3.6}$$

Note that $\Omega_{r\theta}$ is a rectangular domain in the (r, θ)- co-ordinate framework. The Jacobian of the transformation from Cartesian co-ordinates to polar co-ordinates 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 transformed control volume, which is rectangular and hence much easier to work with, to get

$$\int_V -\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}\right) r d\Omega_{r\theta} = \int_V f(rcos\theta, rsin\theta) r d\Omega_{r\theta}. \tag{4.3.8}$$

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 such that

$$\int_V -\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) - \frac{1}{r}\frac{\partial^2 u}{\partial \theta^2} d\Omega_{r\theta} = -\int_V (\frac{\partial}{\partial r}, \frac{\partial}{\partial \theta}) \cdot (r\frac{\partial u}{\partial r}, \frac{1}{r}\frac{\partial u}{\partial \theta}) d\Omega_{r\theta}. \tag{4.3.9}$$

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

$$-\int_{\delta V} (n_r, n_\theta) \cdot (r\frac{\partial u}{\partial r}, \frac{1}{r}\frac{\partial u}{\partial \theta}) d\Gamma = \int_V f(rcos\theta, rsin\theta) r d\Omega_{r\theta}. \tag{4.3.10}$$

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 co-ordinates (r_C, θ_C) in Figure 4.10. The integral at the right-hand side in the above equation is approximated by

$$\int_V f(rcos\theta, rsin\theta) r d\Omega_{r\theta} \approx f(r_C, \theta_C) r_C \Delta r \Delta\theta. \tag{4.3.11}$$

The boundary integral is obtained by the sum of the approximations of integrals over all the boundary segments. Substitution of these approximations into (4.3.10), gives the final result for an internal control volume:

$$\frac{1}{r_C}\frac{u_S - u_C}{\Delta\theta}\Delta r + r_e\frac{u_E - u_C}{\Delta r}\Delta\theta + \frac{1}{r_C}\frac{u_N - u_C}{\Delta\theta}\Delta r + r_w\frac{u_W - u_C}{\Delta r}\Delta\theta = f(r_C, \theta_C) r_C \Delta r \Delta\theta. \tag{4.3.12}$$

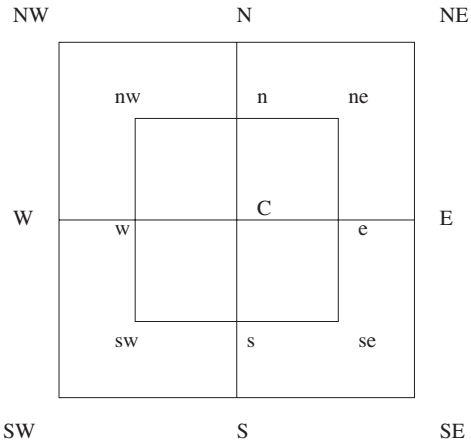


Figure 4.10: General control volume.

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

$$\frac{\partial u}{\partial n} = \alpha(u_0 - u).$$

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. Integrating over the half volume and applying the divergence theorem

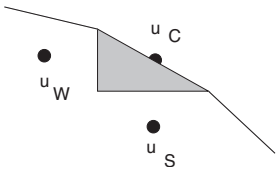


Figure 4.11: Boundary cell.

we get:

$$\frac{1}{r_C}\frac{u_S - u_C}{\Delta\theta}\frac{\Delta r}{2} + r_C\alpha(u_0 - u_C)\Delta\theta + \frac{1}{r_C}\frac{u_N - u_C}{\Delta\theta}\frac{\Delta r}{2} + r_w\frac{u_W - u_C}{\Delta r}\Delta\theta = f(r_C, \theta_C)r_C\Delta r\frac{\Delta\theta}{2}, \tag{4.3.13}$$

where the radiation boundary condition has been substituted into the boundary integral of the right (east) boundary of the control volume.

4.3.4 Error analysis

We did a comparable error analysis of the Laplace equation in Cartesian coordinates in Section 4.2.5. Consider the one point integration of the boundary volume on the right-hand side:

$$\int_V \sqrt{g} f dV = \frac{1}{2}h^2(\sqrt{g}f)_C + O(h^3), \tag{4.3.14}$$

as you can simply verify by Taylor expansion. So this "inaccurate" integration of the right-hand side gives a perturbation of $O(h^3)$ in the approximated right-hand side. The same is true for the integrations along n and s sides, *because they integrate the same integrand and are subtracted from each other.*

Theorem 4.3.1 *For sufficiently smooth f*

$$\int_x^{x+h} f(x, y+h) - f(x, y)dx = h(f(x, y+h) - f(x, y)) + O(h^3). \tag{4.3.15}$$

Proof
Let $F(x, y)$ be such, that $F_x(x, y) = f(x, y)$, then apparently

$$\int_x^{x+h} f(x, y+h) - f(x, y)dx = F(x+h, y+h) - F(x, y+h) - F(x+h, y) + F(x, y). \tag{4.3.16}$$

Now by Taylors theorem:

$$F(x+h,y+h) = F(x,y+h) + hf(x,y+h) + \frac{h^2}{2}f_x(x,y+h) + O(h^3), \quad (4.3.17a)$$

$$F(x+h,y) = F(x,y) + hf(x,y) + \frac{h^2}{2}f_x(x,y) + O(h^3). \quad (4.3.17b)$$

Subtracting the two equations in (4.3.17) we get:

$$\int_x^{x+h} f(x,y+h) - f(x,y)dx = h(f(x,y+h) - f(x,y)) + \frac{h^2}{2}(f_x(x,y+h) - f_x(x,y)) + O(h^3). \quad (4.3.18)$$

From the mean value theorem we note that $f_x(x,y+h) - f_x(x,y) = O(h)$ and the result follows. □

So all "inaccurate" integrations produce an $O(h^3)$ perturbation in the right-hand side. And now we use the same argument as in Section 4.2.5. The perturbation in the solution of the homogeneous Laplacian is always *less* than the perturbation in the boundary condition. But a perturbation of $O(h^3)$ in the right-hand side of a boundary volume equation is equivalent to an $O(h^2)$ perturbation in the boundary condition, hence causes an $O(h^2)$ perturbation in the solution. Because we no longer have the discrete maximum principle at our disposal it is not so easy to formally prove this assertion. But the least we can say is, that if our discrete solution converges to the solution of the continuous problem it is $O(h^2)$ accurate, because we *do* have a maximum principle for the continuous problem.

4.4 Finite volumes on two component fields

We shall show an example of application of the FVM on a two component field. We recall the problem for planar stress from Section 2.4.4. We consider a rectangular plate fixed at the sides ABC and subject to a body force $\mathbf{b}$ inside $\Omega = ABCD$ and boundary stresses t at the two free sides CDA. See Figure 4.12 The equation for the stresses are:

$$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + b_1 = 0, \quad (4.4.1a)$$

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + b_2 = 0, \quad (4.4.1b)$$

We integrate the first equation over a control volume V_1 and the second one over a control volume V_2. We define

$$\mathbf{s}_x = \begin{pmatrix} \sigma_{xx} \\ \tau_{xy} \end{pmatrix} \quad \text{and} \quad \mathbf{s}_y = \begin{pmatrix} \tau_{xy} \\ \sigma_{yy} \end{pmatrix}. \quad (4.4.2)$$

After application of Gauss' divergence theorem we obtain:

$$\oint_{\Gamma_1} \mathbf{s}_x \cdot \mathbf{n}\, d\Gamma + \int_{V_1} b_1\, dV = 0, \quad (4.4.3a)$$

$$\oint_{\Gamma_2} \mathbf{s}_y \cdot \mathbf{n}\, d\Gamma + \int_{V_2} b_2\, dV = 0, \quad (4.4.3b)$$

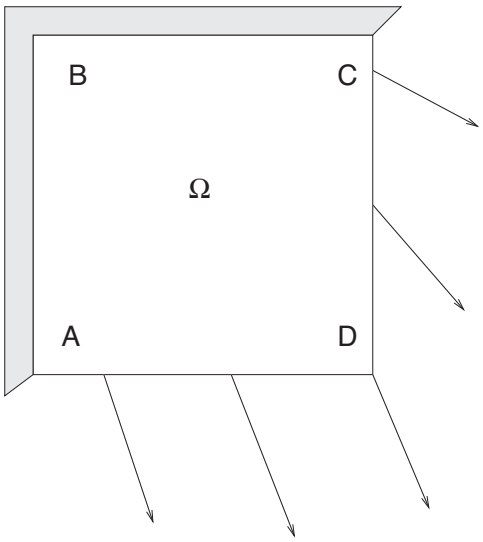


Figure 4.12: Square plate.

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) as for 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 u_x and v_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.

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 u_y and v_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.
Such an arrangement of nodal point is called a *staggered grid*. This means that in general different problem variables reside in different nodes.

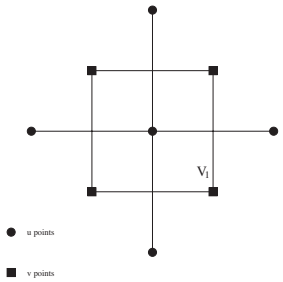


Figure 4.13: V_1-variables.

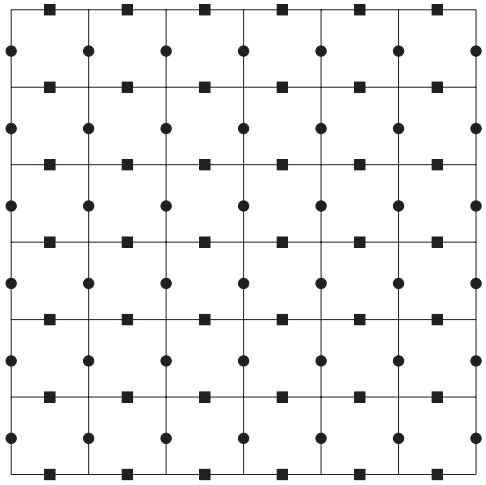


Figure 4.14: Staggered grid.

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 the displacement variables is:

$$B_W v_W + B_{nw} u_{nw} + B_N v_N + B_{ne} u_{ne} + B_E v_e + B_{se} u_{se} + B_S v_S + B_{sw} u_{sw} + B_C v_C = h^2 b_C. \tag{4.4.5}$$

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 $v_W =$

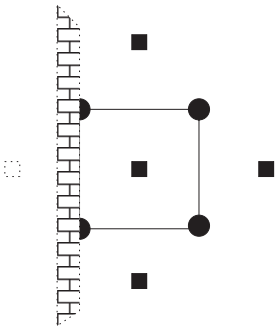


Figure 4.15: Ghost point.

$-v_C$ and Equation (4.4.5) transforms into

$$B_{nw}u_{nw} + B_N v_N + B_{ne}u_{ne} + B_E v_e + B_{se}u_{se} + B_S v_S + B_{sw}u_{sw} + (B_C - B_W)v_C = h^2 b_C. \tag{4.4.6}$$

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. (Figure 4.16)
Let us integrate Equation (4.4.1a) over a half volume V_1 to obtain:

$$h(-s_{xxw} + s_{xxC}) + \frac{1}{2}h(\tau_{xyn} - \tau_{xys}) = -\frac{1}{2}h^2 b_{1C} \tag{4.4.7}$$

Since by the natural boundary conditions $s_{xx} = f_1$ and $\tau_{xy} = f_2$ are given quantities at the boundary this transforms into

$$hs_{xxW} = hf_{1C} + \frac{1}{2}h(f_{2n} - f_{2s})) + \frac{1}{2}h^2 b_{1C}. \tag{4.4.8}$$

Again one point integration of the right-hand side causes a perturbation of $O(h^3)$, 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 $O(h^3)$.

Exercise 4.4.5 *Prove these last two assertions. Compare your results with Section 4.3.4* □

Since this perturbation is of the same order as a perturbation of $O(h^2)$ 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.

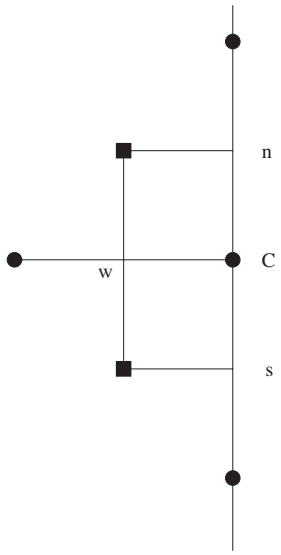


Figure 4.16: Half volume at natural boundary.

4.5 Project: 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).

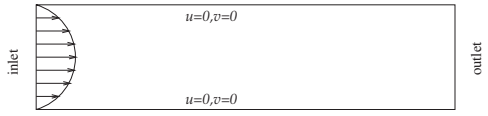


Figure 4.17: Channel for Stokes flow.

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

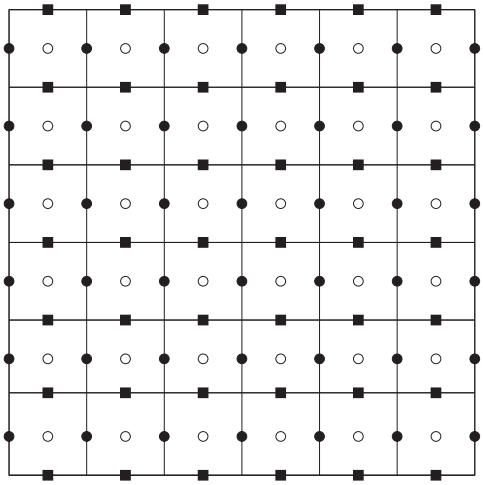


Figure 4.18: The ordering of the unknowns in a staggered approach for the Stokes equations. The solid circles and squares respectively correspond to u and v indicating the horizontal and vertical components of the fluid velocity. The open circles denote the pressure nodes.

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 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 div $\mathbf{u} = 0$. This equation is discretized using a control volume with a pressure node as the center:

$$\int_{\Omega_p} \text{div } \mathbf{u} d\Gamma = \int_{\Gamma_p} \mathbf{u} \cdot \mathbf{n} d\Gamma. \tag{4.5.5}$$

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 is given by

$$\int_{\delta\Omega_u^R} \left(-\mu \frac{\partial u}{\partial x} + p n_x \right) d\Gamma = \int_{\Gamma_u^R} p_0 d\Gamma \approx p_0 h.$$

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 $O(h^3)$. Show this and argue, that this is equivalent to a perturbation of $O(h^2)$ 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 equidistant 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 planar 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

Minimization problems in physics

Objectives

In Chapter 1 we have seen that many physical partial differential equations (PDEs) are the result of conservation laws. A completely different way to derive PDEs is by minimizing an integral. Examples of this approach are: shortest path and minimal potential energy. In this chapter we shall show how to derive a PDE with corresponding boundary conditions starting from a minimization problem.

On the other hand it is possible, under certain conditions, to derive a minimization problem, that in some sense is equivalent to a given PDE. If the PDE has a solution in the classical sense, solution of the minimization problem means also solution of the corresponding PDE and vice versa.

Minimization problems usually admit a larger solution class than a PDE formulation and therefore the solution of the minimization problem is referred to as generalized solution of the PDE. A similar formulation is possible for problems that do not fit the minimization frame work. This formulation, the *weak formulation* will be treated in Chapter 7. It will be shown that this formulation may be considered as a kind of conservation law.

5.1 Introduction

Mathematical models in physics are often derived from conservation laws (see Section 1.3.5), but also from minimization problems (Section 1.4). In this chapter we shall focus on the latter category.

Before analyzing these minimization problems in general, we shall start with the simple example of minimum potential energy, already treated in Section 1.4.1.

5.1.1 Minimal potential energy

In Section 1.4.1 we have seen that the potential energy of an elastic string fixed in $(0, 0)$ and $(0, 1)$ with a given load is defined by

$$\int_0^1 \left\{ \frac{1}{2} k \left(\frac{du}{dx} \right)^2 - uf \right\} dx \,. \tag{5.1.1}$$

Hence the displacement u minimizes the integral (5.1.1) under the conditions:

$$u(0) = 0\,, \quad u(1) = 0\,.$$

We shall consider a slightly more general minimization problem:

Find the function u that minimizes $I(u)$ defined by (5.1.1) such that

$$u(0) = u_0\,. \tag{5.1.2}$$

In the remainder of this chapter we shall always assume that u is sufficiently smooth, which means that implicitly we suppose that all expressions we use and operations we apply are allowed. Later on we shall specify this more precisely.

In Section 5.4 we shall give a number of often classical examples of minimization problems.

The minimization problem (5.1.1), (5.1.2) is different from standard minimization problems in the sense that we have to find a continuous function instead of a finite set of parameters. In fact we may consider this as a problem with an infinite number of unknowns.

5.1.2 Derivation of the differential equation

In order to show that the solution of the minimization problem satisfies a certain differential equation we use a reasoning due to Euler.
We suppose that (5.1.1) with boundary condition (5.1.2) has a smooth solution, which we call $\hat{u}(x)$. We consider a class of functions $u(x)$ defined as

$$u(x) = \hat{u}(x) + \varepsilon\eta(x)\,. \tag{5.1.3}$$

(5.1.3) will be referred to as a variation around $\hat{u}(x)$.

ε is a variable parameter and η is some arbitrary but fixed function. Since both $\hat{u}(x)$ and $u(x)$ must satisfy the boundary condition (5.1.2) it is necessary that

$$\eta(0) = 0\,. \tag{5.1.4}$$

Also η is assumed to be sufficiently smooth.
Substitution of (5.1.3) in (5.1.1) gives

$$I(\hat{u}+\varepsilon\eta) = \int_0^1 \left\{ \frac{1}{2}k\left(\frac{d(\hat{u}+\varepsilon\eta)}{dx}\right)^2 - (\hat{u}+\varepsilon\eta)f \right\} dx\,, \tag{5.1.5}$$

under the conditions (5.1.2) and (5.1.4).

I is a function of ε only (why?), and according to the classical theory of minimization problems, a necessary condition for the existence of an extreme is

$$\frac{dI}{d\varepsilon} = 0\,, \tag{5.1.6}$$

hence

$$\int_0^1 \left\{ k\frac{d(\hat{u}+\varepsilon\eta)}{dx}\frac{d\eta}{dx} - \eta f \right\} dx = 0\,. \tag{5.1.7}$$

Exercise 5.1.1 *Show that Equation (5.1.7) follows from (5.1.5) and (5.1.6)* □

From (5.1.3) it is clear that $I(\varepsilon)$ reaches its minimum for $\varepsilon = 0$, hence (5.1.7) reduces to

$$\int_0^1 \left\{ k\frac{d\hat{u}}{dx}\frac{d\eta}{dx} - \eta f \right\} dx = 0 \,, \qquad (5.1.8)$$
$$\hat{u}(0) = u_0 \,, \quad \eta(0) = 0 \,.$$

Since $\eta(x)$ is an arbitrary function, (5.1.8) must be valid for any η satisfying (5.1.4). Unfortunately we see in (5.1.8) both η and $\frac{d\eta}{dx}$ in the integral. In order to get an expression in η only, we apply integration by parts to the first term. This results in

$$\int_0^1 \left\{ -\eta\frac{d}{dx}\left(k\frac{d\hat{u}}{dx}\right) - \eta f \right\} dx + \eta k\frac{d\hat{u}}{dx}\Big|_0^1 = 0 \,. \qquad (5.1.9)$$

Due to the boundary condition (5.1.4) this reduces to:

$$\int_0^1 \eta\left(-\frac{d}{dx}\left(k\frac{d\hat{u}}{dx}\right) - f \right) dx + \eta(1)k(1)\frac{d\hat{u}}{dx}(1) = 0 \,. \qquad (5.1.10)$$

(5.1.10) must be valid for all $\eta(x)$ with $\eta(0) = 0$. Let us first consider the subset that also satisfies $\eta(0) = \eta(1) = 0$. Now (5.1.10) reduces to

$$\int_0^1 \eta\left(-\frac{d}{dx}\left(k\frac{d\hat{u}}{dx}\right) - f \right) dx = 0 \,, \qquad (5.1.11)$$

for all η with $\eta(0) = \eta(1) = 0$.

Hence the solution $\hat{u}(x)$ must satisfy the differential equation (using Lemma of Dubois-Reymond 5.2.2)

$$-\frac{d}{dx}\left(k\frac{du}{dx}\right) = f \,, \qquad (5.1.12)$$

with boundary condition $u(0) = u_0$.

(We shall show this more rigorously in Section 5.2).

(5.1.12) is a second order linear differential equation. So we need two boundary conditions in order to get a unique solution. To that end we consider the complete class of functions $\eta(x)$, with $\eta(0) = 0$. Substituting (5.1.12) in (5.1.10) gives:

$$\eta(1)k(1)\frac{d\hat{u}}{dx}(1) = 0 \,, \qquad (5.1.13)$$

with $\eta(1)$ arbitrary.

Hence, we arrive at

$$k(1)\frac{du}{dx}(1) = 0 \,, \qquad (5.1.14)$$

which is our second boundary condition.

So we started with the minimization problem (5.1.1) with one boundary condition (5.1.2) and we showed that the solution must satisfy the second order differential equation (5.1.12) with two boundary conditions (5.1.2) and (5.1.14). Apparently boundary condition (5.1.14) is hidden in the minimization problem. Such a boundary condition, that is not imposed explicitly, is called a *natural boundary condition*. Boundary condition (5.1.2), which must be satisfied both by the minimization problem and the differential equation is called an *essential boundary condition*. It limits the class in which to look for a solution.

In the next section we shall consider a more general problem in one dimension.

5.2 A general one-dimensional problem with first order derivatives

In the previous section we have seen how one can derive a differential equation from a minimization problem. In this section we consider a general minimization in 1-d with first order derivatives.

Theorem 5.2.1

Let $f(x,u,p)$ be a sufficiently smooth function.

Consider the minimization problem

$$\min_u \quad l(u) = \min_u \quad \int_{x_0}^{x_1} f(x,u,u') \, dx \,, \tag{5.2.1}$$

with boundary condition

$$u(x_0) = u_0 \,. \tag{5.2.2}$$

u' is a short notation for $\frac{du}{dx}$.
If a solution $\hat{u}$ of problem (5.2.1), (5.2.2) exists, then this solution must satisfy the differential equation

$$\frac{\partial f}{\partial u} - \frac{d}{dx}\frac{\partial f}{\partial u'} = 0 \,, \tag{5.2.3}$$

with boundary conditions

$$\hat{u}(x_0) = u_0 \quad \textit{(essential)} \,, \tag{5.2.4}$$

and

$$\frac{\partial f}{\partial u'}(x_1) = 0 \quad \textit{(natural)} \,. \tag{5.2.5}$$

REMARK: with $\frac{\partial f}{\partial u'}$ we mean: differentiate $f(x,u,p)$ to p and substitute $\frac{du}{dx}$ for p.

Proof
Consider the following family of curves around the solution $\hat{u}(x)$:

$$u(x) = \hat{u}(x) + \varepsilon\eta(x) \,, \tag{5.2.6}$$

with ε an arbitrary parameter and $\eta(x)$ an arbitrary, sufficiently smooth curve satisfying $\eta(x_0) = 0$.

Substitution of (5.2.6) in (5.2.1) gives

$$l(u) = \int_{x_0}^{x_1} f(x, \hat{u} + \varepsilon\eta(x), \hat{u}' + \varepsilon\eta'(x)) \, dx \,. \tag{5.2.7}$$

The integral in (5.2.7) is a function of ε denoted by $I(\varepsilon)$.
$I(\varepsilon)$ is minimal for $u = \hat{u}(x)$, hence $\varepsilon = 0$.
A necessary condition for the existence of a minimum in $\varepsilon = 0$ is

$$\left.\frac{dI(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = 0 \,, \tag{5.2.8}$$

or

$$\int_{x_0}^{x_1} \left\{ \frac{\partial f}{\partial u}\left(x, \hat{u}, \hat{u}'\right)\eta(x) + \frac{\partial f}{\partial u'}(x, \hat{u}, \hat{u}')\eta' \right\} dx = 0 \,. \tag{5.2.9}$$

Integration by parts of the last term results in

$$\int_{x_0}^{x_1} \left[\frac{\partial f}{\partial u} - \frac{d}{dx}\frac{\partial f}{\partial u'} \right] \eta(x)\, dx + \left[\eta(x) \frac{\partial f}{\partial u'} \right]_{x_0}^{x_1} = 0 \,. \tag{5.2.10}$$

$\eta(x)$ is an arbitrary smooth function with $\eta(x_0) = 0$. We first restrict ourselves to the subset of functions that also satisfy $\eta(x_1) = 0$. Then according to the Lemma of Dubois-Reymond, (lemma 5.2.2) it follows that $\hat{u}$ satisfies differential equation (5.2.3).

Subsequently we consider the complete set of functions $\eta(x)$. It is clear that natural boundary condition (5.2.5) must be satisfied. □

REMARK:
Differential equations that follow in this way from a minimization problem are known as *Euler-Lagrange equations*.

Lemma 5.2.2 *(Dubois-Reymond)*

Let $M(x) \in C([a,b])$ *and let*

$$\int_a^b M(x)\eta(x)\, dx = 0 \,, \tag{5.2.11}$$

for all $\eta(x) \in C([a,b])$ *with* $\eta(a) = \eta(b) = 0$.

Then

$$M(x) = 0 \quad on \quad [a,b] \,. \tag{5.2.12}$$

Proof
Suppose there is an $x_0 \in (a,b)$ such that $M(x_0) \neq 0$, for example $M(x_0) > 0$. Since $M(x) \in C(a,b)$ there exists a δ-neighborhood of x_0, $(x_0 - \delta, x_0 + \delta) \subset (a,b)$ such that $M(x) > 0$ if $|x - x_0| < \delta$, $(\delta > 0)$.

Now choose $\eta(x)$ as follows

$$\eta(x) = \begin{cases} (x - x_0 - \delta)^2 (x - x_0 + \delta)^2 & \text{if} \quad |x - x_0| < \delta \\ 0 & \text{elsewhere,} \end{cases}$$

then $\int_a^b M(x)\eta(x)\, dx = \int_{x_0-\delta}^{x_0+\delta} M(x)(x - x_0 - \delta)^2 (x - x_0 + \delta)^2\, dx > 0$.

This contradicts (5.2.11) for $x \in (a,b)$.
So from the continuity of $M(x)$ it follows that $M(x) = 0$ for $x \in [a,b]$. □

In the next section we shall extend the Euler Lagrange equations to $\mathbb{R}^2$.

5.3 A simple two-dimensional case

We have seen how the Euler-Lagrange equations are derived in one dimension. Now we shall extend the theory to two dimensions. First we shall consider a simple two-dimensional example. It will be shown that the only difference with $\mathbb{R}^1$ is that the integration by parts is replaced by Gauss' divergence theorem.

Consider a region Ω (Figure 5.1) in $\mathbb{R}^2$. The boundary Γ is subdivided into 2 parts Γ_1 and Γ_2.

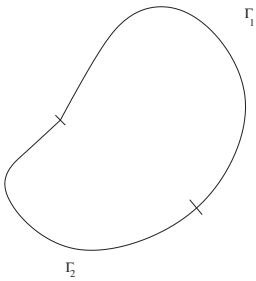


Figure 5.1: Region Ω with 2 boundary parts Γ_1 and Γ_2.

On Ω we consider the following minimization problem:
minimize the integral

$$I(u) = \int_{\Omega} \{\frac{k}{2}|\nabla u|^2 - uf\}\, d\Omega - \int_{\Gamma_2} u\, d\Gamma\,, \tag{5.3.1}$$

over the class of functions satisfying the boundary condition

$$u|_{\Gamma_1} = 0\,, \tag{5.3.2}$$

with $(k : \Omega \to \mathbb{R}^+)$.
With $|\nabla u|^2$ we mean $\nabla u \cdot \nabla u$.
To derive the Euler-Lagrange equations we proceed in exactly the same way as in $\mathbb{R}^1$.

So let $\hat{u}(x, y)$ be the function minimizing (5.3.1), (5.3.2) and consider the set of functions

$$u(\mathbf{x}) = \hat{u}(\mathbf{x}) + \varepsilon\eta(\mathbf{x})\,. \tag{5.3.3}$$

Substitution of (5.3.3) in (5.3.1) gives

$$I(\varepsilon) = \int_{\Omega} \{\frac{k}{2}|\nabla(\hat{u} + \varepsilon\eta)|^2 - (\hat{u} + \varepsilon\eta)f\}\, d\Omega - \int_{\Gamma_2} (\hat{u} + \varepsilon\eta)\, d\Gamma\,. \tag{5.3.4}$$

So the necessary condition for the existence of a minimum of (5.3.4) at $\varepsilon = 0$ is given by:

$$\int_{\Omega} \{k(\nabla\hat{u} \cdot \nabla\eta) - \eta f\}\, d\Omega - \int_{\Gamma_2} \eta\, d\Gamma = 0\,. \tag{5.3.5}$$

Exercise 5.3.1 *Derive formula (5.3.5)* □

In order to apply the two-dimensional version of the Lemma of Dubois-Reymond it is necessary to remove the term $\nabla\eta$.

Instead of classical integration by parts we use Gauss' divergence theorem (1.3.10):

$$\int_{\Omega} \mathrm{div}\, \mathbf{w}\, d\Omega = \oint_{\Gamma} \mathbf{w} \cdot \mathbf{n}\, d\Gamma\,.$$

By substituting $\mathbf{w} = \eta(k\nabla\hat{u})$ we get

$$\int_{\Omega} \nabla\eta \cdot (k\nabla\hat{u})\, d\Omega = -\int_{\Omega} \eta \mathrm{div}\,(k\nabla\hat{u})\, d\Omega + \oint_{\Gamma} \eta k\nabla\hat{u} \cdot \mathbf{n}\, d\Gamma\,. \tag{5.3.6}$$

REMARK: This is in fact the first equation of Green (see Exercise 1.3.8) .

Exercise 5.3.2 *Derive Equation (5.3.6).* □

A combination of (5.3.5) and (5.3.6) results in

$$\int_{\Omega} \{-\text{div}\,(k\nabla\hat{u}) - f\}\eta \; d\Omega + \int_{\Gamma_2} (k\nabla\hat{u}\cdot\mathbf{n} - 1)\eta \; d\Gamma = 0\,. \tag{5.3.7}$$

(Why?)

The two dimensional version of Dubois-Reymond's lemma leads to the PDE

$$-\text{div}\,(k\nabla\hat{u}) = f\,, \tag{5.3.8}$$

with boundary conditions,

$$u|_{\Gamma_1} = 0\,, \quad \text{(essential)} \tag{5.3.9}$$

and

$$k\frac{\partial u}{\partial n}\bigg|_{\Gamma_2} = 1\,. \quad \text{(natural)} \tag{5.3.10}$$

The technique applied here can also be used to solve a more general problem. This is done in Section 5.5. Before doing so we give a number of examples of minimization problems.

Exercise 5.3.3 *Prove Dubois-Reymond's lemma for a bounded region in two dimensions.* □

5.4 Examples of minimization problems

In this section we consider the following examples of minimization problems:

- Minimal surface problem, Section 5.4.1.
- Minimal potential energy, Section 5.4.2. This problem corresponds to a simple Poisson equation and is very suitable to demonstrate numerical techniques.
- Plane stress, Section 5.4.3. This is also a minimum potential energy problem, however, now we have an unknown vector instead of a scalar. As a consequence the corresponding PDE consists of a set of 2 ($\mathbb{R}^2$) or 3 ($\mathbb{R}^3$) coupled PDEs.
- Loaded and clamped plate (normal load), Section 5.4.4.
  Here we have a minimum potential energy problem involving second order derivatives. The corresponding PDE is of order four (Biharmonic equation).

5.4.1 Minimal surface problem

Let $u_c(\mathbf{x})$ be a given closed curve in $\mathbb{R}^3$. Let $c(\mathbf{x})$ be the projection of the curve in the plane ($\mathbb{R}^2$), and define the region Ω as the domain enclosed by c. We assume that u_c has a unique value for every $\mathbf{x} \in c$. The problem is to find the surface in $\mathbb{R}^3$ passing through u_c with minimum area in the domain Ω.

The area of the surface $z = u(x, y)$ is given by

$$s(u) = \int_{\Omega} \sqrt{1 + \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2}\; d\Omega\,. \tag{5.4.1}$$

The problem can be formulated as:
Find u smooth enough, satisfying the boundary conditions:

$$u(c) = u_c \ , \tag{5.4.2}$$

Such that $s(u)$ is minimal. Figure 5.2 shows an example of the solution of such a problem.

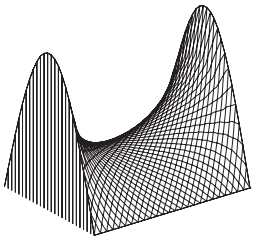


Figure 5.2: Solution of minimal surface problem.

5.4.2 Minimal potential energy

Consider the two rectangular conductors in Figure 5.3.

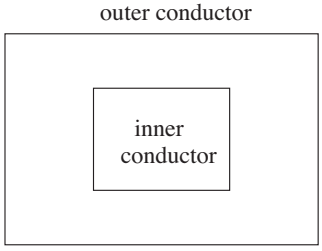


Figure 5.3: Two rectangular conductors.

Let the potential u at the inner conductor be equal to 0, and on the outer conductor be equal to 1. What is the potential u in the region between inner and outer conductor?
Due to symmetry arguments it is sufficient to consider only one quarter of the region, see Figure 5.4.
The principle of minimum potential energy requires that the potential distribution is such that the field energy is minimal.
The energy is given by [32],

$$p(u) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 \, d\Omega \ . \tag{5.4.3}$$

The mathematical formulation of this problem is:

> Minimize the integral $p(u)$ over the class of sufficiently smooth functions with boundary conditions
>
> $$\begin{array}{ll} u = 0 & \text{on } \Gamma_1 \ , \\ u = 1 & \text{on } \Gamma_2 \ . \end{array} \tag{5.4.4}$$
>
> (see Figure 5.4).

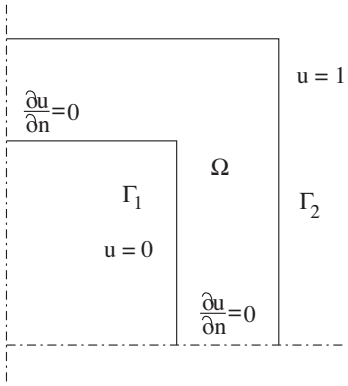


Figure 5.4: One quarter of the potential problem.

5.4.3 Small displacement theory of elasticity (Plane stress)

Consider the flat thin plate of Figure 5.5. See for example [31]. We assume that the thickness of the plate is small compared to its diameter. The outer load is uniform over the cross-section. The load is applied in the same plane as the plate. Along Γ_1 the plate is clamped.

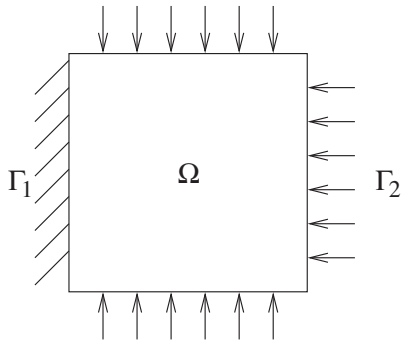


Figure 5.5: Flat plate clamped in Γ_1 and with uniform load on Γ_2.

Unknowns that we want to determine in this problem are the displacement vector $\mathbf{u} = \begin{bmatrix} u \\ v \end{bmatrix}$ and the stress tensor $\boldsymbol{\sigma} = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix}$. These are not independent.

The potential energy of the plate is defined as:

$$P(\mathbf{u}) = \frac{1}{2}\int_{\Omega} (\sigma_{xx}\varepsilon_x + \sigma_{yy}\varepsilon_y + \gamma_{xy}\tau_{xy})\, d\Omega - \int_{\Gamma_2} (t_1 u + t_2 v)\, d\Gamma\,, \tag{5.4.5}$$

where $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_x & \gamma_{xy} \\ \gamma_{xy} & \varepsilon_y \end{bmatrix}$ denotes the strain tensor and

$\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$ the external load vector.

We suppose that are no body forces in this case.

In order to get expression (5.4.5) in one type of unknown we need the *strain-displacement relations*:

$$\varepsilon_x = \frac{\partial u}{\partial x}\,, \quad \varepsilon_y = \frac{\partial v}{\partial y}\,, \quad \gamma_{xy} = [\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}]\,. \tag{5.4.6}$$

We also need a *constitutive equation* which relates stress to strain.
If we assume that the material is elastic, i.e. satisfies Hooke's Law then we get the following relations:

$$\begin{aligned} \sigma_{xx} &= \frac{E}{1-\nu^2}\,(\varepsilon_x + \nu\varepsilon_y), \\ \sigma_{yy} &= \frac{E}{1-\nu^2}\,(\nu\varepsilon_x + \varepsilon_y), \\ \tau_{xy} &= \frac{E}{1-\nu^2}\,\frac{1-\nu}{2}\,\gamma_{xy}, \end{aligned} \tag{5.4.7}$$

with E the elasticity modulus and ν Poisson's ratio.
With these relations we can express the potential energy in the displacements only.

Exercise 5.4.1 *Show that the potential energy (5.4.5) with the relations (5.4.6) and (5.4.7) can be written as:*

$$\begin{aligned} P(\mathbf{u}) &= \frac{1}{2}\int_\Omega \{A\frac{\partial u}{\partial x}(\frac{\partial u}{\partial x} + \nu\frac{\partial v}{\partial y}) + B(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x})(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) \\ &\quad + A\frac{\partial v}{\partial y}(\nu\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y})\}\, d\Omega - \int_{\Gamma_2} (t_1 u + t_2 v)\, d\Gamma\,, \end{aligned} \tag{5.4.8}$$

with $A = \frac{E}{(1-\nu^2)}$ *and* $B = \frac{E}{2(1+\nu)}$ *(E and ν constant).* □

The mathematical formulation of this problem is:

Find u, v with

$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_1 \tag{5.4.9}$$

such that the integral $P(u)$ in (5.4.8) is minimal.

5.4.4 Loaded and clamped plate

Consider the small plate of Figure 5.6.

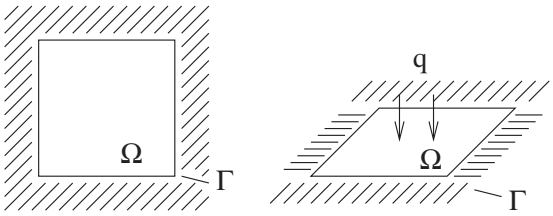


Figure 5.6: Clamped plate with normal load q, domain Ω and boundary Γ.

The load q is normal to the plate and the potential energy is given by

$$I(u) = \int_\Omega \frac{1}{2}[w_{xx}^2 + 2w_{xx}w_{yy} + w_{yy}^2 - 2qw]\, d\Omega\,, \tag{5.4.10}$$

where w is the displacement of the neutral face. We wish to determine w for a given load q. The mathematical formulation of this problem is:

Find w satisfying the boundary conditions:

$$\begin{array}{ll} w = 0 & \text{on } \Gamma\,, \\ \frac{\partial w}{\partial n} = 0 & \text{on } \Gamma\,. \end{array} \tag{5.4.11}$$

such that the integral $I(u)$ in (5.4.10) is minimal.

5.5 A two-dimensional problem

Theorem 5.2.1 can be generalized to two dimensions.

Theorem 5.5.1 *Let Ω be a domain in $\mathbb{R}^2$ with boundary Γ. Let Γ be subdivided into three parts Γ_1, Γ_2 and Γ_3:*

$$\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3\,.$$

The class of functions in which we try to find a solution is given by

$$\Sigma = \{u |\ u(\mathbf{x}) = g(\mathbf{x}),\ \ \forall \mathbf{x} \in \Gamma_1\}\,.$$

Let $F(x, y, u, p, q)$ and $f(x, y, u)$ be sufficiently smooth functions. Consider the following minimization problem

$$\min_{u\in\Sigma} J[u] = \min_{u\in\Sigma} \int_\Omega F(x, y, u, u_x, u_y)\, d\Omega + \int_{\Gamma_2} f(x, y, u)\, d\Gamma\,. \tag{5.5.1}$$

If there exists a solution $\hat{u}$ of this problem then $\hat{u}$ satisfies the PDE

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\frac{\partial F}{\partial u_x} - \frac{\partial}{\partial y}\frac{\partial F}{\partial u_y} = 0\,, \tag{5.5.2}$$

with boundary conditions

$$u = g\,, \quad \forall \mathbf{x} \in \Gamma_1 \tag{5.5.3}$$

$$\frac{\partial F}{\partial u_x} n_1 + \frac{\partial F}{\partial u_y} n_2 + \frac{\partial f}{\partial u} = 0\,, \quad \forall \mathbf{x} \in \Gamma_2 \tag{5.5.4}$$

$$\frac{\partial F}{\partial u_x} n_1 + \frac{\partial F}{\partial u_y} n_2 = 0\,, \quad \forall \mathbf{x} \in \Gamma_3\,. \tag{5.5.5}$$

where $\mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix}$ is the outward normal vector at the boundary.

Exercise 5.5.1 *Prove theorem 5.5.1 using the technique of Section 5.3.* □

5.6 Theoretical remarks

5.6.1 Smoothness requirements

In our derivations of the PDEs we have assumed that the solution is sufficiently smooth. This means that the solution must be so smooth that the differential equation exists. Hence if the PDE is of second order, it is necessary that the solution is in $C^2(\Omega)$. However in the original minimization problem we have only first order derivatives. Hence for the existence of a solution of the minimization problem it is sufficient that the first derivatives exist, or more precisely that the integral

makes sense. Usually this is translated by requiring that both the unknown u as its derivatives u_x and u_y are square integrable, i.e.

$$\int_\Omega u^2 \, d\Omega \,, \quad \int_\Omega u_x^2 \, d\Omega \quad \text{and} \quad \int_\Omega u_y^2 \, d\Omega$$

must exist and be finite.
In general this is even weaker than requiring that the derivatives exist everywhere in Ω.

If the solution of the minimization problem is not twice differentiable, it cannot satisfy the PDE. So actually a minimization problem may have a solution in a larger class of functions than the corresponding PDE. In fact the minimization problem can be seen as a generalization of that PDE.

5.6.2 Boundary conditions

We have seen that we must distinguish between essential and natural boundary conditions. Essential boundary conditions are conditions that have to be satisfied by all functions in the function class where we seek the solution. Natural boundary conditions appear naturally from the minimization problem once we derive the corresponding Euler-Lagrange equations.

In general we can state the following:

> If a minimization problem contains derivatives of first order and not higher, the corresponding Euler-Lagrange equation will be of second order. For such a problem essential boundary conditions have always the form $u = g_0, \mathbf{x} \in \Gamma_0$. If a boundary condition contains first derivatives for this type of problems, it is always a natural boundary condition. See [37].

> If a minimization problem contains derivatives of second order and not higher, the corresponding PDE will be of fourth order. For such problem boundary conditions involving only u or first order derivatives of u are essential, boundary conditions involving second or third derivatives will be natural.

5.6.3 Weak formulation

Consider minimization problem (5.3.1) without the integral over Γ_2 and with Γ_1 equal to the whole boundary Γ, hence

$$\min_{u \in \Sigma} \; I(u) = \int_\Omega \{ \frac{k}{2} |\nabla u|^2 - uf \} \, d\Omega \,, \tag{5.6.1}$$

$$u|_\Gamma = 0 \,. \tag{5.6.2}$$

According to (5.3.7), the solution must satisfy

$$\int_\Omega \{ -\text{div}\,(k \nabla u) - f \} \eta \, d\Omega = 0 \,, \tag{5.6.3}$$

for all η in Σ.

This is precisely the differential equation multiplied by η and integrated over the domain. In Chapter 7 we shall use such a method to arrive at the *weak formulation*. In that case η is called a test function.

5.7 Exercises

Exercise 5.7.1 *Find the Euler-Lagrange equation for the minimal surface problem (5.4.1), with boundary conditions (5.4.2). Do not use theorem 5.5.1.* □

Exercise 5.7.2 *Find the Euler-Lagrange equations for the minimization problem in Section 5.4.2 by direct variation around the solution. Which boundary conditions are essential and which are natural?* □

Exercise 5.7.3 *Find the Euler-Lagrange equations for the minimization problem of Exercise 5.4.1 in Section 5.4.3. (assume $u = \hat{u} + \varepsilon\eta, v = \hat{v} + \varepsilon\xi$).*

Use the strain-displacement relations (5.4.6) and stress-strain relations (5.4.7) to rewrite the Euler-Lagrange equations in the form

$$-\frac{\partial \sigma_{xx}}{\partial x} - \frac{\partial \tau_{xy}}{\partial y} = 0\,,$$
$$-\frac{\partial \tau_{xy}}{\partial x} - \frac{\partial \sigma_{yy}}{\partial y} = 0\,.$$

□

Exercise 5.7.4 *Find the Euler-Lagrange equations for the minimization problem in Section 5.4.4.* □

Exercise 5.7.5 *Find the Euler-Lagrange equations for the rotation surface with minimal area defined by*

$$\min J[u] = 2\pi \int_{x_0}^{x_1} u\sqrt{1 + \left(\frac{du}{dx}\right)^2}\, dx\,,$$
$$u(x_0) = y_0\,.$$

□

Exercise 5.7.6 *Consider the region Ω of Figure 5.7.*

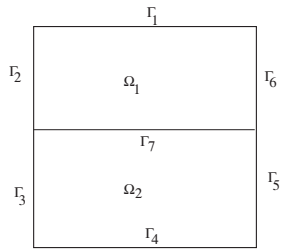


Figure 5.7: Region consisting of 2 layers.

In this region we have two layers Ω_1 and Ω_2 with different values of the permeability κ (κ_1 and κ_2). The pressure p in this layer satisfies the minimization problem

$$\min_{p\in\Sigma} \int_{\Omega} \frac{1}{2}\kappa |\nabla p|^2\, d\Omega, \tag{5.7.1}$$

subject to the essential boundary condition $p|_{\Gamma_1} = g(x)$.
Find the Euler-Lagrange equations for this problem. What are the natural boundary conditions on Γ_3. Derive also the interface conditions on Γ_7.
Hint: Split the integral into two parts. □

5.8 From PDE to minimization problem

5.8.1 Introduction

We have seen in Section 5.3 that if the solution of a minimization problem is smooth, it satisfies a partial differential equation. Furthermore, on those parts of the boundary where no essential boundary condition has been prescribed, natural boundary conditions result from the minimization problem. Before trying to solve these minimization problems numerically we ask ourselves the question: is it always possible to find a minimization problem corresponding to a (partial) differential equation? The answer to this question is no. Under certain conditions only, one can find an equivalent minimization problem. The key property will be symmetry, which will be defined later on in this section. Nevertheless a large class of important PDEs satisfies the requirements necessary to derive an equivalent minimization problem. In Chapter 7 we shall generalize the theory in such a way that the numerical techniques of Chapter 6 can be applied even for cases where no minimization problem can be found. For simplicity we shall restrict ourselves to linear problems only. However, we have seen in 5.1 that also non-linear PDEs may correspond to minimization problems.

In first instance we consider only homogeneous boundary conditions. The general case will be treated later.

5.8.2 Linear problems with homogeneous boundary conditions

Consider the linear PDE (5.3.8)-(5.3.10), but with homogeneous boundary conditions

$$-\operatorname{div} k\nabla u = f\,, \tag{5.8.1}$$

$$u|_{\Gamma_1} = 0\,, \tag{5.8.2}$$

$$k\frac{\partial u}{\partial n}\bigg|_{\Gamma_2} = 0\,. \tag{5.8.3}$$

The solution of (5.8.1)-(5.8.3) must be found in the vector space Σ:

$$\Sigma = \{u \text{ smooth} \mid \; u|_{\Gamma_1} = 0\,; \;\; \frac{\partial u}{\partial n}\bigg|_{\Gamma_2} = 0\}\,.$$

In general we shall write a linear PDE like (5.8.1) in the form

$$Lu = f\,. \tag{5.8.4}$$

Hence in (5.8.1) we have $Lu \equiv -\operatorname{div}(k\nabla u)$.
It can be shown that a minimization problem for (5.8.4) can be found if L satisfies the two following properties:

$$\text{symmetry (self adjointness)} \quad \int_\Omega uLv\,d\Omega = \int_\Omega vLu\,d\Omega\,, \quad \forall u, v \in \Sigma\,. \tag{5.8.5}$$

$$\text{positiveness} \quad \int_\Omega uLu\,d\Omega \geq 0\,,\ \forall u \in \Sigma\,. \tag{5.8.6}$$

In practice it turns out that the two properties are also necessary for the existence of a corresponding minimization problem.
Differential operators satisfying properties (5.8.5) and (5.8.6) are called *strongly elliptic*.

Before constructing a minimization problem we shall check if properties (5.8.5), (5.8.6) are satisfied by problem (5.8.1)-(5.8.3). To this end we multiply (5.8.1) by $v \in \Sigma$, integrate over Ω and apply the divergence theorem twice:

$$\int_\Omega -v(\operatorname{div} k\nabla u)\, d\Omega = \int_\Omega k\nabla u \cdot \nabla v\, d\Omega - \oint_\Gamma vk\frac{\partial u}{\partial n}\, d\Gamma\,. \tag{5.8.7}$$

Due to the boundary conditions the boundary integral vanishes.

$$\int_\Omega k\nabla u \cdot \nabla v\, d\Omega = -\int_\Omega u(\operatorname{div} k\nabla v)\, d\Omega + \oint_\Gamma uk\frac{\partial v}{\partial n}\, d\Gamma\,. \tag{5.8.8}$$

Again the boundary integral vanishes.
In fact (5.8.7)-(5.8.8) already demonstrate symmetry.

To prove positivity it is sufficient to substitute u for v in (5.8.7):

$$\int_\Omega -u(\operatorname{div} k\nabla u)\, d\Omega = \int_\Omega k\nabla u \cdot \nabla u\, d\Omega \geq 0\,, \text{ for } u \in \Sigma\,. \tag{5.8.9}$$

With properties (5.8.5) and (5.8.6) it is easy to prove the following theorem:

Theorem 5.8.1 *Let L be a linear, symmetric, positive differential operator defined over a space Σ and let*

$$Lu = f\,. \tag{5.8.10}$$

Then the solution u minimizes the functional

$$I(u) = \int_\Omega \{\frac{1}{2}uLu - uf\}\, d\Omega\,, \text{ over the space } \Sigma\,. \tag{5.8.11}$$

On the other hand if u minimizes (5.8.11) then u satisfies (5.8.10).

Proof
First suppose that u_0 is the solution of (5.8.10), hence $Lu_0 = f$.
Substituting this in (5.8.11) gives (using the symmetry of L)

$$\begin{aligned} I(u) &= \int_\Omega \{\frac{1}{2}uLu - uLu_0\}\, d\Omega = \\ &\int_\Omega \{\frac{1}{2}(u-u_0)L(u-u_0) - \frac{1}{2}u_0Lu_0\}\, d\Omega\,. \end{aligned} \tag{5.8.12}$$

Since $\int_u \frac{1}{2}u_0Lu_0\, d\Omega$ is fixed and L is positive we know that the minimum is reached if

$$\int_\Omega \frac{1}{2}(u-u_0)L(u-u_0)\, d\Omega = 0\,.$$

Hence u_0 minimizes (5.8.11). □

Exercise 5.8.1 *Show that the minimum of $I(u)$ over Σ satisfies (5.8.10). Use the standard Euler-Lagrange approach and symmetry of L.* □

If we apply this theorem to example (5.8.1)-(5.8.3), it immediately follows that the corresponding functional $I(u)$ is given by

$$\begin{aligned} I(u) &= \int_\Omega \{\frac{1}{2}u(-\text{div}\, k\nabla u) - uf\}\, d\Omega \\ &= \int_\Omega \{\frac{1}{2}k|\nabla u|^2 - uf\}\, d\Omega\,, \end{aligned}$$

and this is the same as (5.3.1) except for the boundary integral. Actually with respect to the minimization problem it is not necessary to satisfy the natural boundary condition and it is sufficient to consider the space

$$\Sigma = \{u \text{ smooth} \mid \ u|_{\Gamma_1} = 0\}\,.$$

REMARK:

The proof of this theorem is based upon the symmetry and the positivity of the differential operator. These properties are sufficient. In practice these properties are also necessary. As a consequence no equivalent minimization problem for the convection-diffusion equation equation can be found.

In this section we have restricted ourselves to homogeneous boundary conditions, because they are necessary for the symmetry property (5.8.5). Otherwise the boundary integral in (5.8.7) would not vanish. It is only a small extension to consider also non-homogeneous boundary conditions, as will be demonstrated in Section 5.8.3.

Exercise 5.8.2 *Show that the operator in the convection-diffusion equation*

$$-\text{div}\,(k\nabla c) + \mathbf{u}\cdot\nabla c = f$$

is not symmetric. □

5.8.3 Linear problems with non-homogeneous boundary conditions

Theorem 5.8.1 relates a PDE with an equivalent minimization problem. However, this theorem is only applicable for homogeneous boundary conditions. In case of non-homogeneous boundary conditions we have to adapt the theorem or make the boundary conditions homogeneous. The last solution is the most simple one.

Theorem 5.8.2 *Let $Lu = f$ with non-homogeneous boundary conditions. Suppose that there is a smooth function w satisfying the non-homogeneous boundary conditions. If this function does not exist, the original problem has no solution.*
Then u satisfies the minimization problem

$$I(u) = \frac{1}{2}\int_\Omega (u-w)(Lu+Lw)\, d\Omega - \int_\Omega fu\, d\Omega\,. \tag{5.8.13}$$

Proof
Consider

$$v = u - w. \tag{5.8.14}$$

Clearly v satisfies homogeneous boundary conditions and since

$$Lv = Lu - Lw = f - Lw, \tag{5.8.15}$$

Theorem 5.8.1 can be applied for v with right-hand side $f - Lw$.

So the corresponding minimization problem is:

$$\min_{v\in\Sigma} I(v) = \frac{1}{2}\int_\Omega vLv\, d\Omega - \int_\Omega v(f - Lw)\, d\Omega\,, \tag{5.8.16}$$

with Σ provided with homogeneous boundary conditions.
Substituting (5.8.14) in (5.8.16) it is easy to see that

$$\tilde{I}(u) = \frac{1}{2}\int_\Omega (u - w)(Lu + Lw)\, d\Omega - \int_\Omega fu\, d\Omega + \int_\Omega fw\, d\Omega\,. \tag{5.8.17}$$

Since w and f are given functions independent of the solution, the minimum of (5.8.17) does not change if we skip the last term and we end up with (5.8.13). □
Mark that in this case we do not have $\int_\Omega uLw\, d\Omega = \int_\Omega wLu\, d\Omega$ (why?). As a consequence Equation (5.8.13) can not be simplified furthermore.
It is of course necessary to remove w from this expression, since w is unknown. This will be done in the following example.

Theorem 5.8.3 *Consider a region Ω with boundary Γ. Γ consists of 3 parts Γ_1, Γ_2 and Γ_3 such that*

$$\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3\,.$$

Consider the differential equation

$$-\mathrm{div}\, k\nabla u = f \quad (k > 0) \text{ in } \Omega\,, \tag{5.8.18}$$

with boundary conditions

$$u = g_1 \quad on \quad \Gamma_1\,, \tag{5.8.19}$$

$$k\frac{\partial u}{\partial n} = g_2 \quad on \quad \Gamma_2\,, \tag{5.8.20}$$

$$cu + k\frac{\partial u}{\partial n} = g_3 \quad on \quad \Gamma_3 \quad (c > 0)\,. \tag{5.8.21}$$

Then u satisfies the minimization problem

$$\min_{u\in\Sigma} \int_\Omega \{\frac{1}{2}k|\nabla u|^2 - uf\}\, d\Omega - \int_{\Gamma_2} g_2 u\, d\Gamma + \int_{\Gamma_3} \{\frac{1}{2}cu^2 - g_3 u\}\, d\Gamma\,, \tag{5.8.22}$$

with $\Sigma : \{u \mid u|_{\Gamma_1} = g_1\}$.

Proof
The function w satisfies (5.8.19) to (5.8.21).
Substitution of these terms in (5.8.13) gives

$$\min_u I(u) = \frac{1}{2}\int_\Omega (u - w)(-\mathrm{div}\,(k\nabla(u + w))\, d\Omega - \int_\Omega fu\, d\Omega\,.$$

Gauss theorem applied to the first terms gives

$$\begin{aligned} I(u) \;=\; & \frac{1}{2}\int_\Omega \nabla(u - w)\cdot k\nabla(u + w)\, d\Omega - \int_\Omega fu\, d\Omega \\ & -\frac{1}{2}\int_\Gamma (u - w)k\nabla(u + w)\cdot \mathbf{n}\, d\Gamma\,. \end{aligned} \tag{5.8.23}$$

The first term can be written as

$$\frac{1}{2}\int_{\Omega} \nabla(u-w)\cdot k\nabla(u+w)\,d\Omega = \frac{1}{2}\int_{\Omega} \nabla u\cdot k\nabla u\,d\Omega - \frac{1}{2}\int_{\Omega} \nabla w\cdot k\nabla w\,d\Omega\,. \tag{5.8.24}$$

Since the last term does not depend on u it can be removed from the minimization problem, without effect on u.
The last term of (5.8.23) is split over the 3 boundaries Γ_1, Γ_2 and Γ_3.
On Γ_1 this term is equal to 0 because of $u - w = 0$.
On Γ_2 it can be written as:

$$-\frac{1}{2}\int_{\Gamma_2}\{uk\frac{\partial(u+w)}{\partial n} - wk\frac{\partial(u+w)}{\partial n}\}\,d\Gamma = -\int_{\Gamma_2}\{ug_2 - wg_2\}\,d\Gamma$$

and again the last term can be skipped from the minimization problem since it does not depend on u.
On Γ_3 as

$$-\int_{\Gamma_3}\{ug_3 - wg_3 - \frac{1}{2}cu^2 + \frac{1}{2}cw^2\}\,d\Gamma\,,$$

and now the second and fourth term can be removed (why ?).
So at last we arrive at the minimization problem (5.8.22).

So the Dirichlet boundary condition (5.8.19) is an essential boundary condition. □

5.8.4 Exercises

Exercise 5.8.3 *Find the equivalent minimization problem of the three-dimensional Poisson equation*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f(x,y,z) \quad in\ \Omega\,,$$

with boundary condition

$$u = g \quad on\ \Gamma\,.$$

□

Exercise 5.8.4 *Find the minimization problem corresponding to the differential equation*

$$-\frac{d}{dx}(p(x)\frac{du}{dx}) = f$$

with boundary condition

$$u(0) = 1\,, \quad u(1) + \frac{du}{dx}(1) = a\,.$$

□

Exercise 5.8.5 *Find the minimization problem corresponding to the system of s ordinary differential equations*

$$\sum_{k=1}^{s}\left[-\frac{d}{dx}(p_{jk}(x)\frac{du_k}{dx}) + q_{jk}(x)u_k\right] = f_j(x) \qquad j = 1,2,...,s\,,$$
$$a < x < b\,,$$

with boundary conditions $u_j(a) = u_j(b) = 0$ $(j = 1, 2, ..., s)$ and P a symmetric positive definite matrix with elements $p_{jk}(x)$ and Q a symmetric positive semi-definite matrix with elements $q_{jk}(x)$.
Use theorem 5.8.1 with u a vector instead of a scalar. □

Exercise 5.8.6 *Find the minimization problem corresponding to the differential equation*

$$\frac{d^4u}{dx^4} = f \,,$$

with boundary conditions

$$u(0) = \frac{du}{dx}(0) = 0 \,, \quad \frac{du}{dx}(1) = 0 \,, \quad \frac{d^3u}{dx^3}(1) = 1 \,.$$

□

5.9 Mathematical theory of minimization

Section 5.8 shows that linear PDEs which satisfy some extra properties like symmetry and positivity, are equivalent to a minimization problem. In the proof of Theorem 5.8.1 we needed a function space Σ satisfying some smoothness requirements. In this section we shall consider this theory from a more fundamental (mathematical) point of view. However, this will not be a complete and thorough mathematical treatment of the problem, since that is beyond the scope of this book.

Let us first introduce some notations.

In Section 5.8 we have used expressions like $\int_\Omega uf \, d\Omega$. So it is naturally to use the L^2 inner product

$$(u, v) = \int_\Omega uv \, d\Omega \,, \tag{5.9.1}$$

which is defined for all functions $u, v \in L^2(\Omega)$.

Besides that we have introduced the integral

$$\int_\Omega uLv \, d\Omega \tag{5.9.2}$$

in (5.8.5). For this integral we have required some properties like symmetry (5.8.5) and positivity (5.8.6).

Definition 5.9.1
The operator L is called positive definite if there exist a constant $\gamma > 0$, such that

$$\int_\Omega uLu \, d\Omega \geq \gamma \int_\Omega u^2 \, d\Omega \,, \quad \forall u \in \Sigma \,. \tag{5.9.3}$$

If (5.8.5), (5.8.6) and (5.9.3) are satisfied we can define a new inner product, the *energy product*, by:

$$(u, v)_L = \int_\Omega uLv \, d\Omega \tag{5.9.4}$$

and corresponding (energy) norm $\|u\|_L^2 = (u, u)_L$.

Exercise 5.9.1 *Prove that (5.9.4) satisfies all the requirements of an inner product.* □

It is necessary that the definition space Σ is such that the integral in (5.9.4) makes sense (i.e. is finite) and besides that, the space must be a vector space. This means that elements in Σ must satisfy the following linearity property

$$\begin{aligned} &\text{if} \quad u,\ v \in \Sigma \quad \text{then also} \\ &\alpha u + \beta v \in \Sigma \quad \text{with} \quad \alpha, \beta \in \mathbb{R}^1 \,. \end{aligned}$$

The space Σ is a space with smoothness requirements for its elements, but also each function in Σ must satisfy essential boundary conditions.

Exercise 5.9.2 *Show that Σ can be a vector space only if homogeneous boundary conditions are satisfied.* □

Now we can formulate Theorem 5.8.1 in a more mathematical way:

Let L be a linear operator defined on a Hilbert space Σ satisfying

$$(Lu, v) = (v, Lu) \quad \forall u, v \in \Sigma \text{ i.e. } L \text{ is self-adjoint .} \tag{5.9.5}$$

$$(u, Lu) \geq 0 \quad \forall u \in \Sigma \quad \text{i.e } L \text{ is positive .} \tag{5.9.6}$$

$$(u, Lu) \geq \gamma(u, u) \quad \forall u \in \Sigma \quad \text{i.e } L \text{ is positive definite .} \tag{5.9.7}$$

In fact (5.9.7) implies (5.9.6).
(u, u) is the L^2 inner product and (Lu, v) is the inner product in Σ.

Then the solution of

$$Lu = f \,, \quad u \in \Sigma \,, \quad f \in L^2(\Omega) \,, \tag{5.9.8}$$

minimizes the functional with $J(u)$

$$\min_{u \in \Sigma} J(u) = \frac{1}{2}(u, Lu) - (u, f) \,, \tag{5.9.9}$$

and the minimum of (5.9.9) satisfies (5.9.8).

Exercise 5.9.3 *Prove this theorem in the same way as in Section 5.8.* □

REMARK:
The property that L must be positive definite is not necessary in the theorem. It is only important in order to define an inner product. Also it enables us to prove uniqueness of the solution.

Theorem 5.9.1 *There is exactly one $u \in \Sigma$ that minimizes the functional $J(u)$ defined in (5.9.9).*

Proof

$$(u, f) \leq \|u\| \, \|f\| \,, \tag{5.9.10}$$

where $\|u\|$ is the L^2-norm. ($\|u\| = (u, u)^{1/2}$).

From (5.9.3) it follows that

$$\|u\|^2 \leq \frac{1}{\gamma} \|u\|_L^2 \,. \tag{5.9.11}$$

(5.9.10) together with (5.9.11) gives

$$(u, f) \leq \frac{1}{\sqrt{\gamma}} \|u\|_L \|f\| \,. \tag{5.9.12}$$

Now we apply Riesz' representation theorem (see [22]), This theorem states:

for every bounded linear functional $\ell(u)$ defined on a Hilbert space H there is exactly one element $u_0 \in H$ such that

$$\ell(u) = (u, u_0)_H \quad \forall u \in H\,,$$

with $(u, v)_H$ the inner product in H.

Hence there is exactly one $u_0 \in \Sigma$ such that

$$(u, f) = (u, u_0)_L \quad \forall u \in \Sigma\,. \tag{5.9.13}$$

Now consider the minimization problem:

$$\begin{aligned} J[u] &= \frac{1}{2}\|u\|_L^2 - (u, f) \\ &= \frac{1}{2}\|u\|_L^2 - (u, u_0)_L \\ &= \frac{1}{2}(u - u_0, u - u_0)_L - \frac{1}{2}(u_0, u_0)_L\,. \end{aligned} \tag{5.9.14}$$

Since $(v, v)_L > 0\ \forall v \in \Sigma$, $J[u]$ takes its minimum for $u = u_0$. Since u_0 is unique (from the Riesz' representation theorem) we have proven the theorem. □

So the minimization problem has always a unique solution in the Hilbert space Σ. However, this solution does not have to be the solution of the original PDE. The reason is that for second order PDEs, the inner product (u, Lv) only contains first order derivatives.
For elements in Σ it is sufficient that (u, Lu) is finite, hence the first derivative must be in $L^2(\Omega)$. For a second order PDE it is necessary that the second derivatives exist. So we may have a solution of the minimization problem in Σ that does not satisfy the PDE in classical sense.

According to Theorem 5.8.1 a solution of the minimization problem satisfies

$$Lu = f \quad u \in \Sigma\,. \tag{5.9.15}$$

But if u is not smooth we cannot consider this as a classical solution. So a solution of (5.9.9) is called a 'generalized' or 'weak' solution of the PDE. If the solution is also sufficiently smooth, it is called a 'strong' solution. In fact we have proved that there is always a weak solution. To prove that there is a strong solution we need extra smoothness requirements for both f and the boundary of Ω. Such a proof is general not simple. See for example [12]. But if a strong solution exists, it is equal to the weak solution. (Why?).

The Hilbert spaces introduced in this Chapter are usually *Sobolev spaces* denoted as $H^k(\Omega)$, where k refers to the highest order derivatives in the inner product. For example the space $H^1(\Omega)$ is defined in Theorem (1.6.1). More theory about Sobolev spaces can for example be found in [1].

5.10 Summary of Chapter 5

A number of physical problems can be formulated in the following form:

find a function $u(\mathbf{x})$ in a class of functions such that an integral of a function of $u(\mathbf{x})$ and some of its derivatives is minimal.

It has been demonstrated by variation, that the solution of the minimization problems satisfies a PDE, the *Euler-Lagrange equation*.

Some boundary conditions can be prescribed on the solution class of the minimization problems; these are called essential boundary conditions. Others arise naturally when the Euler-Lagrange equations are derived from the minimization problem. These boundary conditions are called natural. They always involve first order derivatives for second order equations and second and third order derivatives for fourth order equations.

On the other hand it has been shown that under certain conditions (symmetry and positiveness) an equivalent minimization problem can be derived from a PDE. The minimization problem has always lower order derivatives than the PDE. For example if the minimization problem contains first order derivatives the PDE is of second order and in case the minimization problem contains second order derivatives, the PDE is of order four. As a consequence the smoothness requirements for the solution of the PDE are more restrictive than those of the minimization problem.

If the PDE and minimization problem are equivalent, solution of one of the two automatically solves the other.

Chapter 6

The numerical solution of minimization problems

Objectives

Chapter 5 showed the equivalence between a certain class of PDEs and minimization problems. As a consequence solving the PDE also solves the minimization problem and vice versa. Chapters 3, 4 were devoted to solving the PDE directly by finite differences or, after integration over a volume, by finite volumes. In this chapter we shall solve the corresponding minimization problem numerically. Hence the PDE is solved in an indirect way.
The numerical technique that will be applied is the classical *Ritz's method* based on expressing the solution as a linear combination of previously chosen functions: the basis functions. These are in general not related to the problem, but chosen beforehand. This method itself is not very practical, but combined with a clever choice of basis functions we arrive at the finite element method (FEM). The FEM is well suited for unstructured grids and has a strict local character. All information in one element is used, without considering neighbors. This makes the method very attractive for computer implementation. For certain types of PDEs, for example those arising from elasticity and plasticity problems, the FEM is the most popular method at this moment.
Another way of looking at the FEM, is to consider it as an automatic tool to derive finite difference formula for unstructured grids. An important advantage of the FEM is that the treatment of boundary conditions is almost always very natural and therefore simpler than in classical difference methods.

6.1 Ritz's method

6.1.1 Introduction

Suppose we want to solve the general minimization problem

$$\min_u J[u] \; ; \quad J[u] = \int_\Omega F(x, y, u, u_x, u_y) \; d\Omega \; , \tag{6.1.1}$$

where the minimum must be found over a class of functions in the target space Σ:

$$\Sigma = \{u \text{ sufficiently smooth}; \;\; u|_\Gamma = g\} \; . \tag{6.1.2}$$

Chapter 5 already demonstrated that the solution of this problem is not simple. Actually we transformed it to a minimization problem with one unknown (ε), thus deriving the Euler-Lagrange equations.
Direct minimization of (6.1.1), (6.1.2) is in general only possible if we have a finite number of unknowns.

This can be achieved by approximating the solution by a linear combination of a finite fixed set of functions $\varphi_i(\mathbf{x})$:

$$u^n(\mathbf{x}) = \sum_{j=1}^{n} a_j \varphi_j(\mathbf{x}) \,. \tag{6.1.3}$$

In the remainder of this section we assume homogeneous essential boundary conditions, i.e. $g = 0$.

The functions $\varphi_i(\mathbf{x})$ (the *basis functions*), must be chosen such that:

$$\varphi_i(\mathbf{x}) \in \Sigma \quad \text{for all } i \,.$$

This means that $\varphi_i(\mathbf{x})$ must be sufficiently smooth, such that (6.1.1) makes sense, and also that $\varphi_i(\mathbf{x})$ must satisfy the homogeneous boundary conditions. So in fact the functions $\varphi_i(\mathbf{x})$ span a subspace of Σ. Moreover, the functions $\varphi_i(x)$ should preferably be linearly independent (why?). Now Ritz's method consists of solving the minimization problem over this subspace.

Since only the a_i in (6.1.3) are unknown, this means that the problem reduces to minimizing over the set $a_1 ... a_n$:

$$\min_{a_i \in \mathbb{R}^n} J[a_1, a_2, ..., a_n] \,. \tag{6.1.4}$$

The necessary condition for the existence of a minimum is

$$\frac{\partial J[u^n]}{\partial a_i} = 0 \,, \quad i = 1, 2, ..., n \,. \tag{6.1.5}$$

(6.1.5) forms a set of n equations with n unknowns, which under certain conditions, can be solved uniquely. This produces a solution $u^n(\mathbf{x})$. By increasing the number of basis functions we hope that $u^n(\mathbf{x})$ converges to the solution $u(\mathbf{x})$ of (6.1.1) and (6.1.2). It is clear that the choice of the basis functions $\varphi_i(\mathbf{x})$ is essential for the convergence and especially for the speed of convergence of Ritz's method.

Let us first consider a simple one-dimensional example to show how Ritz's method behaves in practice.

6.1.2 A simple one-dimensional example

Theorem 6.1.1 *Let u satisfy the following minimization problem (cf. Section 5.1.1)*

$$\min_{u \in \Sigma} J[u] = \int_0^1 \{ \frac{1}{2} \left(\frac{du}{dx} \right)^2 - f(x) u(x) \} \, dx \,, \tag{6.1.6}$$

$$\Sigma : \; \{ u \mid u \text{ sufficiently smooth; } u(0) = 0 \} \,.$$

and let $u^n(\mathbf{x})$ be defined by (6.1.3)*, then the set of Ritz equations is given by*

$$\sum_{j=1}^{n} a_j \int_0^1 \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx} \, dx = \int_0^1 f(x) \varphi_i(x) \, dx, \quad i = 1, 2, ..., n. \tag{6.1.7}$$

Proof
Substitution of (6.1.5) in (6.1.6), (6.1.5) gives

$$\frac{\partial}{\partial a_i}\int_0^1 \{\frac{1}{2}\left(\frac{d\sum_{j=1}^{n} a_j\varphi_j(x)}{dx}\right)^2 - f(x)(\sum_{j=1}^{n} a_j\varphi_j(x))\}\,dx = 0. \tag{6.1.8}$$

Hence a_i satisfies (6.1.7) □

Exercise 6.1.1 *Verify Equation (6.1.7)* □

Exercise 6.1.2 *Show that the solution of (6.1.6) satisfies the DE*

$$-\frac{d^2u}{dx_2} = f(x)\,, \tag{6.1.9}$$

with boundary conditions

$$u(0) = 0\,, \quad \frac{du}{dx}(1) = 0\,, \tag{6.1.10}$$

provided the solution of (6.1.6) is twice differentiable. □

The system of equations (6.1.7) is uniquely solvable if the coefficient matrix S is non-singular. This system can be written in matrix-vector notation by

$$S\mathbf{a} = \mathbf{f}\,, \tag{6.1.11}$$

with S an $(n \times n)$ matrix with elements $s_{ij} = \int_0^1 \frac{d\varphi_i}{dx}\frac{d\varphi_j}{dx}\,dx$,

a an $(n \times 1)$ vector with elements a_j,

f an $(n \times 1)$ vector with elements $f_i = \int_0^1 f(x)\varphi_i(x)\,dx$.

There are many possible choices for the basis functions $\varphi_i(\mathbf{x})$, but we shall restrict ourselves to 2 specific ones.

Theorem 6.1.2 *Let the basis functions $\varphi_i(\mathbf{x})$ be given by*

$$\varphi_k(x) = \sin k\pi x \tag{6.1.12}$$

then the matrix **S** *in* (6.1.11) *has elements*

$$S_{kk} = \frac{k^2\pi^2}{2}\,. \tag{6.1.13}$$

and the solution a_k satisfies

$$a_k = \frac{2}{k^2\pi^2}\int_0^1 f(x)\sin(k\pi x)\,dx\,. \tag{6.1.14}$$

□

The basis functions $\varphi_k(x)$ are elements of Σ, since they are analytical functions and satisfy $\varphi_k(0) = 0$. Note that none of them satisfies the natural boundary condition. Using the orthogonality relations of the cosine we see that the basis functions $\varphi_k(x)$ produce a diagonal matrix S, with diagonal elements (6.1.13)

Exercise 6.1.3 *Prove that Equation (6.1.14) is the result of substituting the basis functions (6.1.12) into (6.1.7)* □

Exercise 6.1.4 *Show that the set a_k defined by (6.1.14) form the coefficients of the Fourier expansion of the exact solution $u(x)$ using functions* $\sin(k\pi x)$.
Hint: substitute 6.1.3 into 6.1.9. □

Theorem 6.1.3 *Let the basis functions $\varphi_i(\mathbf{x})$ be given by*

$$\varphi_k(x) = x^k \tag{6.1.15}$$

then the matrix $\mathbf{S}$ *in* (6.1.11) *is given by*

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & \frac{4}{3} & \frac{6}{4} & \frac{8}{5} & \frac{10}{6} \\ 1 & \frac{6}{4} & \frac{9}{5} & \frac{12}{6} & \cdots \\ 1 & \frac{8}{5} & \frac{12}{6} & \frac{16}{7} & \cdots \\ 1 & \frac{10}{6} & \cdots & \cdots & \cdots \\ \vdots & & & & \end{bmatrix} . \tag{6.1.16}$$

□

Exercise 6.1.5 *Derive Equation (6.1.16).* □

Matrix S in (6.1.16) is a *Hilbert matrix*. Not only is this matrix full, it is also very badly conditioned. Although the matrix is non-singular, numerically it is not invertible for relatively small values of n (order 10 to 20), on a 16 digits computer.

From these two specific choices for the basis functions we can draw some conclusions with respect to requirements for the basis functions.

6.1.3 Some observations concerning the basis functions

- With respect to the basis function $\varphi_k(x)$ defined in (6.1.12) it is clear that the solution $u^n(x)$ converges to the minimization problem, because the Fourier series is convergent (Exercise 6.1.4).
  One can also prove convergence in case of basis functions $\varphi_k(x)$ in (6.1.15), provided the system of linear equations can be solved.
- Even though the basis functions themselves do not satisfy the natural boundary condition, in the limit the linear combination does in some way, if there is convergence to the exact solution. In practice $\frac{du^n}{dx}(x)$ will be small in some sense, for n large enough.
- We have seen that with the specific choice (6.1.12) of the basis functions, the coefficient matrix is diagonal, and therefore the solution of the system of equations is trivial.
  This is not a coincidence: these functions form the eigenfunctions of the continuous eigenvalue problem

$$-\frac{d^2u}{dx_2} = \lambda u \; ; \; u(0) = 0 \, , \; u(1) = 0 \, . \tag{6.1.17}$$

  These eigenfunctions are orthogonal with respect to the inner product

$$\int_0^1 \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx} \, dx,$$

which implies that these inner products vanish if $i \neq j$. Also for more general problems one can define such an inner product and again the eigenfunctions have the same property. Unfortunately in practice it is almost impossible to find an analytical expression for the eigenfunctions. Numerical computation of the eigenfunctions is in general a harder task than solving the system of equations (6.1.11).

- On the other hand choosing an arbitrary set of basis functions leads to a full matrix. Unless the number of basis functions is very small, solution of such a system is very expensive.
  In finite difference methods and finite volume methods we always arrived at systems of equations with a sparse structure. If we want a sparse matrix in Ritz's method, it is necessary that most of the integrals

$$\int_0^1 \frac{d\varphi_i}{dx}\frac{d\varphi_j}{dx}\,dx$$

  vanish. So the majority of the basis functions must be orthogonal with respect to the inner product defined by these integrals. We shall call such a set "nearly orthogonal".

- It is obvious that in the limit, the set of basis functions must span the complete space Σ, otherwise there are elements in Σ that can not be represented as linear combination of basis functions. Besides that, it would be nice if arbitrary functions in Σ could be approximated accurately with a small number of basis functions. The basis functions $\varphi_k(x)$ in (6.1.12) do not satisfy this property.

Combining all this we come to the following requirements for our set of basis functions:

1 the basis functions must be linearly independent

2 the basis functions must span the complete space Σ

3 the basis functions should be "nearly orthogonal"

4 arbitrary functions in Σ must be approximated accurately by a limited number of basis functions

At first sight it seems very difficult to satisfy all these demands. However, in Section 6.2 we shall show how to construct such basis functions by the finite element method.

We have treated lightly over the convergence of Ritz's method for a good reason: this is very hard to prove in general. For a specific case of practical importance we provide a proof: strongly elliptic operators (see Section 5.8.2).

6.1.4 Mathematical theory: convergence of Ritz's method

We consider the convergence of Ritz's method for the specific case of a linear operator satisfying properties (5.9.5) – (5.9.7). In order to do that we need a few tools. First we recall the definition of a basis for a Hilbert space.

Definition 6.1.1 *A family* $\{\varphi_\alpha\} \in \Sigma$ *is called a basis for the Hilbert space* Σ, *if the following two properties are satisfied.*

1. *Linear independence*

$$\sum_{i=1}^{N} \beta_i \varphi_i = 0 \ \textit{implies} \ \beta_i = 0\,, \ \ i = 1,..,N.$$

2. *Completeness*
For every $u \in \Sigma$ and a given $\varepsilon > 0$, there is a finite linear combination of basis functions such that the distance between u and this combination is smaller than ε. In formula:
$\forall \varepsilon > 0 \; \exists \; \{\varphi_{\alpha_1}, \varphi_{\alpha_2}, ..., \varphi_{\alpha_N}\}$ and $\{\beta_1, \beta_2, ..., \beta_N\}$, $N < \infty$, such that

$$\|u - \sum_{i=1}^{N} \beta_i \varphi_{\alpha_i}\|_\Sigma < \varepsilon\,,$$

in which $\| \cdot \|_\Sigma$ is the norm in Σ.

□

Theorem 6.1.4 *The Ritz equations with approximate solution* (6.1.3) *to solve the minimization problem (5.9.9):*

$$\min_{u \in \Sigma} J[u]\,, \ \ \textit{with} \ \ J[u] = \frac{1}{2}\|u\|_L^2 - (u, f)\,, \tag{6.1.18}$$

are given by

$$\sum_{j=1}^{n} a_j(\varphi_i, \varphi_j)_L = (f, \varphi_i) \quad i = 1, ..., n\,. \tag{6.1.19}$$

Exercise 6.1.6 *Prove Equation (6.1.19)* □

The system of linear equations (6.1.19) has a unique solution if and only if the coefficient matrix S defined by

$$S = \begin{bmatrix} (\varphi_1, \varphi_1)_L & (\varphi_2, \varphi_1)_L & \cdots & (\varphi_n, \varphi_1)_L \\ (\varphi_1, \varphi_2)_L & (\varphi_2, \varphi_2)_L & \cdots & (\varphi_n, \varphi_2)_L \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ (\varphi_1, \varphi_n)_L & (\varphi_2, \varphi_n)_L & \cdots & (\varphi_n, \varphi_n)_L \end{bmatrix}. \tag{6.1.20}$$

is non-singular.
S is a *Gramm matrix* for the set of functions $\varphi_1, \varphi_2, ..., \varphi_n$ in the space Σ. In the following we assume that $\{\varphi_i\}$ is a basis for Σ.

Theorem 6.1.5 *S defined by (6.1.20) is not singular.*

Proof
Suppose there is a non-zero vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n)$ such that $S\boldsymbol{\alpha} = \mathbf{0}$.
Then $(\boldsymbol{\alpha}, S\boldsymbol{\alpha}) = 0$ so

$$\sum_i \sum_j \alpha_i \alpha_j (\varphi_i, \varphi_j)_L = 0\,.$$

Since the inner product is bilinear this implies

$$(\sum_i \alpha_i \varphi_i, \sum_j \alpha_j \varphi_j)_L = 0 \ \text{ or}$$
$$\|\sum_i \alpha \varphi_i\|_L = 0 \ \text{ and because } ||.||_L \text{ is a norm } \sum_i \alpha \varphi_i = 0\,.$$

By the linear independence of the basis functions, this implies $\alpha_i = 0$. So $S\boldsymbol{\alpha} = 0$ implies $\boldsymbol{\alpha} = 0$ and S is non-singular. □

Theorem 6.1.6 *If $\{\varphi_i\}$ is a basis for Σ, then approximation (6.1.3) converges to the solution u_0 of the minimization problem (6.1.18), .*

Proof
According to (5.9.14), J[u] can be written as

$$J[u] = \frac{1}{2}\|u - u_0\|_L^2 - \frac{1}{2}\|u_0\|_L^2 \,, \tag{6.1.21}$$

where $u_0 \in \Sigma$ minimizes $J[u]$ over Σ. This is a continuous function of the energy norm (why?), that is

$$\begin{aligned} &\forall \varepsilon > 0 \quad \exists\, \delta > 0 \text{ such that} \\ &\|u - u_0\|_L < \delta \Rightarrow |J[u] - J[u_0]| < \varepsilon \,. \end{aligned} \tag{6.1.22}$$

Let $u_0^n \in \Sigma^n = \text{Span}\,\{\varphi_j\}_{j=1}^n$ minimize $J[u]$ over Σ^n, then

$$J[u_0] \leq J[u_0^n] \leq J[u^n], \;\; \forall u^n \in \Sigma^n. \tag{6.1.23}$$

We choose u^n using completeness:

$$\begin{aligned} &\exists N \geq 1, \alpha_1, \ldots, \alpha_N \text{ such that} \\ &\|u_0 - u^n\|_L < \delta, \text{ with } u^n = \sum_{j=1}^{n} \alpha_j \varphi_j, \forall n \geq N. \end{aligned} \tag{6.1.24}$$

Continuity of $J[u]$ gives $|J[u^n] - J[u_0]| < \varepsilon$. Note that $\varepsilon > 0$ is arbitrary, for which $\delta > 0$ and $N \geq 1$ exist, and hence $J[u^n] \to J[u_0]$ as $n \to \infty$.
The Squeeze Theorem is applied to (6.1.23) to conclude that

$$J[u_0^n] \to J[u_0] \text{ as } n \to \infty. \tag{6.1.25}$$

Equation (6.1.21) finally implies $\|u_0^n - u_0\| \to 0$ as $n \to \infty$. □

6.2 The finite element method in $\mathbb{R}^1$

6.2.1 Introduction

Ritz's method can be used to solve the minimization problem and therefore also the corresponding PDE. The main issue in Ritz' method is the choice of the basis functions. In Section 6.1.3 we have formulated a number of properties the basis functions should satisfy, in order to get an attractive solution method.
We derive a construction technique that creates basis functions satisfying all these properties. The key to this method is the subdivision of the region Ω into subparts (elements) and an element-wise polynomial approximation of the unknown function.

First we demonstrate this construction in $\mathbb{R}^1$, subsequently it will be extended to $\mathbb{R}^2$.

6.2.2 The Poisson equation in $\mathbb{R}^1$

As first example we consider Poisson's equation in one dimension:

$$\begin{aligned} -\frac{d^2u}{dx_2} &= f(x) \,, \\ u(0) &= 0 \,, \\ \frac{du}{dx}(1) &= 0 \,. \end{aligned} \tag{6.2.1}$$

Exercise 6.2.1 *Show that the solution of (6.2.1) satisfies the minimization problem*

$$\min_{u\in\Sigma} J[u]\ ;\ \ J[u] = \int_0^1 \{\frac{1}{2}(\frac{du}{dx})^2 - u(x)f(x)\}\ dx\ . \tag{6.2.2}$$
$$\Sigma : \{u \text{ sufficiently smooth };\ u(0) = 0\}$$

□

The smoothness requirement implies that the integral in (6.2.2) makes sense.

The system of Ritz equations is given by (6.1.7).
In order to construct the basis functions we subdivide the interval [0, 1] into subintervals $e_k = [x_{k-1}, x_k]$ the *elements*, as shown in Figure 6.1.

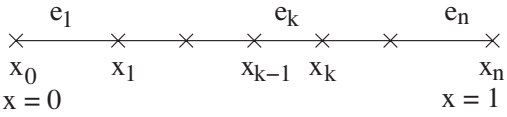


Figure 6.1: Subdivision of the interval [0, 1] in elements.

The solution u is approximated by a piecewise (lower order) polynomial defined element-wise. The most simple approximation is piecewise linear per element. Figure 6.2 shows a typical approximation $\tilde{u}$ of a function u by a piecewise linear polynomial. Note that the boundary condition $u(0) = 0$ is already satisfied by $\tilde{u}(0) = 0$.

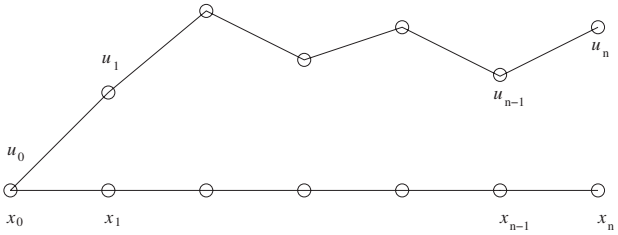


Figure 6.2: Approximation of $u(x)$ by $\tilde{u}(x)$.

Exercise 6.2.2 *Let $\tilde{u}$ be a piecewise linear approximation of u. Then $\tilde{u}$ does not belong to $C^1(0,1)$. Let $f(x)$ be a continuous function.*
Show that the integral in (6.2.2) remains finite when u is replaced by $\tilde{u}$. □

The linear interpolation polynomial of the function $u(x)$ over the element e_k is defined by

$$u_k(x) = \frac{x - x_k}{x_{k-1} - x_k} u(x_{k-1}) + \frac{x - x_{k-1}}{x_k - x_{k-1}} u(x_k)\ . \tag{6.2.3}$$

Exercise 6.2.3 *Show that Formula (6.2.3) is indeed the linear interpolation polynomial.*
□

Formally speaking it is not correct to use $u(x_k)$ since $u(x)$ is unknown. It would be better to use $\tilde{u}(x_k)$. However, as long as there is no confusion possible, we will omit the tilde.

We define linear *Lagrangian polynomials* $l_k(x)$

$$\ell_{k-1}(x) = \frac{x - x_k}{x_{k-1} - x_k}\,; \quad \ell_k(x) = \frac{x - x_{k-1}}{x_k - x_{k-1}}\,, \tag{6.2.4}$$

and write (6.2.3) as

$$u_k(x) = \ell_{k-1}(x)u_{k-1} + \ell_k(x)u_k\,. \tag{6.2.5}$$

u_k denotes $u(x_k)$.

Clearly $\ell_{k-1}(x)$ and $\ell_k(x)$ are linear on e_k and are defined by the relations

$$\ell_j(x_i) = \delta_{ij}\,; \quad i, j = k-1, k\,. \tag{6.2.6}$$

δ_{ij} is the Kronecker delta, defined by

$$\delta_{ij} = 0 \quad if \quad i \neq j \tag{6.2.7}$$

$$\delta_{ij} = 1 \quad if \quad i = j. \tag{6.2.8}$$

These relations define $\ell_j(x)$ uniquely (why?).

From (6.2.5) it is clear that $\tilde{u}(x)$ is a linear function of $u_0, u_1, ..., u_n$ so that we can write

$$\tilde{u}(x) = \sum_{j=0}^{n} u_j \varphi_j(x)\,. \tag{6.2.9}$$

The function $\varphi_i(x)$ consist of piecewise linear *Lagrangian polynomials* and may be considered as a generalized Lagrangian polynomials defined over the whole region Ω.
A typical $\varphi_i(x)$ has been sketched in Figure 6.3.

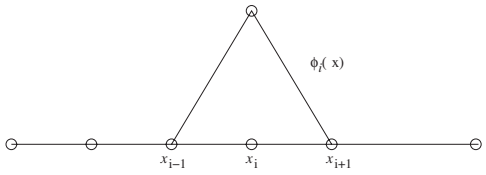


Figure 6.3: Example of a typical generalized Lagrangian polynomial.

φ_i is found taking all coefficients $u_k = 0$ $(i \neq k)$ and $u_i = 1$.

Exercise 6.2.4 *Sketch the basis functions $\varphi_0(x)$ and $\varphi_n(x)$.* □

Note that $\varphi_i(x)$ is only non-zero in the elements that contain the node x_i.
It is immediately clear that $\varphi_i(x)$ is defined by the following rules:

$$\begin{array}{ll} \text{a.} & \varphi_i(x) \text{ is linear in each element.} \\ \text{b.} & \varphi_i(x_j) = \delta_{ij}. \end{array} \tag{6.2.10}$$

Since $u_0 = 0$, (6.2.9) can be written as

$$\tilde{u}(x) = \sum_{j=1}^{n} u_j \varphi_j(x)\,. \tag{6.2.11}$$

The basis function $\varphi_0(x)$ will be used for non-homogeneous boundary conditions (see Section 6.2.4).

Theorem 6.2.1 *Suppose that an equidistant grid is used* $(x_{i+1} - x_i = h)$. *The system of Ritz equations* (6.1.7) *with the basis functions defined by (6.2.10) leads to the following system of equations:*

$$\frac{1}{h}\begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & \bigcirc & \\ & -1 & 2 & -1 & & & \\ & & & \ddots & \ddots & & \\ & \bigcirc & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \\ \vdots \\ \\ u_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \\ \vdots \\ \\ f_n \end{bmatrix}, \tag{6.2.12}$$

with $f_i = \int_0^1 f(x)\varphi_i(x)\,dx$.

Exercise 6.2.5 *Prove Theorem 6.2.1.* □

Exercise 6.2.6 *Compare system (6.2.12) with the system obtained by the FDM.* □

6.2.3 Numerical integration

The right-hand-side vector of (6.2.12) contains an integral over a function $f(x)$. In general one can not compute such an integral analytically, so a numerical approximation is required. Since we are integrating over each element separately an obvious choice is to use a numerical rule based on the same element.

Well-known integration rules are for example

mid-point rule:
$$\int_{x_{k-1}}^{x_k} g(x)\,dx \approx (x_k - x_{k-1})g(x_{k-1/2})\,, \tag{6.2.13}$$

trapezoid rule:
$$\int_{x_{k-1}}^{x_k} g(x)\,dx \approx \frac{x_k - x_{k-1}}{2}\{g(x_{k-1}) + g(x_k)\}\,, \tag{6.2.14}$$

Simpson's rule:
$$\int_{x_{k-1}}^{x_k} g(x)\,dx \approx \frac{x_k - x_{k-1}}{6}\{g(x_{k-1}) + 4g(x_{k-1/2}) + g(x_k)\}\,. \tag{6.2.15}$$

All these rules can be written in the general form:

$$\int_{x_{k-1}}^{x_k} g(x)\,dx \approx \sum_{k=1}^{r} w_k g(v_k)\,, \tag{6.2.16}$$

with r the number of quadrature points,
w_k the weights, and
v_k quadrature points.

Exercise 6.2.7 *Give* r, w_k *and* v_k *for the midpoint rule, the trapezoid rule and Simpson's rule.* □

Another class of integration rules of the shape (6.2.16) are the *Gaussian rules*. These methods are characterized by the fact that integration points and weights are chosen such that the highest order of accuracy is reached with a particular number of

integration points. Weights and integration points of Gaussian integration rules can be found in various text books, like for example [50] and [37].

f_i in (6.2.12) is defined as

$$\int_0^1 f(x)\varphi_i(x)\,dx = \int_{x_{i-1}}^{x_i} f(x)\varphi_i(x)\,dx + \int_{x_i}^{x_{i+1}} f(x)\varphi_i(x)\,dx\,. \tag{6.2.17}$$

The integrand $g(x)$ in (6.2.17) is defined by $f(x)\varphi_i(x)$. We could use every possible integration rule of type (6.2.16) to compute (6.2.17).
We consider integration over the element $[x_{k-1}, x_k]$. In the Finite Element Method, one represents the numerical solution in terms of a linear combination of basis functions. For the case of linear basis functions, one approximates the function $g(x)$ by linear interpolation, that is

$$g(x) \approx g(x_{k-1})\varphi_{k-1}(x) + g(x_k)\varphi_k(x), \tag{6.2.18}$$

over the interval $[x_{k-1}, x_k]$. Subsequently, integration over $[x_{k-1}, x_k]$ gives

$$\begin{aligned} \int_{x_k}^{x_{k-1}} g(x)\,dx &\approx \int_{x_k}^{x_{k-1}} g(x_{k-1})\varphi_{k-1}(x) + g(x_k)\varphi_k(x)\,dx & (6.2.19)\\ &= g(x_{k-1})\int_{x_k}^{x_{k-1}} \varphi_{k-1}(x)\,dx + g(x_k)\int_{x_k}^{x_{k-1}} \varphi_k(x)\,dx. & (6.2.20) \end{aligned}$$

Using linearity of the basis functions and the relation $\varphi_i(x_j) = \delta_{ij}$, we get

$$\int_{x_k}^{x_{k-1}} g(x)\,dx \approx \frac{x_k - x_{k-1}}{2}(g(x_{k-1}) + g(x_k)). \tag{6.2.21}$$

Similar rules are derived for higher order basis functions with more quadrature points. Imagine that one integrates over interval $[x_{k-l}, x_{k+m}]$, $l, m \geq 0$, $l \cdot m \neq 0$. Let this interval contain nodes $x_{k-l}, x_{k-l+1}, \ldots, x_{k+m}$, then using the basis functions $\varphi_{k-l}, \varphi_{k-l+1}, \ldots, \varphi_{k+m}$ one can write the following interpolating approximation for $g(x)$

$$g(x) \approx \sum_{p=k-l}^{k+m} g(x_p)\varphi_p(x). \tag{6.2.22}$$

This interpolation is substituted into the integral over $g(x)$

$$\int_{x_{k-l}}^{x_{k+m}} g(x)\,dx \approx \sum_{p=k-l}^{k+m} g(x_p)\int_{x_{k-l}}^{x_{k+m}} \varphi_p(x)\,dx. \tag{6.2.23}$$

This type of quadrature based on interpolation on the FEM basis functions is called *Newton-Cotes rule*.

Theorem 6.2.2 *The Newton-Cotes rule applied to (6.2.12), the right-hand side vector can be written as*

$$h \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ f(x_n)/2 \end{bmatrix}. \tag{6.2.24}$$

Exercise 6.2.8 *Prove Theorem 6.2.2.* □

Note that the Newton-Cotes rule applied for a linear approximation in $\mathbb{R}^1$ is identical to the trapezoid rule. If a quadratic approximation is used this rule is identical to Simpson's rule.

Not only the type of interpolation, also the type of integration rule influences the accuracy of the solution. This subject will be considered in Section 8.7.

6.2.4 Boundary conditions

In our example (6.2.1) we have seen how homogeneous boundary conditions had to be treated. In summary:

- homogeneous natural boundary conditions pose no problem at all. They are an implicit part of the minimization problem.
- homogeneous essential boundary conditions fix the parameters on the boundary. The corresponding interpolation functions are not used as basis functions. In this way all basis functions satisfy the essential boundary conditions.

Non-homogeneous boundary conditions require only a small adaptation. We shall demonstrate this by extending example (6.2.1) with non-homogeneous boundary conditions:

$$\begin{aligned} -\frac{d^2u}{dx_2} &= f(x)\,, \\ u(0) &= a\,, \\ \frac{du}{dx}(1) &= b\,. \end{aligned} \tag{6.2.25}$$

The solution of (6.2.25) satisfies the minimization problem

$$\min_{u\in\Sigma} J[u] = \int_0^1 \{\frac{1}{2}(\frac{du}{dx})^2 - u(x)f(x)\}\,dx - bu(1)\,, \tag{6.2.26}$$
$$\Sigma : \{u \mid u \text{ sufficiently smooth; } \mathrm{u}(0) = \mathrm{a}\}$$

Exercise 6.2.9 *Show that the solution of (6.2.25) satisfies the minimization problem (6.2.26).* □

In order to apply Ritz's method we define

$$\tilde{u}(x) = \sum_{j=0}^{n} u_j\varphi_j(x) = \sum_{j=1}^{n} u_j\varphi_j(x) + u_0\varphi_0(x) \tag{6.2.27}$$

Again we use the linear Lagrangian polynomials $\ell_i(x)$ as basis functions, so $\varphi_i(x)$ is defined by (6.2.10). Now it is clear that $u_0 = a$ (why?).
If we use the approximation (6.2.27), the Ritz equations corresponding to (6.2.26) are equal to

$$\sum_{j=1}^{n} u_j \int_0^1 \frac{d\varphi_i}{dx}\frac{d\varphi_j}{dx}\,dx = \int_0^1 f\varphi_i\,dx - u_0\int_0^1 \frac{d\varphi_i}{dx}\frac{d\varphi_0}{dx}\,dx + b\varphi_i(1) \tag{6.2.28}$$
$$i = 1, 2, ..., n\,.$$

$$\frac{1}{h}\begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & & & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 1 \end{bmatrix}\begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix} = h\begin{bmatrix} f_0/2 \\ f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n/2 + b/h \end{bmatrix}$$

Figure 6.4: System of equation before applying essential boundary conditions.

Exercise 6.2.10

a. Derive (6.2.28).

b. Why is $i = 0$ not part of (6.2.28)?

c. Which of the functions $\varphi_i(x)$ is non-zero in $x = 1$?

□

From Formula 6.2.28 it will be clear that the non-homogeneous essential boundary condition gives a contribution to the right-hand side. To compute this contribution we first build the matrix and right-hand side as if there are no essential boundary conditions (see Figure 6.4). Following Exercise 6.2.10 row 1 (corresponding to $\varphi_i = \varphi_0$), must be removed. Since the matrix must be square also column 1 must be removed. This is done by multiplying this column by the given value u_0 and subtracting it from the right-hand side as sketched in Figure 6.5.

$$\frac{1}{h}\begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & 1 & 2 & -1 & & & \\ & & \ddots & & & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 1 \end{bmatrix}\begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix} = h\begin{bmatrix} f_0/2 \\ f_1 \\ f2 \\ \vdots \\ f_{n-1} \\ f_n/2 + b/h \end{bmatrix} - \begin{bmatrix} u_0/h \\ -u_0/h \\ 0 \\ \vdots \\ \cdot \\ 0 \end{bmatrix}$$

Figure 6.5: Remove row 1. Multiply column 1 by u_0 and put it into the right-hand side.

The result of this operation is in Figure 6.6. The inhomogeneous natural boundary

$$\frac{1}{h}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & & & & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix} = h\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n/2 + b/h \end{bmatrix} - \frac{u_0}{h}\begin{bmatrix} -1 \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix}$$

Figure 6.6: System of equations after application of essential boundary conditions.

condition also contributes to the right-hand side. This contribution is an immediate consequence of the minimization problem.

6.2.5 Element matrices and element vectors

In order to construct the matrix in (6.2.12) it was necessary to evaluate the integrals in (6.1.11). Since $\varphi_i(x)$ is defined in an element-wise manner, the natural way to do this is by an element-wise way.

$$\int_0^1 \frac{d\varphi_i}{dx}\frac{d\varphi_j}{dx}\,dx = \sum_{k=1}^{n} \int_{e_k} \frac{d\varphi_i}{dx}\frac{d\varphi_j}{dx}\,dx\,. \tag{6.2.29}$$

So instead of computing the left-hand side for all i and j, one might first compute all integrals

$$\int_{e_k} \frac{d\varphi_i}{dx}\frac{d\varphi_j}{dx}\,dx\,, \tag{6.2.30}$$

for all i and j and add these integrals afterwards to get (6.2.29). This seems a very complicated way to compute the integrals. However at most 4 of the integrals in (6.2.30) are different from zero (Why?). We store these four integrals in a small matrix, the *element matrix*:

$$S^{e_k} = \begin{bmatrix} \int_{e_k} \frac{d\varphi_{k-1}}{dx}\frac{d\varphi_{k-1}}{dx}\,dx & \int_{e_k} \frac{d\varphi_{k-1}}{dx}\frac{d\varphi_k}{dx}\,dx \\ \int_{e_k} \frac{d\varphi_k}{dx}\frac{d\varphi_{k-1}}{dx}\,dx & \int_{e_k} \frac{d\varphi_k}{dx}\frac{d\varphi_k}{dx}\,dx \end{bmatrix}. \tag{6.2.31}$$

In the same way we create the *element vector*:

$$\mathbf{f}^{e_k} = \begin{bmatrix} \int_{e_k} f(x)\varphi_{k-1}(x)\,dx \\ \int_{e_k} f(x)\varphi_k(x)\,dx \end{bmatrix}. \tag{6.2.32}$$

Once all element matrices and vectors are computed, it is a matter of addition to compute the large matrix S and the large right-hand side $\mathbf{F}$. The main advantage of this approach is that all information of the minimization problem, the type of approximation in the element as well as the numerical integration rule applied, is stored locally.
To create the large matrix it is sufficient to know which unknowns are present in the element and to which entries the entries of the element matrix must be added. This is called the *topology* of the problem. The same holds for the large vector on the right-hand side.

This is a big advantage of the FEM. Once the region is subdivided into elements, it is sufficient to give a generic algorithm for the contributions of an arbitrary element. There is no need to worry about neighboring elements. Especially for more-dimensional unstructured grids, this is very attractive.

6.2.6 Assembly of the large matrix and vector

We have seen that all information for the FEM is stored in element matrices, element vectors and problem topology. The question is now: how can we construct the large matrix and vector from this information. The process of creating the large matrix and vector is known as assembly. To demonstrate this process we reuse minimization problem (6.2.2).
We consider the subdivision of the region [0, 1] into 4 elements as shown in Figure 6.7. Element e_i is defined by $e_i = [x_{i-1}, x_i]$ and the nodes are numbered from 0

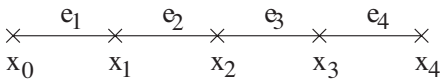


Figure 6.7: Subdivision of [0, 1] into 4 elements, and corresponding numbering of nodes and elements.

to 4. The unknowns have in this special case exactly the same numbering, where we know that $u_0 = 0$, so that the real unknowns are numbered from 1 to 4. In first instance the large matrix has size (5×5) and the large vector (5×1). The actual essential boundary condition is eliminated afterwards. The problem topology of this case is very simple; each element contains two unknowns.

$$\begin{array}{ll} e_1: & (0,1)\,, \\ e_2: & (1,2)\,, \\ e_3: & (2,3)\,, \\ e_4: & (3,4)\,. \end{array} \tag{6.2.33}$$

In the first step the large matrix and vector are cleared:

$$S^0 = \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \end{array} \\ \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{array}{l} \leftarrow 1 \\ \leftarrow 2 \\ \leftarrow 3 \\ \leftarrow 4 \\ \leftarrow 5 \end{array} \end{array} \quad \mathbf{f}^0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6.2.34}$$

The element matrix for an arbitrary element e_k has shape

$$S^{e_k} = \frac{1}{x_k - x_{k-1}} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{6.2.35}$$

We apply the Newton-Cotes rule, to obtain the element vector

$$\mathbf{f}^{e_k} = \frac{x_k - x_{k-1}}{2} \begin{bmatrix} f(x_{k-1}) \\ f(x_k) \end{bmatrix}. \tag{6.2.36}$$

For the sake of simplicity we assume an equidistant grid with step size $x_k - x_{k-1} = h$.

Adding the first element matrix and right-hand side to (6.2.34) gives

$$S^1 = \frac{1}{h} \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{f}^1 = \frac{h}{2} \begin{bmatrix} f(x_0) \\ f(x_1) \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6.2.37}$$

Next we add S^{e_2} and f^{e_2} to S^1 and f^1:

$$S^2 = \frac{1}{h} \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{f}^2 = \frac{h}{2} \begin{bmatrix} f(x_0) \\ 2f(x_1) \\ f(x_2) \\ 0 \\ 0 \end{bmatrix}. \tag{6.2.38}$$

Repeating this process for e_3 and e_4 gives:

$$S = S^4 = \frac{1}{h}\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}, \quad \mathbf{f} = \mathbf{f}^4 = \frac{h}{2}\begin{bmatrix} f(x_0) \\ 2f(x_1) \\ 2f(x_2) \\ 2f(x_3) \\ f(x_4) \end{bmatrix}. \tag{6.2.39}$$

This is of course the same expression as (6.2.12) and (6.2.24).
After the elimination of $u_0 = 0$ as described in Figure 6.6 the matrix, S, and the right-hand side, $\mathbf{f}$, become

$$S = \frac{1}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 1 \end{bmatrix}, \quad \mathbf{f} = \frac{h}{2}\begin{bmatrix} 2f(x_1) \\ 2f(x_2) \\ 2f(x_3) \\ f(x_4) \end{bmatrix}. \tag{6.2.40}$$

This construction seems very long-winded, especially for such a simple one dimensional problem. However, it is very well suited for computer implementation. All one needs is a topology formed by a subdivision in elements, as well as a procedure to compute an element matrix and element vector for an arbitrary element. The rest is a matter of book keeping. How complicated the mesh may be, the assembly process is always the same. All finite element codes work according to this principle.

6.2.7 Boundary conditions and assembly

We would like to apply the same procedure as in Section 6.2.6, even in the case of non-homogeneous boundary conditions. To that end we consider the DE (6.2.25) with corresponding minimization problem (6.2.26).

In the right-hand side of (6.2.28) we see two extra terms compared to (6.1.11):

$$-u_0 \int_0^1 \frac{d\varphi_i}{dx}\frac{d\varphi_0}{dx}\,dx + b\varphi_i(1)\,. \tag{6.2.41}$$

The integral in the first of these terms is already present in the element matrix of element e_1:

$$S^{e_1} = \begin{bmatrix} \int_{e_1} \frac{d\varphi_0}{dx}\frac{d\varphi_0}{dx}\,dx & \int_{e_1} \frac{d\varphi_1}{dx}\frac{d\varphi_0}{dx}\,dx \\ \int_{e_1} \frac{d\varphi_1}{dx}\frac{d\varphi_0}{dx}\,dx & \int_{e_1} \frac{d\varphi_1}{dx}\frac{d\varphi_1}{dx}\,dx \end{bmatrix}. \tag{6.2.42}$$

If we skip the first row of S^{e_1} and multiply the remaining part of the first column by u_0 and subtract this term of the right-hand side vector, then we get precisely the first term in (6.2.41).
This step can easily be performed by a finite element program, provided point 0 is marked as a point with essential boundary condition.
So, even if the first row of S^{e_1} is not used, it is conceptually simpler always to create a 2×2 matrix for all elements e_k.
The term $-b\varphi(1)$ only influences the element vector in the last element. However, also in this case it is better not to worry about boundary conditions in the element vector.

In order to create this extra term we introduce an extra boundary element (in this case a point element), consisting of 1 point $(x = 1)$ only. This element is solely meant to incorporate the term $b\varphi_i(1)$

Exercise 6.2.11 *Show that the element matrix and element vector for the boundary condition*

$$\left.\frac{du}{dx}\right|_{(x=1)} = b$$

are given by:

$$S^e = [0]\,, \quad \mathbf{f}^e = [b]\,. \tag{6.2.43}$$

□

The elimination of the essential boundary conditions can be described in the following formulae.

Suppose we renumber the unknowns such that we have first all non-prescribed unknowns $(\mathbf{u}_i)$ (also called *degrees of freedom*) and subsequently all unknowns given by the essential boundary conditions $(\mathbf{u}_b)$.

The system of equations can be written as:

$$\begin{bmatrix} S_{ii} & S_{ib} \\ S_{bi} & S_{bb} \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{u}_b \end{bmatrix} = \begin{bmatrix} \mathbf{f}_i \\ \mathbf{f}_b \end{bmatrix}\,. \tag{6.2.44}$$

Since $\mathbf{u}_b$ is given (6.2.44) can be reduced to

$$S_{ii}\mathbf{u}_i = \mathbf{f}_i - S_{ib}\mathbf{u}_b\,, \tag{6.2.45}$$

and this is the actual system to be solved.

The last set of equations in (6.2.44) contains also some useful information. Suppose that $\mathbf{u}_b$ is not given, but that the flux (natural boundary condition) is prescribed. In that case the last equation would be:

$$S_{bi}\mathbf{u}_i + S_{bb}\mathbf{u}_b = \mathbf{f}_b + \mathbf{b}\,, \tag{6.2.46}$$

where $\mathbf{b}$ is the given flux (see 6.2.43)).

The consequence is that if $\mathbf{u}_b$ is given and $\mathbf{u}_i$ has been solved from (6.2.44), the flux can be approximated by

$$\mathbf{b} = S_{bi}\mathbf{u}_i + S_{bb}\mathbf{u}_b - \mathbf{f}_b\,. \tag{6.2.47}$$

This term is also known under the name *reaction force*. It represents the flux through the boundary with essential boundary conditions.

In the next Section we shall extend our example to two-dimensions.

6.2.8 Periodical boundary conditions

Consider the Poisson equation with periodical boundary conditions.

$$\frac{d^2u}{dx^2} = f, \qquad u(0) = u(1), \quad \frac{du(0)}{dx} = \frac{du(1)}{dx}. \tag{6.2.48}$$

Theorem 6.2.3 *The minimization problem corresponding to* (6.2.48) *is given by*

$$\min_{u\in\Sigma} J[u] = \int_0^1 \{\frac{1}{2}\left(\frac{du}{dx}\right)^2 - f(x)u(x)\}\,dx\,, \tag{6.2.49}$$

$$\Sigma:\ \{u \mid u \text{ sufficiently smooth};\ u(0) = u(1)\}\,.$$

Exercise 6.2.12 *Prove Theorem 6.2.3.* □

Note that the boundary condition $\frac{du(0)}{dx} = \frac{du(1)}{dx}$ is a natural boundary condition for this minimization problem.
In order to apply the Ritz method we set

$$u^n(\mathbf{x}) = \sum_{j=0}^{n} a_j \varphi_j(\mathbf{x}) \,, \tag{6.2.50}$$

with $a_0 = a_n$. Hence the unknowns in the first and last node are identified. By doing so, the first and last element are coupled to each other, which is precisely the idea of periodical boundary conditions.

Exercise 6.2.13 *Compute the matrix and right-hand side for the solution of 6.2.49 using linear basis functions and Newton Cotes quadrature.* □

6.2.9 The structure of finite element packages

In the previous sections it has been made clear that the finite element method is well suited for automatization. As a consequence a lot of (commercial) packages have been developed over the last decades. Most packages subdivide the finite element process in three steps.

- Preprocessing: usually the mesh generation
- Solving: the actual FEM
- Postprocessing: showing the results

The solve part consists globally of the following steps:

```
Read input and mesh
Compute the structure of the large matrix from the topology
Clear large matrix and vector
for all elements (including boundary elements) do
    Compute element matrix and vector
    Add element matrix to large matrix
    Add element vector to large vector
end for
Apply essential boundary conditions
Solve system of equations
Write results for postprocessing
```

The crucial step is the computation of element matrix and vector. In fact this part defines the actual differential equation and type of approximation.
In general one uses preprogrammed finite element subroutines to compute element matrix and vector, however, it is also possible that the user supplies his own element matrix and vector. In this way she may use the general concept of the FEM, and still solve her own specific problem.

6.3 The finite element method in $\mathbb{R}^2$

6.3.1 The Poisson equation in $\mathbb{R}^2$

We have demonstrated the FEM for the one-dimensional Poisson equation using linear interpolations. And even that simple equation showed many of the issues of the FEM. In this section we shall extended that example to $\mathbb{R}^2$. More general cases will be the subject of Chapter 7.

Consider Poisson's equation defined on a bounded region $\Omega \subset \mathbb{R}^2$ with boundary $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$.

$$-\Delta u = f\,, \qquad \mathbf{x} \in \Omega \tag{6.3.1}$$

with boundary conditions

$$\begin{aligned} u &= g_1(\mathbf{x})\,, \qquad \mathbf{x} \in \Gamma_1 \\ \frac{\partial u}{\partial n} &= g_2(\mathbf{x})\,, \qquad \mathbf{x} \in \Gamma_2 \\ \alpha u + \frac{\partial u}{\partial n} &= g_3(\mathbf{x})\,, \qquad \mathbf{x} \in \Gamma_3 \quad (\alpha \geq 0)\,. \end{aligned} \tag{6.3.2}$$

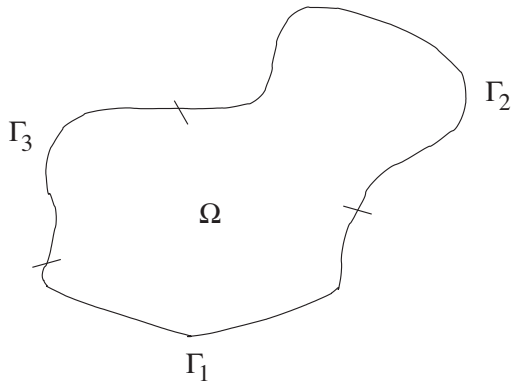


Figure 6.8: Region Ω with boundary $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$.

The minimization problem corresponding to (6.3.1), (6.3.2) is given by

$$\min_{u \in \Sigma} J[u] \tag{6.3.3}$$

with

$$J[u] = \int_\Omega \{\frac{1}{2}|\nabla u|^2 - uf\}\, d\Omega - \int_{\Gamma_2} g_2 u\, d\Gamma - \int_{\Gamma_3} g_3 u\, d\Gamma + \frac{1}{2}\int_{\Gamma_3} \alpha u^2\, d\Gamma$$

and $\Sigma = \{\text{sufficiently smooth} \mid u = g_1|_{\Gamma_1}\}$.

Exercise 6.3.1 *Prove that the PDE formulation (6.3.1) together with (6.3.2) is 'equivalent' to the minimization form (6.3.3).* □

To provide a general framework we first apply Ritz's method formally.

First we choose a set of basis functions $\varphi_i(\mathbf{x}) \in \Sigma_0$ with

$$\Sigma_0 = \{u \mid \ u|_{\Gamma_1} = 0\}\,. \tag{6.3.4}$$

Next we choose an arbitrary but known function u_B that satisfies

$$u_B(\mathbf{x}) = g_1(\mathbf{x}),\ \mathbf{x} \in \Gamma_1\,. \tag{6.3.5}$$

The solution $u(\mathbf{x})$ is approximated by a finite dimensional subset of Σ:

$$u^n(\mathbf{x}) = \sum_{j=1}^{n} u_j \varphi_j(\mathbf{x}) + u_B(\mathbf{x}) \tag{6.3.6}$$

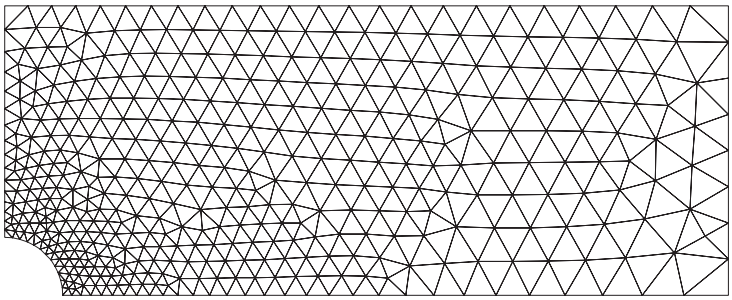


Figure 6.9: Subdivision in triangles.

Clearly we have $u^n(\mathbf{x}) \in \Sigma$. The set of Ritz equations to approximate the minimization problem (6.3.3) by (6.3.6) is given by:

$$\sum_{j=1}^{n} u_j \{ \int_{\Omega} (\nabla \varphi_i \cdot \nabla \varphi_j)\, d\Omega + \int_{\Gamma_3} \alpha \varphi_i \varphi_j \, d\Gamma \} = \int_{\Omega} f \varphi_i \, d\Omega + \int_{\Gamma_2} g_2 \varphi_i \, d\Gamma +$$
$$\int_{\Gamma_3} g_3 \varphi_i \, d\Gamma - \int_{\Omega} \nabla \varphi_i \cdot \nabla u_B \, d\Omega - \int_{\Gamma_3} \alpha \varphi_i u_B \, d\Gamma \,. \tag{6.3.7}$$

Exercise 6.3.2 *Derive (6.3.7).* □

The next step is to provide FEM basis functions. To this end we subdivide the region into elements and define a polynomial approximation on each element.

6.3.2 Linear elements in $\mathbb{R}^2$

The extension of the linear line element in $\mathbb{R}^1$ is the triangle in $\mathbb{R}^2$. Figure 6.9 shows a typical subdivision of a region into triangles. In order to construct a linear polynomial on each triangle we need 3 parameters. A natural choice is to use the function values in the three vertices of the triangle (Figure 6.10).

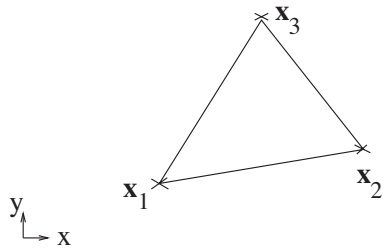


Figure 6.10: Linear triangle with nodal points.

This has the added benefit of making the approximation continuous across element boundaries.
Following the same procedure as in $\mathbb{R}^1$, it will be clear that the corresponding basis functions φ_i have the properties:

$$\begin{aligned} &1) \quad \varphi_i(\mathbf{x}) \quad \text{is linear per triangle ,} \\ &2) \quad \varphi_i(\mathbf{x}_j) = \delta_{ij} \,. \end{aligned} \tag{6.3.8}$$

A typical basis function is sketched in Figure 6.11.

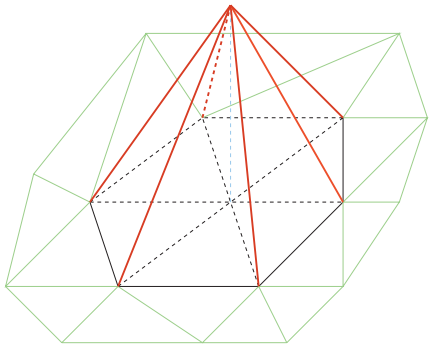


Figure 6.11: Sketch of a typical linear basis function.

(6.3.8) defines the basis functions implicitly. In order to compute the integral in (6.3.7), it is necessary to have an explicit expression per element. Consider the triangle in Figure 6.10).
A linear polynomial is defined by

$$\varphi_i(\mathbf{x}) = \alpha_i + \beta_i x + \gamma_i y \,. \tag{6.3.9}$$

(6.3.8) defines 3 equations for each i to compute the parameters $\alpha_i, \beta_i, \gamma_i$. Substitution of (6.3.8) in (6.3.9) leads to the following system of linear equations:

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} . \tag{6.3.10}$$

Exercise 6.3.3 *Verify (6.3.10)* □

The system of equations (6.3.10) has a solution if the coefficient determinant Δ (see (6.3.11)) does not vanish

$$\Delta = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} . \tag{6.3.11}$$

Δ in (6.3.11) can be expressed as

$$\Delta = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1) \,, \tag{6.3.12}$$

which is twice the area of the triangle in Figure 6.10, as will be shown in Section 8.2.

Exercise 6.3.4 *Prove (6.3.12).*
Hint: subtract the first row from the second and the third row. □

If the orientation of the nodes is counterclockwise Δ is positive, otherwise it is negative.

Exercise 6.3.4 shows that the system is regular as long as the area of the triangle differs from 0.
The solution of system of equations (6.3.10) is given by

$$\begin{aligned} \beta_1 &= \frac{1}{\Delta}(y_2 - y_3) \,, \quad \beta_2 = \frac{1}{\Delta}(y_3 - y_1) \,, \quad \beta_3 = \frac{1}{\Delta}(y_1 - y_2) \,, \\ \gamma_1 &= \frac{1}{\Delta}(x_3 - x_2) \,, \quad \gamma_2 = \frac{1}{\Delta}(x_1 - x_3) \,, \quad \gamma_3 = \frac{1}{\Delta}(x_2 - x_1) \,, \\ \alpha_i &= 1 - \beta_i x_i - \gamma_i y_i \,. \end{aligned} \tag{6.3.13}$$

Exercise 6.3.5 *Show that (6.3.13) is the solution of (6.3.10).*
Hint: formulate the equations for α_1, β_1 and γ_1 and subtract the first equation from the second and third one. Repeat this process for the other unknowns. □

Now we have all ingredients to evaluate the integrals in formula (6.3.7). As we have seen in Section 6.2.5, we only need to compute the element matrix and element vector.
First of all we shall consider the case that α, g_2 and g_3 are all equal to zero, so that all boundary integrals in (6.3.7) vanish. Later on we shall pay attention to these boundary integrals in inhomogeneous boundary problems.
The element matrix for the linear triangle corresponding to (6.3.7) is given by

$$S^{e_k} = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix} , \tag{6.3.14}$$

With $S_{ij} = \int_{e_k} \nabla\varphi_i \cdot \nabla\varphi_j \, d\Omega$.

Exercise 6.3.6 *Show that (6.3.14) is the element matrix corresponding to (6.3.7).* □

From (6.3.9) - (6.3.14) it follows that

$$S_{ij} = \frac{|\Delta|}{2}(\beta_i\beta_j + \gamma_i\gamma_j) \, . \tag{6.3.15}$$

The element vector for the linear triangle corresponding to (6.3.7) is given by:

$$\mathbf{f}^{e_k} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} , \tag{6.3.16}$$

with

$$f_i = \int_{e_k} f(\mathbf{x})\varphi_i(\mathbf{x}) \, d\Omega \, . \tag{6.3.17}$$

Exercise 6.3.7 *Verify (6.3.16) and (6.3.17).* □

We shall have to evaluate (6.3.17) numerically.

6.3.3 Numerical integration in $\mathbb{R}^n$

Numerical integration in $\mathbb{R}^1$ has been the subject of Section 6.2.3. In this section we consider the more general case of integration over triangles in $\mathbb{R}^2$ or tetrahedrons in $\mathbb{R}^3$. Integration over other types of elements will be the subject of Section 8.7.
In $\mathbb{R}^2$ and $\mathbb{R}^3$ we can derive integration rules of the same type as mid-point rule, trapezoidal rule or Simpson's rule, by integrating polynomials of a certain degree exactly. Besides that, for these triangles and tetrahedrons it is possible to construct Gaussian integration rules. Weights and integration points can be found in numerous text books. (See for example [50]).

Definition 6.3.1 *A simplex in $\mathbb{R}^n$ is the convex hull of $n+1$ points in $\mathbb{R}^n$.*

A simplex in $\mathbb{R}^1$ is an interval, in $\mathbb{R}^2$ a triangle and in $\mathbb{R}^3$ a tetrahedron. □

The next theorems gives a general formula for integration of powers of linear basis functions over simplices. It is very useful.

Theorem 6.3.1 *Let S be a triangle in* $\mathbb{R}^2$ *and let* Δ *be the determinant defined by*

$$\Delta = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}, \tag{6.3.18}$$

with $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ *the vertices of S.*

Let $\lambda_i(\mathbf{x})$ *be the linear basis functions over S defined by*

$$\begin{aligned} &\lambda_i(\mathbf{x}) \quad \textit{linear} \\ &\lambda_i(\mathbf{x}_j) = \delta_{ij} \quad i, j, = 1, 2, 3\,. \end{aligned} \tag{6.3.19}$$

Then the following general integration rule holds:

$$\int_S \lambda_1^{m_1} \lambda_2^{m_2} \lambda_3^{m_3} = \frac{m_1! m_2! m_3!}{(m_1 + m_2 + m_3 + 2)!} |\Delta|\,,$$

for all $m_i \geq 0$.

Proof: See Holand and Bell (1969)[20], page 84.

Exercise 6.3.8 *Use Theorem 6.3.1 to show that*

$$\int_S \lambda_i = \frac{|\Delta|}{6}. \tag{6.3.20}$$

□

This theorem can be extended to n dimensions:

Theorem 6.3.2 *Let S be a simplex in* $\mathbb{R}^n$ *and let* Δ *be the determinant defined by*

$$\Delta = \begin{vmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{n,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{n,2} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{n+1,1} & x_{n+1,2} & \cdots & x_{n+1,n} \end{vmatrix}, \tag{6.3.21}$$

with $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{n+1}$ *the vertices of S, and* $x_{i,j}$ *the* j^{th} *component of* $\mathbf{x}_i$.

Let $\lambda_i(\mathbf{x})$ *be the linear basis functions over S defined by*

$$\begin{aligned} &\lambda_i(\mathbf{x}) \quad \textit{linear} \\ &\lambda_i(\mathbf{x}^j) = \delta_{ij} \quad i, j, = 1, 2, \cdots, n+1\,. \end{aligned} \tag{6.3.22}$$

Then the following general integration rule holds:

$$\int_S \lambda_1^{m_1} \lambda_2^{m_2} \cdots \lambda_{n+1}^{m_{n+1}}\, d\Omega = \frac{m_1! m_2! \cdots m_{n+1}!}{(\sum_i m_i + n)!} |\Delta|\,,$$

for all $m_i \geq 0$.

Exercise 6.3.9 *Apply Theorem 6.3.2 to show that*

$$\int_{x_1}^{x_2} \lambda_i \, dx = \frac{h}{2}. \tag{6.3.23}$$

□

Theorem 6.3.3 *Let $\lambda_i(\mathbf{x})$ be defined as in (6.3.22). Then*

$$\sum_{i=1}^{n+1} \lambda_i(\mathbf{x}) = 1 \, . \tag{6.3.24}$$

Exercise 6.3.10 *Prove Theorem 6.3.3.*
□

Exercise 6.3.11 *Show that* $\int_S d\Omega = \frac{|\Delta|}{n!}$.
Hint: use Theorem 6.3.2. □

Exercise 6.3.12 *Find the midpoint rule for a triangle and a tetrahedron.*
Hint: the midpoint rule is a one point integration rule that is exact for linear polynomials. Determine this point by integrating the linear basis functions. □

Exercise 6.3.13 *Prove that the Newton-Cotes rule for a triangle in $\mathbb{R}^2$ with linear basis functions is given by*

$$\int_S g(\mathbf{x}) \, d\Omega = \frac{|\Delta|}{6}(g(\mathbf{x}_1) + g(\mathbf{x}_2) + g(\mathbf{x}_3)) \, . \tag{6.3.25}$$

Hint: use Theorem 6.3.2. □

Exercise 6.3.14 *Show that if the Newton-Cotes rule is applied to (6.3.16), (6.3.17), the element vector is given by*

$$\mathbf{f}^{e_k} = \frac{|\Delta|}{6} \begin{bmatrix} f(\mathbf{x}^1) \\ f(\mathbf{x}^2) \\ f(\mathbf{x}^3) \end{bmatrix} . \tag{6.3.26}$$

□

6.3.4 Boundary conditions

The way in which essential boundary conditions are treated is independent of the dimension of the space. With respect to natural boundary conditions we follow a similar approach as in $\mathbb{R}^1$. In that case we introduced point elements to treat the extra term $b\varphi_i(1)$. In equation (6.3.7) we find four boundary integrals, three of which are related to natural boundary conditions.

$$\int_{\Gamma_3} \alpha \varphi_i \varphi_j \, d\Gamma \, , \tag{6.3.27}$$

$$\int_{\Gamma_2} g_2 \varphi_i \, d\Gamma \, , \tag{6.3.28}$$

$$\int_{\Gamma_3} g_3 \varphi_i \, d\Gamma \, . \tag{6.3.29}$$

Since we use linear triangles we actually approximate the boundary by straight lines. In Section 8.7 we shall return to the consequences of this approximation.

For the moment we assume that the boundary is exactly given by the straight boundary lines of the subdivision. Of course it is possible to add the contribution of the integrals (6.3.27)-(6.3.29) to all element matrices and vectors that correspond to boundary triangles that have a side in common with Γ_2 or Γ_3. From a computational point of view this is not so desirable because this means that not all triangles have the same type of element matrix and element vector. So following our discussion in $\mathbb{R}^1$ it is natural to introduce extra line elements just for the computation of the integrals in (6.3.27)-(6.3.29). These line elements (also called boundary elements) are implicitly defined by the boundary and the subdivision in triangles, see for example Figure 6.12.

Figure 6.12: Subdivision in triangles and line elements.

A typical line element is sketched in Figure 6.13.

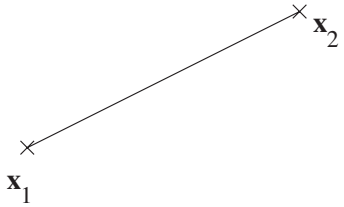


Figure 6.13: Example of a linear line element.

Only two base functions differ from zero on this element (why?), so the element matrix must have size (2×2) and the element vector (2×1). The element matrix for the boundary elements along Γ_3 is given by

$$S^{e_k} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

with

$$S_{ij} = \int_{e^k} \alpha \varphi_i \varphi_j \, d\Gamma \,. \tag{6.3.30}$$

The element vector for the Γ_3 boundary elements are defined by

$$\mathbf{f}^{e_k} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad f_i = \int_{e^k} g_3 \varphi_i \, d\Gamma \,.$$

Exercise 6.3.15 *Give the line element matrices and vectors along* Γ_2. □

To compute the line integrals along element e_k we map the element $(\mathbf{x}_1, \mathbf{x}_2)$ onto $(0, h)$, with h the length of the element given by

$$h = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}\,. \tag{6.3.31}$$

Hence

$$S_{ij} = \int_{e^k} \alpha \varphi_i \varphi_j \, d\Gamma \;\; = \;\; \int_0^h \alpha(t) \varphi_i(t) \varphi_j(t) \, dt, \tag{6.3.32}$$

where $t = 0$ corresponds to x_1 and $t = h$ to x_2.
Application of Newton-Cotes to (6.3.32) gives

$$S_{ij} = \frac{h}{2} \alpha(t_i) \delta_{ij} \;\; = \;\; \alpha(\mathbf{x}_i) \delta_{ij}\,. \tag{6.3.33}$$

Exercise 6.3.16 *Prove* (6.3.33). □

In the same way we can approximate the elements of the element vector along Γ_3 by

$$f_i = \frac{h}{2} g_3(x_i). \tag{6.3.34}$$

Exercise 6.3.17 *Prove* (6.3.34). □

Exercise 6.3.18 *Compute the element matrix and element vector for the line elements along* Γ_2. □

In case of essential boundary conditions we have to choose $u_B(x)$. It is natural to approximate $u_B(x)$ by a linear combination of basis functions corresponding to the points on Γ_1. Hence

$$u_B(x) = \sum_{j=n+1}^{n+n_B} u_j \phi_j(x) \tag{6.3.35}$$

and the approximation can be written as

$$u^n = \sum_{j=1}^{n+n_B} u_j \phi_j(x), \tag{6.3.36}$$

where the last n_B parameters u_j are prescribed. So Equation (6.3.7) reduces to

$$\sum_{j=1}^{n+n_B} u_j \{ \int_\Omega (\nabla \varphi_i \cdot \nabla \varphi_j) \, d\Omega + \int_{\Gamma_3} \alpha \varphi_i \varphi_j \, d\Gamma \} = \int_\Omega f \varphi_i \, d\Omega + \int_{\Gamma_2} g_2 \varphi_i \, d\Gamma + \int_{\Gamma_3} g_3 \varphi_i \, d\Gamma \,, i = 1, 2, \ldots, n\,. \tag{6.3.37}$$

The implementation of essential boundary conditions is exactly the same as described in (6.2.44) and (6.2.45).

6.4 Theoretical remarks

6.4.1 Smoothness requirements

In Section 5.6 we have seen that it is necessary that integrals like $\int_\Omega u_x^2 \, d\Omega$ and $\int_\Omega u_y^2 \, d\Omega$ must exist and be finite. This must also be true for the approximation and hence for the basis functions $\varphi_i(\mathbf{x})$.

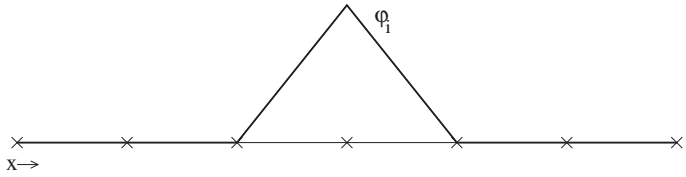


Figure 6.14: One-dimensional basis function $\varphi_i(x)$.

If we consider the one-dimensional basis function $\varphi_i(\mathbf{x})$ sketched in Figure 6.14, then we see that this function is infinitely often differentiable in the interior of each element, but not differentiable on some of the element boundaries (why?). On these boundaries the basis function is continuous. If we split the integral

$$\int_0^1 \frac{d\varphi_i}{dx}\, dx \tag{6.4.1}$$

into

$$\int_0^1 \frac{d\varphi_i}{dx}\, dx = \sum_{k=1}^{m} \int_{e_k} \frac{d\varphi_i}{dx}\, dx\,, \tag{6.4.2}$$

then each of the integrals exists and is finite. This operation is allowed as long as the contribution for the element boundaries in equal to zero.
This is the case if $\varphi_i(x)$ is continuous since the "length" of a point is zero, and therefore the point has no contribution to the integral. Mathematically speaking we say that a point has "zero measure".

However, if $\varphi_i(x)$ would be discontinuous like the one sketched in Figure 6.15, then the derivative on the element interface will be infinite. In fact the derivative is a *delta* function and the contribution of the point with the discontinuity does not vanish. The integral $\int_\Omega u_x^2\, d\Omega$ is no longer finite and such basis functions are not allowed.

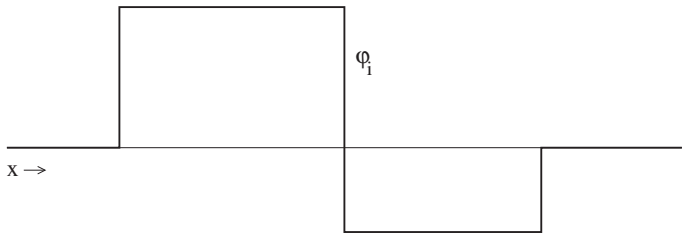


Figure 6.15: Example of a discontinuous basis function.

So for a second order problem in $\mathbb{R}^1$ it is necessary that the basis functions are not only piecewise smooth but also globally continuous.

In $\mathbb{R}^2$ the theory is slightly more complicated but one can say in general that a basis function that is a polynomial per element and continuous over the element boundaries may be used for second order problems. Discontinuous basis functions are in

general not allowed. For fourth order problems we have to require continuity of the first derivative (why?).

Elements with basis functions that satisfy the continuity requirement are called *conforming*, and the basis functions are referred to as admissible. Elements not satisfying this requirement are *non-conforming*. Sometimes they are used for special applications. See Section 8.8.1.

Exercise 6.4.1 *Show that the two-dimensional basis function derived in Section 6.3.2 is admissible.* □

6.4.2 Mathematical theory of FEM

Consider the linear equation

$$Lu = f \quad u \in \Sigma. \tag{6.4.3}$$

The corresponding minimization problem is defined as (see 5.9.9)

$$\min_{u\in\Sigma} J[u], \text{ with } J[u] = \frac{1}{2}\|u\|_L^2 - (u, f). \tag{6.4.4}$$

We have seen that Ritz's method may be formulated as (see 6.1.3)

$$\min_{u_h\in\Sigma_h} J[u_h], \text{ with } \Sigma_h \text{ a finite dimensional subspace of } \Sigma. \tag{6.4.5}$$

Now we can prove the following theorem:

Theorem 6.4.1 *Let $\hat{u}$ be the solution of (6.4.4) and $\hat{u}_h$ be the solution of (6.4.5) then*

$$(\hat{u}, v)_L = (f, v) \quad \forall v \in \Sigma. \tag{6.4.6}$$

$$(\hat{u}_h, v_h)_L = (f, v_h) \quad \forall v_h \in \Sigma_h. \tag{6.4.7}$$

Proof Equation (6.4.6) follows immediately by substituting $u = \hat{u} + \epsilon v$ in (6.4.4), differentiating with respect to ϵ and putting $\epsilon = 0$ like in the derivation of the Euler-Lagrange equations. According to (6.1.19) Ritz's equations $\mathbf{Su} = \mathbf{f}$ can be written as

$$\sum_{j=1}^{n} u_j(\varphi_i, \varphi_j)_L = (f, \varphi_i) \quad (i = 1, ..., n). \tag{6.4.8}$$

Let $v_h \in \Sigma_h$ be given by $v_h = \sum_{i=1}^{n} v_i$. We take the inner product of $\mathbf{Su} = \mathbf{f}$ with the vector $\mathbf{v} = (v_1, v_2, ..., v_n)$ to get

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\varphi_i, \varphi_j)_L u_j v_i = \sum_{i=1}^{n} v_i(f, \varphi_i). \tag{6.4.9}$$

Using the linearity of the inner product we get

$$(\sum_{i=1}^{n} v_i\varphi_i, \sum_{j=1}^{n} u_j\varphi_j)_L = (f, \sum_{i=1}^{n} v_i\varphi_i). \tag{6.4.10}$$

And this of course equal to

$$(\hat{u}_h, v_h)_L = (f, v_h) \quad \forall v_h \in \Sigma_h. \tag{6.4.11}$$

Since v_h is arbitrary we have proved (6.4.7).

With Theorem 6.4.1 we can prove the following

Theorem 6.4.2 *Let $\hat{u}$ be the solution of (6.4.4) over Σ, $\hat{u}_h$ be the solution of (6.4.5) and let $\tilde{u}$ be the interpolation of $\hat{u}$ by the FEM basis functions. So $\tilde{u}$ is defined by*

$$\tilde{u} = \sum_{k=1}^{n} \hat{u}(x_k, y_k)\varphi_k, \tag{6.4.12}$$

with (x_k, y_k) the nodes of the FEM approximation. Then

$$\|\hat{u} - \hat{u}_h\|_L^2 \leq \|\tilde{u} - \hat{u}\|_L^2. \tag{6.4.13}$$

In other words the error in finite element solution is smaller than the error that we would have if we interpolate the solution by the same set of FEM basis functions, at least measured in the energy norm.
In fact the FEM minimizes $\hat{u} - \hat{u}_h$ in energy norm.

Proof:
Since $v_h \in \Sigma$, (6.4.6) is true for each $v_h \in \Sigma_h$. Subtraction of (6.4.7) from (6.4.6) gives

$$(\hat{u} - \hat{u}_h, v_h)_L = 0, \ \forall v_h \in \Sigma_h. \tag{6.4.14}$$

Now choose

$$v_h = \hat{u} - \hat{u}_h - (\hat{u} - w_h), \text{ with } w_h \text{ arbitrary } \in \Sigma_h. \tag{6.4.15}$$

Then

$$(\hat{u} - \hat{u}_h, \hat{u} - \hat{u}_h)_L = (\hat{u} - \hat{u}_h, \hat{u} - w_h)_L, \ \forall w_h \in \Sigma_h. \tag{6.4.16}$$

Using $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ it follows that

$$\|\hat{u} - \hat{u}_h\|_L^2 \leq \frac{1}{2}\|\hat{u} - \hat{u}_h\|_L^2 + \frac{1}{2}\|\hat{u} - w_h\|_L^2. \tag{6.4.17}$$

So

$$\|\hat{u} - \hat{u}_h\|_L^2 \leq \|\hat{u} - w_h\|_L^2, \ \forall w_h \in \Sigma_h. \tag{6.4.18}$$

Substitution of $\tilde{u}$ for w_h proves the theorem.

Remark: this error estimate is only true if we use exact integration and also the region is completely identical to the union of all finite elements. In other words when we use linear triangles like in this chapter, the boundary of the region must consist of piecewise straight lines in order that this error estimate holds (i.e. be polygonal).

6.4.3 Approximation errors

By applying the FEM we make a number of errors. First of all we approximate the solution by a polynomial. This produces an approximation error. We might expect that a higher order polynomial reduces this error. We shall return to this subject in Chapter 8.
Besides that we have approximated the region Ω by straight lines. This too introduces an error. Finally the integrals are approximated by a numerical integration rule. Hence another error is made.
It is clear that these errors must be in balance. Each error should be of the same order, thus producing an optimal result. We shall return to this matter in Chapter 8.

6.5 Summary of Chapter 6

The equivalence of a certain class of PDEs and minimization problems has been proven in Chapter 5. A method to approximate the solution of the minimization problem (Ritz) has been derived. The solution is approximated by a finite set of basis functions.
The FEM is a numerical method that constructs the basis functions by subdividing the region into elements and using a simple polynomial approximation per element. In this chapter we have limited ourselves to 1D and 2D linear elements. The most important property of the FEM basis functions is that they are non-zero in a very limited number of elements.
Since all integrals are computed element-wise it is possible to store all contributions per element in an *element matrix* and *element vector* of small size. By using the generic form of these element matrices and vectors it is very simple to construct the large matrix and right-hand side automatically. In order to approximate integrals per element, numerical integration rules are applied. The Newton-Cotes rule derived in Section 6.2.3 is a very attractive rule since it is based on the FEM basis functions.
Essential boundary conditions in the FEM are implemented by direct substitution. Natural boundary conditions require boundary elements or when homogeneous, no special arrangement at all.

Chapter 7

The weak formulation and Galerkin's method

Objectives

Chapter 5 showed that under certain conditions, solving a PDE is equivalent to solving a minimization problem. For an important class of PDEs, for instance those containing a convective term, these conditions are not met. In order to apply the FEM for such problems, it is necessary to have an alternative formulation. This alternative is based on the weak formulation already mentioned in Section 5.6.3. This formulation is applicable for all kinds of PDEs. Usually it is equivalent to the original conservation law used to derive the PDE.

To solve the weak formulation numerically, the Galerkin method is applied. This method is a direct generalization of Ritz. In case an equivalent minimization method exists, Ritz and Galerkin are identical. Since Galerkin is also based on an expansion in basis functions, the FEM is immediately applicable.

As an extension we shall consider the possibility to introduce upwinding in the FEM by using a special variant of Galerkin, the so-called streamline upwind Petrov Galerkin method (SUPG).

7.1 The weak formulation for a symmetrical problem

7.1.1 Introduction

Let us recall the minimization problem (5.3.1) with boundary condition (5.3.2):

$$\min_{u\in\Sigma} I(u) = \int_{\Omega} \{\frac{k}{2}|\nabla u|^2 - uf\}\, d\Omega, \tag{7.1.1}$$

in which the minimization class Σ is defined by $\Sigma = \{u \text{ smooth} \mid u|_\Gamma = 0\}$.

Γ is the complete boundary of Ω. According to (5.3.5) the solution of (7.1.1) must satisfy

$$\int_{\Omega} \{k(\nabla u \cdot \nabla \eta) - \eta f\}\, d\Omega = 0\,, \forall \eta \in \Sigma. \tag{7.1.2}$$

Integration by parts (5.3.7) resulted in

$$\int_{\Omega} \{-\text{div}\,(k\nabla u) - f\}\eta\, d\Omega = 0, \forall \eta \in \Sigma. \tag{7.1.3}$$

And finally in (5.3.8) we arrived at the differential equation:

$$-\text{div}\,(k\nabla u) = f\,, \tag{7.1.4}$$

with boundary condition,

$$u|_\Gamma = 0\,. \tag{7.1.5}$$

In the derivation of the weak formulation we follow the opposite direction. We start with the differential equation (7.1.4), (7.1.5). Next we multiply this equation by an arbitrary function $\eta \in \Sigma$, and we integrate over the domain Ω. This yields exactly formulation (7.1.3). Integration by parts (Gauss' theorem) applied to (7.1.3) results in (7.1.2).

The arbitrary function η is known under the name *test function* and (7.1.2) is called *weak formulation*. Strictly speaking (7.1.3) is also a form of a weak formulation, but in this book we shall limit ourselves to those forms in which by integration by parts the derivatives have been reduced to the lowest order possible.

Note that the form (7.1.2) is symmetric, whereas (7.1.3) is non-symmetric.

There are several reasons to introduce the weak formulation (7.1.2) instead of the differential equation (7.1.4). First of all, it is easier to prove existence and uniqueness of a solution satisfying (7.1.2) than for one satisfying (7.1.4), (7.1.5). It is clear that a solution that satisfies (7.1.4), (7.1.5) is always a solution of (7.1.2). On the other hand a solution of (7.1.2) requires only the existence of the integral over the first derivatives, and it may be possible that the second derivative does not exist at all. For that reason the term *generalized* or *weak* formulation is used.
The second reason to introduce the weak formulation is that it naturally leads to the FEM. Without weak formulation we are not able to derive a FEM for general PDEs. Of course in this specific case there is no need to use a weak formulation, since we can use the minimization problem (7.1.1) and apply Ritz's method.

7.1.2 Natural boundary conditions

In Section 7.1.1 we have seen a simple example with essential boundary conditions only. Let us extend (7.1.4), (7.1.5) to the complete problem treated in Section 5.3.

So we start with the PDE (5.3.8) with boundary conditions (5.3.9) and (5.3.10):

$$-\text{div}\,(k\nabla u) = f\,, \tag{7.1.6}$$

with boundary conditions,

$$u|_{\Gamma_1} = 0\,, \tag{7.1.7}$$

and

$$k\frac{\partial u}{\partial n}|_{\Gamma_2} = 1\,. \tag{7.1.8}$$

In order to derive the weak formulation we use the solution space Σ of functions satisfying the essential boundary conditions (7.1.7): $\Sigma = \{u \text{ smooth } |u|_{\Gamma_1} = 0\ \}$.

Multiplication of (7.1.6) by a test function η and integration over Ω yields:

$$\int_\Omega \{-\text{div}\,(k\nabla u) - f\}\eta\, d\Omega \;=\; 0, \quad \eta \in \Sigma. \tag{7.1.9}$$

Gauss' theorem applied to (7.1.9), while substituting (7.1.8) gives

$$\int_\Omega \{k(\nabla u \cdot \nabla \eta) - \eta f\}\, d\Omega \;-\; \int_{\Gamma_2} \eta\, d\Gamma = \; 0, \quad \forall \eta \in \Sigma, \tag{7.1.10}$$

and this is precisely Equation (5.3.5). So the natural boundary condition (7.1.8) gives rise to a boundary integral in (7.1.10) but does not influence the solution space nor the space of test functions. In fact the natural boundary condition has been applied by replacing $k\frac{\partial u}{\partial n}$ in the boundary integral on Γ_2.

7.1.3 Non-homogeneous essential boundary conditions

The case of non-homogeneous essential boundary conditions has been considered in Chapter 5. For a minimization problem there is no difficulty in applying such a boundary condition. Already in the one-dimensional example of Sections 5.1.1 and 5.1.2, we have seen that the solution u must satisfy the non-homogeneous essential boundary condition, but that the test function $\eta(\mathbf{x})$ must satisfy a homogeneous essential boundary condition.

The derivation of the minimization problem from a PDE with homogeneous boundary conditions was much more complicated (see Section 5.8.3), but the final result is simple.

From these observations it is logical to derive the weak formulation corresponding to the differential equation (7.1.6) with inhomogeneous boundary conditions by demanding that the test functions satisfy the homogeneous essential boundary conditions.

Consider the PDE (7.1.11) with boundary conditions (7.1.12), (7.1.13),

$$-\text{div}\,(k\nabla u) \;=\; f, \tag{7.1.11}$$

$$u|_{\Gamma_1} \;=\; g_1(\mathbf{x}), \tag{7.1.12}$$

$$\sigma u + k\frac{\partial u}{\partial n}|_{\Gamma_2} \;=\; g_2(\mathbf{x}), \quad \sigma \geq 0. \tag{7.1.13}$$

In order to get the weak formulation we multiply (7.1.11) by a test function $\eta(\mathbf{x}) \in \Sigma = \{\eta \mid \eta|_{\Gamma_1} \;=\; 0\}$, and integrate over Ω.

$$\int_\Omega \eta\{-\text{div}\,(k\nabla u) - f\}\, d\Omega \;=\; 0. \tag{7.1.14}$$

Gauss' theorem gives:

$$\int_\Omega (k\nabla u \cdot \nabla\eta - f\eta)\, d\Omega - \int_\Gamma k\frac{\partial u}{\partial n}\eta\, d\Gamma \;=\; 0. \tag{7.1.15}$$

Since $\eta|_{\Gamma_1} = 0$ and $k\frac{\partial u}{\partial n} \;=\; g_2 - \sigma u$, (7.1.15) can be written as:

$$\int_\Omega k\nabla u \cdot \nabla\eta\, d\Omega + \int_{\Gamma_2} \sigma u\eta\, d\Gamma \;=\; \int_\Omega f\eta\, d\Omega + \int_{\Gamma_2} g_2\eta\, d\Gamma \quad \forall\eta \in \Sigma \tag{7.1.16}$$

and

$$u|_{\Gamma_1} = \; g_1. \tag{7.1.17}$$

(7.1.16), (7.1.17) form our weak formulation.

7.1.4 Periodical boundary conditions

Consider the PDE (7.1.11) with boundary conditions (7.1.12) and periodical boundary conditions on the opposite boundaries Γ_2 and Γ_3

$$-\text{div}\,(k\nabla u) \;=\; f, \tag{7.1.18}$$

$$u|_{\Gamma_1} \;=\; g_1(\mathbf{x}), \tag{7.1.19}$$

$$u|_{\Gamma_2} = u|_{\Gamma_3}, \;\; \frac{\partial u}{\partial n}|_{\Gamma_2} = -\frac{\partial u}{\partial n}|_{\Gamma_3}. \tag{7.1.20}$$

Exercise 7.1.1 *Explain the minus sign in 7.1.20* □

In order to get the weak formulation we multiply (7.1.11) by a test function $\eta(\mathbf{x}) \in \Sigma = \{\eta \mid \eta|_{\Gamma_1} \;=\; 0\}$, and integrate over Ω.

$$\int_\Omega \eta\{-\text{div}\,(k\nabla u) - f\}\; d\Omega \;=\; 0. \tag{7.1.21}$$

Gauss' theorem gives:

$$\int_\Omega (k\nabla u \cdot \nabla\eta - f\eta)\; d\Omega - \int_\Gamma k\frac{\partial u}{\partial n}\eta\; d\Gamma \;=\; 0. \tag{7.1.22}$$

Application of the boundary conditions 7.1.19 and 7.1.20 gives

$$\int_\Omega k\nabla u \cdot \nabla\eta\; d\Omega = \int_\Omega f\eta\; d\Omega \;\;\; \forall\eta \in \Sigma \tag{7.1.23}$$

and

$$u|_{\Gamma_1} = \; g_1, \;\; u|_{\Gamma_2} = u|_{\Gamma_3}. \tag{7.1.24}$$

(7.1.23), (7.1.24) form our weak formulation.

Exercise 7.1.2 *Prove 7.1.23, 7.1.24.* □

The extension to non-symmetric problems is straight-forward as will be shown in Section 7.2. The examples in this section, however, already show that the weak formulation is much easier than deriving the corresponding minimization problem if it exists.

7.2 The weak formulation for a non-symmetric problem

As a generalization of the preceding theory we consider the convection-diffusion equation in two space dimensions:

$$-\text{div}\,(\kappa\nabla T) + \rho c_p(\mathbf{u} \cdot \nabla T) + cT \;=\; f. \tag{7.2.1}$$

T is the temperature, κ the heat conduction, ρc_p the heat capacity, c some non-negative constant and f a source term.

We assume that the boundary Γ is subdivided into three parts Γ_1, Γ_2 and Γ_3.

On Γ_1 we prescribe the essential boundary condition

$$T|_{\Gamma_1} \;=\; g_1(\mathbf{x}). \tag{7.2.2}$$

On Γ_2 the flux is given

$$\kappa \frac{\partial T}{\partial n}|_{\Gamma_2} = g_2(\mathbf{x}). \tag{7.2.3}$$

Finally on Γ_3 we assume a mixed boundary condition

$$\sigma T + \kappa \frac{\partial T}{\partial n}|_{\Gamma_3} = g_3(\mathbf{x}), \quad \sigma \geq 0. \tag{7.2.4}$$

In order to derive the weak formulation we proceed as in the symmetrical case. Equation (7.2.1) is multiplied by a test function η satisfying the homogeneous essential boundary condition $\eta|_{\Gamma_1} = 0$ and integrated over the domain Ω. This results in

$$\int_\Omega \{-\text{div}\,(\kappa \nabla T) + \rho c_p(\mathbf{u} \cdot \nabla T) + cT - f\}\eta \, d\Omega = 0. \tag{7.2.5}$$

Now we apply Gauss' theorem, but only on the second derivative. Application to the first order term would not result in lower order derivatives, since the first derivative of the temperature would be replaced by a first derivative of the test function.

$$\int_\Omega \kappa(\nabla T \cdot \nabla \eta) + \{\rho c_p(\mathbf{u} \cdot \nabla T) + cT - f\}\eta \, d\Omega - \int_\Gamma \kappa \frac{\partial T}{\partial n}\eta \, d\Gamma = 0. \tag{7.2.6}$$

Substituting the boundary conditions (7.2.3) and (7.2.4) as well as the essential boundary condition for the test function:

$$\eta|_{\Gamma_1} = 0. \tag{7.2.7}$$

leads to

$$\int_\Omega \kappa(\nabla T \cdot \nabla \eta) + \rho c_p(\mathbf{u} \cdot \nabla T)\eta + cT\eta \, d\Omega + \int_{\Gamma_3} \sigma T \eta \, d\Gamma = \int_\Omega f\eta \, d\Omega + \int_{\Gamma_2} g_2 \eta \, d\Gamma + \int_{\Gamma_3} g_3 \eta \, d\Gamma. \tag{7.2.8}$$

(7.2.8) together with the boundary conditions (7.2.2) and (7.2.7) forms the weak formulation of Equations (7.2.1) to (7.2.4).

We see that the highest derivative in (7.2.8) is of first order which means that it is sufficient to require that the integrals over the first derivatives exist. Strict mathematically speaking the integral over the square of the first derivatives must exist.

If we suppose the existence of a function $T_1(x)$ which is smooth enough and satisfies the boundary condition (7.2.2), then the weak formulation can be stated as:

> Find T such that $T - T_1 \in \Sigma$ and (7.2.8) is satisfied $\forall \eta \in \Sigma$, with Σ the space of sufficiently smooth functions that satisfy (7.2.7).

7.3 Galerkin's method

7.3.1 Introduction

In Section (6.1) we have introduced Ritz's method as a numerical procedure to solve the minimization problem. The idea was based on the approximation of the unknown solution by a finite linear combination of basis functions:

$$u^n(\mathbf{x}) = \sum_{j=1}^{n} a_j \varphi_j(\mathbf{x}), \tag{7.3.1}$$

and to substitute this in the minimization problem. Minimizing over the set of unknown parameters a_j resulted in a system of linear equations to be solved.

Before considering the general convection-diffusion equation, we start with the symmetrical problem (7.1.4-7.1.5). The weak formulation is given in (7.1.2):

$$\int_\Omega \{k(\nabla u \cdot \nabla \eta) - \eta f\}\, d\Omega = 0\,, \forall \eta \in \Sigma. \tag{7.3.2}$$

Substitution of (7.3.1) in (7.3.2) gives

$$\int_\Omega \{k(\nabla u^n \cdot \nabla \eta) - \eta f\}\, d\Omega = 0\,, \forall \eta \in \Sigma. \tag{7.3.3}$$

(7.3.3) contains n unknown parameters a_j. So for a unique solution we need n equations. Since η is in the same space as u it is natural to demand that η is a linear combination of the n basis functions $\varphi_j(\mathbf{x})$:

$$\eta = \sum_{i=1}^{n} b_i \varphi_i(\mathbf{x}). \tag{7.3.4}$$

η is arbitrary, hence a natural choice is to make one of the coefficients b_i equal to 1 and all others to 0. If i runs from 1 to n this results in exactly n linear equations

$$\sum_{j=1}^{n} a_j \int_\Omega \{k(\nabla \varphi_j \cdot \nabla \varphi_i)\, d\Omega = \int_\Omega f \varphi_i\, d\Omega \quad (i = 1, ..., n). \tag{7.3.5}$$

This is identical to using (7.3.4) for each b_i. Why?

Mark that (7.3.5) is precisely the set of Ritz equation corresponding to the PDE (7.1.1). This method, which is in fact a generalization of Ritz is called Galerkin's method.

Summarizing, the method consists of the following steps:

- Derive the weak formulation corresponding to the PDE.
- Approximate the solution by a linear combination of basis functions.
- Replace the test function by each of the basis function separately.

In mathematical terms we may say that we are solving the weak formulation in the function space Σ, which is expanded by an infinite number of basis functions. In Galerkin's method we are looking for a solution in a finite dimensional subspace of Σ.

7.3.2 Galerkin's method applied to the convection-diffusion equation

The extension of Galerkin's method to more general problems like for example the convection-diffusion equation is straightforward. First we have to derive the weak formulation. For the convection-diffusion Equation (7.2.1) with boundary conditions (7.2.2) to (7.2.4), the weak formulation is given in (7.2.8):

$$\int_\Omega \kappa(\nabla T \cdot \nabla \eta) + \rho c_p(\mathbf{u} \cdot \nabla T)\eta + cT\eta\, d\Omega + \int_{\Gamma_3} \sigma T \eta\, d\Gamma = \int_\Omega f\eta\, d\Omega + \int_{\Gamma_2} g_2 \eta\, d\Gamma + \int_{\Gamma_3} g_3 \eta\, d\Gamma. \tag{7.3.6}$$

The next step is to approximate T by T^n:

$$T^n = \sum_{j=1}^{n+n_b} T_j \varphi_j(\mathbf{x}), \tag{7.3.7}$$

where n_b refers to the prescribed (essential) boundary conditions, and to substitute $\eta = \varphi_i(\mathbf{x})$ for i from 1 to n. This yields the following system of equations:

$$\sum_{j=1}^{n+n_b} T_j \{ \int_\Omega \kappa(\nabla\varphi_j \cdot \nabla\varphi_i) + \rho c_p(\mathbf{u}\cdot\nabla\varphi_j)\varphi_i + c\varphi_j\varphi_i \, d\Omega + \int_{\Gamma_3} \sigma\varphi_j\varphi_i \, d\Gamma \} =$$

$$\int_\Omega f\varphi_i \, d\Omega + \int_{\Gamma_2} g_2\varphi_i \, d\Gamma + \int_{\Gamma_3} g_3\varphi_i \, d\Gamma, \quad i = 1, \ldots, n. \tag{7.3.8}$$

In matrix-vector notation this can be written as $\mathbf{ST} = \mathbf{F}$.

Exercise 7.3.1
Give the elements of the matrix $\mathbf{S}$ *and the right-hand side vector* $\mathbf{F}$.

Why do we have to use j *in the summation (7.3.7) and* i *for the test function and not vice versa?* □

7.3.3 The convection-diffusion equation in $\mathbb{R}^1$ by finite elements

Once the Galerkin equations are derived, we can apply the finite element method since the FEM is just a tool to construct basis functions. In this section we shall limit ourselves to the 1D convection-diffusion equation:

$$-\frac{d}{dx}\kappa\frac{dT}{dx} + \rho c_p u\frac{dT}{dx} = f, \tag{7.3.9}$$

with boundary conditions

$$\begin{aligned} T(0) &= T_0, \\ \kappa\tfrac{dT}{dx}(1) &= 0. \end{aligned} \tag{7.3.10}$$

The weak formulation corresponding to Equation (7.3.9) with boundary conditions (7.3.10) is given by

$$\int_0^1 (\kappa\frac{dT}{dx}\frac{d\eta}{dx} + \rho c_p u\frac{dT}{dx}\eta) \, dx = \int_0^1 f\eta \, dx \tag{7.3.11}$$

with $T(0) = T_0$ and $\eta(0) = 0$.

Hence the Galerkin equations corresponding to the weak formulation (7.3.11) are given by

$$\sum_{j=0}^{n} T_j \int_0^1 (\kappa\frac{d\varphi_j}{dx}\frac{d\varphi_i}{dx} + \rho c_p u\frac{d\varphi_j}{dx}\varphi_i) \, dx = \int_0^1 f\varphi_i \, dx, \quad i = 1, \ldots, n. \tag{7.3.12}$$

In order to apply the finite element method we use the linear basis functions defined in (6.2.10). Again we can introduce finite element matrices and vectors to store the contribution for each element.

Example 7.3.1

Let $\rho c_p = 1$ and κ and u be constant.
Then the element matrix corresponding to (7.3.12) is given by

$$S^{e_k} = \frac{\kappa}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{u}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \tag{7.3.13}$$

with h the length of the element.

For constant values of h the large matrix corresponding to (7.3.4) is given by

$$\frac{\kappa}{h}\begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & & \ddots & & & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix} + \frac{u}{2}\begin{bmatrix} 0 & 1 & & & & & \\ -1 & 0 & 1 & & & & \\ & -1 & 0 & 1 & & & \\ & & & \ddots & & & \\ & & & & -1 & 0 & 1 \\ & & & & & -1 & 1 \end{bmatrix}. \tag{7.3.14}$$

Compare this result with a central finite difference scheme. (Choose end point $x = 1$ between two nodes of the grid).

So from (7.3.14) we may conclude that discretization of the convection term by Galerkin finite elements results in a scheme which is very similar to central differences. How to upwind in case of finite elements will be the subject of Section (7.4.2).

Exercise 7.3.2 *Prove Equation (7.3.11)* □

Exercise 7.3.3 *Prove Equation (7.3.12)* □

Exercise 7.3.4 *Prove Equation (7.3.13)* □

Exercise 7.3.5 *Prove Equation (7.3.14)* □

7.3.4 The convection-diffusion equation in $\mathbb{R}^2$ by finite elements

Consider the convection-diffusion equation (7.2.1) with boundary conditions (7.2.2) to (7.2.4). The weak formulation is given in (7.2.8) and the corresponding Galerkin equations in (7.3.8). For the application of the FEM we subdivide the domain in triangles and use the linear basis functions defined in (6.3.8). The boundary integrals in (7.3.8) are approximated by line elements in exactly the same way as in Section 6.3.4. The element vector for the internal element (triangle) is of course the same as in Formula (6.3.17). Only the element matrix for the triangle is different from that in Section 6.3.1.

Exercise 7.3.6
Show that the element matrix S^{e_k} for the triangle corresponding to (7.3.8) is a 3×3 matrix with elements s_{ij} given by

$$s_{ij} = \int_{e_k} \kappa(\nabla\varphi_j \cdot \nabla\varphi_i) + \rho c_p(\mathbf{u} \cdot \nabla\varphi_j)\varphi_i + c\varphi_i\varphi_j \, d\Omega. \tag{7.3.15}$$

□

Exercise 7.3.7 *Let κ and c be constant scalars, ρc_p be equal to 1 and* $\mathbf{u}$ *be a vector with constant components. Express the elements of the element matrix S^{e_k} in terms of Δ (6.3.11) and the coefficients β_i and γ_i given in (6.3.13).*
Hint: use the Newton Cotes formula to approximate the integrals. □

7.4 Petrov-Galerkin

7.4.1 Introduction

In Section 7.3 we have introduced the Galerkin method. Galerkin is based on the discretization of the weak formulation, where the solution space and the test space are identical. This approach is very common in finite elements and it has many advantages. However, it is not necessary that solution space and space of test functions are the same. In the literature one can find for example a method called collocation, where the test functions are in fact delta functions. This means that one satisfies the differential equation point-wise in the nodal points. In that case it is of course necessary to require more smoothness of the approximation to the solution since integration by parts, to reduce the order of the weak formulation, is no longer possible. For that reason this approach has never become very popular. However, sometimes one uses test functions that do not have the same shape as the basis functions, without affecting the continuity requirements. So starting point is the same weak formulation as for Galerkin. Such methods are for example applied to stabilize the numerical solution. Methods in which the test functions and the basis functions for the solution have different shapes are called Petrov-Galerkin methods.

A typical application in which Petrov-Galerkin methods are used, is convection dominated flow. In finite difference methods it is necessary to use upwind schemes to stabilize the solution, in finite elements Petrov-Galerkin plays the same role.

In the remainder of this chapter we shall use SGA for the Standard Galerkin Approach and SUPG for Petrov-Galerkin. The letters SU will be explained in Section 7.4.3.

7.4.2 Upwinding in $\mathbb{R}^1$ by Petrov-Galerkin

In Section 3.3.2.2 we introduced upwind differencing using the artificial model problem:

$$-\varepsilon\frac{d^2c}{dx^2} + u\frac{dc}{dx} = 0, \tag{7.4.1}$$

with boundary conditions

$$c(0) = 0, \quad c(1) = 1. \tag{7.4.2}$$

Figure 3.8 shows the exact solution for $\varepsilon = 0.01$ and $u = 1$. If we use SGA as in Section 7.3.3 we get a scheme that is almost identical to a central difference scheme (see Exercise 7.3.3). So one may expect the same behavior as for central differences. Figure 7.1 shows that this is indeed the case.

The central difference scheme for Equation (7.4.1) with boundary conditions (7.4.2) is given by:

$$-\varepsilon\frac{c_{i+1} - 2c_i + c_{i-1}}{h^2} + u\frac{c_{i+1} - c_{i-1}}{2h} = 0, \quad i = 1, ..., n. \tag{7.4.3}$$

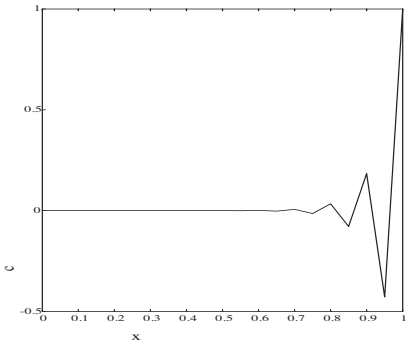


Figure 7.1: Solution of 1d convection-diffusion equation by SGA.

To avoid the unrealistic wiggles we use upwind to approximate the convective terms. The most simple one is first order upwind, where the convection term is discretized by a backward difference scheme for positive velocity u.
The first order upwind scheme for Equation (7.4.1) reads ($u > 0$)

$$-\varepsilon\frac{c_{i+1} - 2c_i + c_{i-1}}{h^2} + u\frac{c_i - c_{i-1}}{h} = 0, \quad i = 1, ..., n. \tag{7.4.4}$$

The local truncation error for Equation (7.4.4) is equal to

$$-\frac{uh}{2}\frac{d^2c}{dx^2} + O(h^2). \tag{7.4.5}$$

Exercise 7.4.1 *Prove (7.4.5)* □

(7.4.5) shows that first order upwind in fact introduces an artificial diffusion of $\varepsilon + \frac{uh}{2}$. There are many other upwind schemes that in fact introduce an artificial diffusion in some clever way.

This observation has inspired FEM researchers to simulate this behavior by a suitable choice of the test functions. To derive the weak formulation for Equation (7.4.1) we multiply by a test function η:

$$\int_0^1 (-\varepsilon\frac{d^2c}{dx^2} + u\frac{dc}{dx})\eta \, dx = 0. \tag{7.4.6}$$

Now the trick is to split $\eta(x)$ into two parts $w(x)$ and $p(x)$ ($\eta(x) = w(x) + p(x)$), where $w(x)$ is the classical test function from the same space as the solution and $p(x)$ is used to take care of the upwind behavior. The $w(x)$ part ensures the consistency of the scheme. This function must be so smooth that integration by parts is allowed. $p(x)$ on the other hand will be defined elementwise, which means that

it may be discontinuous over the element boundaries. In this way Equation (7.4.6) is written as

$$\int_0^1 (\varepsilon \frac{dc}{dx}\frac{dw}{dx} + u\frac{dc}{dx}w)\,dx + \int_0^1 (-\varepsilon\frac{d^2c}{dx^2} + u\frac{dc}{dx})p\,dx = 0. \tag{7.4.7}$$

The last integral is replaced by the sum over the elements, so in fact contributions over the element boundaries are neglected. Hence

$$\int_0^1 (\varepsilon \frac{dc}{dx}\frac{dw}{dx} + u\frac{dc}{dx}w)\,dx + \sum_{elements}\int_{element} (-\varepsilon\frac{d^2c}{dx^2} + u\frac{dc}{dx})p\,dx = 0. \tag{7.4.8}$$

Note that the last term is just a correction to the standard SGA equations. This correction goes to zero if c approaches the exact solution. Of course the choice of p is essential for the behavior of the scheme. Since we are dealing with linear basis functions, the term $\varepsilon\frac{d^2c}{dx^2}$ is zero per element. So the extra term reduces to

$$\sum_{elements}\int_{element} (u\frac{dc}{dx})p\,dx. \tag{7.4.9}$$

Now we choose $p(x)$ such that we get an artificial diffusion of the size $\frac{uh}{2}$.

Exercise 7.4.2
Show that if we choose $p(x) = \frac{h}{2}\frac{d\varphi_i}{dx}$ per element, the artificial diffusion is equal to $\frac{uh}{2}$.

Hint compare Expression (7.4.9) with the discretization of the diffusion term. □

In practice one chooses $p(x) = \frac{h\xi}{2}\frac{d\varphi_i}{dx}$, with ξ some parameter depending on the ratio of u and ε.

ξ equal to sign(u) corresponds to the classical upwind scheme. Popular choices for ξ can be found for example in Brooks and Hughes [6].

7.4.3 SUPG: stream line upwinding in $\mathbb{R}^2$ by Petrov-Galerkin

To extend the upwind Petrov-Galerkin method to 2D one might consider to use the same approach as in Section (7.4.2). An alternative is to apply the upwind technique of Section (7.4.2) to both coordinate directions. However, both approaches have the disadvantage that we have a diffusion term in all directions. The wiggles we get are due to convection, so it makes sense to apply upwind only in the direction of the flow. So a natural choice for upwind in more dimensions is to apply the one-dimensional upwind in the velocity direction. Brooks and Hughes [6] achieved this by replacing the term $p(x) = \frac{h\xi}{2}\frac{d\varphi_i}{dx}$ by $p(\mathbf{x}) = \frac{h\xi}{2}\frac{\nabla\varphi_i\cdot\mathbf{u}}{||\mathbf{u}||}$. This means that the x-derivative of the basis function in the one-dimensional problem is replaced by the directional derivative of the basis function in the direction of the velocity. For h one takes some representative distance in the element, preferably in the direction of $\mathbf{u}$. Since streamlines are always in the direction of the velocity this method is commonly called the Streamline Upwind Petrov Galerkin method (SUPG).

To show the difference between SGA and SUPG we consider the following benchmark problem.

Rotating cone problem