Lecture Notes
Scientific Computing, advanced

Discretisation of PDEs, Finite Element Method

P.W. Hemker

Contents

1 Introduction to PDEs **1**
1.1 Classification 1
1.1.1 Ordinary and partial differential equations 1
1.1.2 Elliptic, parabolic or hyperbolic equations 4
1.1.3 Conservation laws 6
1.2 Hyperbolic equations 7
1.2.1 Characteristics 8
1.2.2 Discontinuous solutions 10

2 Discretisation Principles **17**
2.1 Discrete representation of the solution 17
2.1.1 Discretisation of the domain, solution and equation 19
2.1.2 Finite difference approximation 19
2.1.3 Finite element approximation 21
2.1.4 Finite volume approximation 21
2.1.5 Spectral methods 21
2.1.6 Staggered grid 22
2.2 Techniques for discretisation of PDEs 24
2.2.1 Finite difference methods 24
2.2.2 Variational method 26
2.2.3 Weighted residual methods 27
2.2.4 Collocation methods 27
2.2.5 Galerkin methods 28
2.2.6 Box methods = Finite Volume methods 28
2.3 Examples 29
2.3.1 The diffusion equation 29
2.3.2 The source term and solution of the system 30
2.3.3 The convection equation 31
2.4 Techniques for time-discretisation 31
2.4.1 Time-space discretisation or semi-discretisation 31
2.4.2 FDM for a linear hyperbolic problem 32
2.4.3 The equivalent differential equation 35
2.5 Criteria for good discretisation methods 42

3 Finite Element Methods **45**
3.1 Introduction . . . 45
3.2 Examples of boundary value problems . . . 45
3.2.1 One-dimensional second order . . . 45
3.2.2 Two-dimensional second order . . . 46
3.2.3 Fourth order problems . . . 48
3.3 Abstract Elliptic Theory . . . 48
3.3.1 Introduction to Functional Analysis . . . 49
3.3.2 The Generalised Lax-Milgram Theorem . . . 56
3.3.3 Distributions and Sobolev Spaces . . . 58
3.3.4 The Poincaré-Friedrichs Inequality. . . . 64
3.3.5 Variational formulations for differential equations . . . 65
3.4 The technique of the finite element method . . . 69
3.4.1 The principles . . . 69
3.4.2 Piecewise Lagrange Interpolation in one dimension . . . 70
3.4.3 The construction of the discrete equations . . . 73
3.4.4 Other piecewise polynomial approximations in one dimension . . . 81
3.4.5 Approximation in two dimensions . . . 83
3.4.6 Isoparametric elements . . . 86
3.5 Error estimates for the finite element method . . . 87
3.5.1 The discrete version of the Generalised Lax-Milgram . . . 87
3.5.2 More general error estimate . . . 90
3.5.3 The formalisation of the finite element method . . . 92
3.5.4 Interpolation theory and applications . . . 94
3.5.5 Pointwise error estimate and superconvergence . . . 102
3.5.6 The influence of the quadrature rule . . . 104

References **109**

Index **111**

Chapter 1

Introduction to PDEs

1.1 Classification

In these lectures we study numerical methods for partial differential equations (PDEs). In the later part, in particular, we will concentrate on so called elliptic equations. These equations are used in many areas of application, as e.g. mechanics, electromagnetics, etc..

To get insight into the quantitative behaviour of the solutions of the PDEs, and to determine the solutions with some accuracy, in most cases analytic methods fail and we have to rely on numerical methods. The numerical techniques to compute (approximations to) the solution are often split in two parts. First, the PDE (the solution of which is a continuous function or set of functions) is transformed into a set of linear or nonlinear algebraic equations, usually with a large number of unknowns. This transformation is called the discretisation process. The second step is the solution of the large algebraic system of equations, in order to determine the unknowns that describe (approximately) the solution of the PDE.

In the present chapter different kinds of differential equations are classified, and some elementary properties of the PDEs are explained. In Chapter 2 we will give a survey of the methods that are used for discretisation, for different kinds of PDEs. In Chapter 3 we study the Finite Element Method for the discretisation of elliptic equations.

1.1.1 Ordinary and partial differential equations

An *ordinary differential equation* (ODE) implicitly describes a function depending on a single variable and the ODE expresses a relation between the solution and one or more of its derivatives. The *order of the differential equation* is the order of the highest derivative in the equation. Beside the ODE, usually one or more additional (initial) conditions are needed to determine the unknown function uniquely.

Example 1.1.1
An example of a first order ODE is

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

An example of a second order ODE is

$$u(x) = -c^2\, u''(x), \quad u(0) = 0, u'(0) = 1.$$

In many interesting cases a function depends on more independent variables and a relation is given for the function and its partial derivatives. This relation describes a *partial differential equation* (PDE). In many practical problems the independent variables represent the time and space coordinates (t, x, y, z). To determine the unknown function, again additional relations are required: initial and/or boundary conditions.

Example 1.1.2
Let $u(t, x, y, z)$ denote the mass fraction of a chemical species, m, in a medium with density $\rho(t, x, y, z)$. In the presence of a velocity field $\mathbf{v}(t, x, y, z)$ the conservation of the species m is expressed by

$$\frac{\partial}{\partial t}(\rho u) + \text{div } (\rho u \mathbf{v}) = s, \tag{1.1}$$

where $s(t, x, y, z)$ is a possible source [1].

If diffusion also plays a role, then the *diffusive flux* $\mathbf{J}_d$ is described by Fick's law of diffusion [2]

$$\mathbf{J}_d = -d \,\text{grad } u,$$

where d is the diffusion coefficient, and the complete equation describing the behaviour of $u(t, x, y, z)$ reads

$$\frac{\partial}{\partial t}(\rho u) + \text{div } (\rho \mathbf{v} u) + \text{div } \mathbf{J}_d = s. \tag{1.2}$$

In this equation we distinguish between the diffusive flux, $\mathbf{J}_d$, and the *convective flux*, $\mathbf{J}_c = \rho u \mathbf{v}$.

[1] The *divergence* of the vector field $\mathbf{v}$ is defined by $\text{div } \mathbf{v} = \nabla \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$. An important relation for the divergence is *Gauss' theorem*

$$\int_\Omega \text{div } \mathbf{u}\, d\Omega = \oint_\Gamma \mathbf{u} \cdot \mathbf{n}\, d\Gamma,$$

where $\Omega \subset \mathbb{R}^3$ is a bounded volume with surface Γ and $\mathbf{n}$ is the outward pointing normal on Γ.

[2] The *gradient gradient* of a scalar function $\phi(x, y, z)$ is the vector field defined by

$$\text{grad } \phi = \nabla \phi = \begin{pmatrix} \partial\phi/\partial x \\ \partial\phi/\partial y \\ \partial\phi/\partial z \end{pmatrix}.$$

The *principle part* of a differential equation is the part (the set of terms) that contain the highest derivatives. E.g. the principle part of (1.1) is $(\rho u)_t + \text{div}\,(\rho u \mathbf{v})$ and the principle part of (1.2) is $(\rho u)_t - \text{div}\,(d\,\text{grad}\,u)$. PDEs can have quite diverse properties. A few of such typical forms are given in the next example. Most typical properties are related with the shape of the principle part of the equation.

Example 1.1.3

- With $\Omega \subset \mathbb{R}^2$ an open domain, the *Poisson equation* is

$$u_{xx} + u_{yy} = f(x, y), \tag{1.3}$$

  where $f(x, y)$ is a prescribed function. In the special case of $f \equiv 0$ equation (1.3) is called *Laplace's equation*. The solution of these equations can be determined if the value of $u(x, y)$ is given at the boundary of Ω. Notice that the equation (in its three-dimensional form) is obtained from (1.2) in the case of a time-independent solution with $\mathbf{v} = 0$ and $s = 0$.

- The *diffusion equation*

$$u_t = d\,u_{xx}, \tag{1.4}$$

  describes the one-dimensional unsteady case of simple diffusion, without a velocity field or a source term.

- A different kind of PDE is the *wave equation*

$$u_{tt} = a^2 u_{xx}, \tag{1.5}$$

  where a is a positive real number. It is simple to see that a solution of this equation is given by

$$u(x, y) = \phi(at - x) + \psi(at + x), \tag{1.6}$$

  where ϕ and ψ are arbitrary (twice differentiable) functions.

All the above mentioned examples are differential equations that are linear in the dependent variable u. Although most theory available is for such linear partial differential equations, many important equations are nonlinear. E.g. the equation (1.1) or (1.2) would be nonlinear if the velocity field $\mathbf{v}$ was dependent on u or the source s was a nonlinear function of u. An important subclass of the nonlinear differential equations are the *quasi-linear* equations, i.e. equations in which the highest derivatives appear linearly.

In general, the dependent variable u in the PDE is a function of time, t, and the three space coordinates, (x, y, z). In numerical methods we have to choose how to represent the function u, and often we select values of the independent variables at which the values of u will be calculated. The fewer the number of degrees of freedom in the approximation of $u(t, x, y, z)$, the less computational

effort is needed to determine an approximation of $u(t, x, y, z)$. Fortunately, not all problems require all independent variables. By symmetry considerations (i.e. by a judicious choice of the coordinate system) the number of space dimensions can often be reduced to two or one.

If a solution is sought that is not time-dependent the solution is called steady and the time-variable can be neglected. The corresponding equations are the *steady equations*.

To understand the behaviour of the solution it is important to notice that in some equations we can distinguish a direction of "flow of information". For the solution (1.6) of equation (1.5) we see that the information contained in the part ϕ of the solution flows forward (i.e. in the positive x-direction) with speed a, whereas the information in the part ψ flows backward as time proceeds. Also, in equation (1.1) the flow of information follows the vector field $\mathbf{v}$. However, no such direction is found in (1.3). For time-dependent PDEs such as (1.2) or (1.4) we can always distinguish a unique direction "forward in time".

The motivation for the discussion of "direction of flow of information" is that, if such a direction can be identified, the behaviour of the solution can be better understood. The knowledge also may be used to reduce the computer storage and/or the computer time needed to find the numerical solution. E.g. in a time-dependent problem as (1.1), for which the solution has to be computed over a long time-interval $[t_0, t_2]$, the solution at time $t = t_1$, with $t_0 < t_1 < t_2$, contains all information to determine the solution for $t > t_1$. Having computed the solution for $t \leq t_1$, it may be efficient -from the point of view of computer resources- to disregard all or most information that was obtained about the solution for $t < t_1$ (or $t < t_1 - \tau$, for some small $\tau > 0$) before the solution on $(t_1, t_2]$ is determined.

The notion of direction "forward in time" shows also that in time-dependent problems all or part of the side-conditions will be given as "initial conditions", that determine the solution (i.e. the state of the physical system) at a later stage.

1.1.2 Elliptic, parabolic or hyperbolic equations

In the previous section we noticed that different PDEs may show a different character. This character is mainly determined by the highest derivatives that appear in the equation. To characterise some typical classes of PDEs, we consider the general linear second order differential equation in two space dimensions:

$$\begin{aligned} Lu := \quad & a(x,y)u_{xx} + 2\,b(x,y)u_{xy} + c(x,y)u_{yy} \\ & + 2\,d(x,y)u_x + 2\,e(x,y)u_y + f(x,y)u = g(x,y). \end{aligned} \tag{1.7}$$

L is a linear differential operator of order two. With this differential operator we associate the *characteristic polynomial*

$$\begin{aligned} P_{xy} = \quad & a(x,y)\xi^2 + 2\,b(x,y)\xi\eta + c(x,y)\eta^2 + \\ & + 2\,d(x,y)\xi + 2\,e(x,y)\eta + f(x,y). \end{aligned} \tag{1.8}$$

The second order differential equation (1.7) is called *elliptic* at (x, y) if $a(x, y)c(x, y) - b^2(x, y) > 0$; it is *hyperbolic* at (x, y) if $a(x, y)c(x, y) - b^2(x, y) < 0$, and it is *parabolic* at (x, y) if $a(x, y)c(x, y) - b^2(x, y) = 0$. Thus, this terminology is related with the geometric interpretation of the characteristic polynomial. The differential equation is called elliptic, parabolic or hyperbolic (without reference to a specific point) if it satisfies the corresponding (in)equalities for all $(x, y) \in \Omega$, where Ω is the domain on which the differential equation is defined. We see that the Poisson equation, the diffusion equation and the wave equation are elliptic, parabolic and hyperbolic, respectively.

For an equation of more variables $(x_1, x_2, ..., x_n)$, the general second order equation reads

$$Lu := \sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) u_{x_i x_j} + \sum_{i=1}^{n} a_i(\mathbf{x}) u_{x_i} + a(\mathbf{x}) u. \tag{1.9}$$

Now, for the *principle part* of the equation, the *characteristic polynomial* in $\xi_1, \xi_2, ..., \xi_n$ is

$$P_{\mathbf{x}}(\xi_1, ..., \xi_n) = \sum_{i,j=1}^{n} a_{ij}(\mathbf{x})\, \xi_i \xi_j. \tag{1.10}$$

The operator is elliptic if the matrix $(a_{ij}(\mathbf{x}))$ is positive or negative definite; it is hyperbolic if $(a_{ij}(\mathbf{x}))$ has one negative (positive) eigenvalue and $n-1$ positive (negative) eigenvalues; it is parabolic if one eigenvalue vanishes and the other $n - 1$ have the same sign. It is clear that in more dimensions not all linear second order PDEs can be classified as elliptic, parabolic or hyperbolic.

The characterisation of the different PDEs is important for distinguishing the kind of additional conditions that is needed to specify a (unique) solution for the equation and for understanding the "flow of information" that may appear in its solution. We show this for the three linear examples given earlier in this section.

The hyperbolic equation

For the wave equation (1.5) with initial conditions $u(0, x) = f(x)$, $u_t(0, x) = g(x)$, the general solution for $t \geq 0$ reads

$$u(t, x) = \frac{1}{2} f(x + at) + \frac{1}{2} f(x - at) + \frac{1}{2a} \int_{x-at}^{x+at} g(\xi)\, d\xi. \tag{1.11}$$

This means that the solution at (t, x) depends only on the initial solution on the interval $[x - at, x + at]$.

The parabolic equation

For the linear diffusion equation (1.4), with initial condition $u(0, x) = f(x)$, with $x \in \mathbb{R}$, the solution reads

$$u(t, x) = \frac{1}{2\sqrt{\pi dt}} \int_{-\infty}^{+\infty} f(\xi)\, e^{-(\xi - x)^2/(4dt)}\, d\xi. \tag{1.12}$$

This implies that the solution at $x \in \mathbb{R}$ for any $t > 0$ depends on all of the solution $u(0, x)$ over the interval $-\infty < x < +\infty$.

The elliptic equation

For Laplace's equation, written in polar coordinates (r, ϕ), it is known from the theory of harmonic functions that the solution $u(x, y) = \tilde{u}(r, \phi)$ in the interior of a circle with radius R, is given by

$$\tilde{u}(r,\phi) = \frac{1}{2\pi} \oint_0^{2\pi} \frac{R^2 - r^2}{R^2 - 2Rr\cos(\theta - \phi) + r^2}\, \tilde{u}(R,\theta)\, d\theta. \tag{1.13}$$

This shows that the behaviour of the solution inside a circle, $r = R$, is completely determined by *all* the values of the solution at the boundary of that circle. Generally, the function of an elliptic second order equation is determined by its values at the boundary of its domain of definition.

1.1.3 Conservation laws

A slight extension of equation (1.1) is the system of equations

$$\frac{\partial}{\partial t}\mathbf{u}(t,\mathbf{x},\mathbf{v}) + \text{div } \mathbf{J}(t,\mathbf{x},\mathbf{v})) = \mathbf{s}(t,\mathbf{x},\mathbf{v}), \tag{1.14}$$

where $\mathbf{v}(t, \mathbf{x})$ is the solution sought, $\mathbf{u}(t, \mathbf{x}, \mathbf{v})$ is a state vector and $\mathbf{J}(t, \mathbf{x}, \mathbf{v})$ is the flux of the state variables[3]; $\mathbf{s}(t, \mathbf{x}, \mathbf{v})$ is a possible source term.

Equation(1.14) is called a system of conservation laws because the equations describe the conservation of the quantities $\mathbf{u}$. This is seen by considering an arbitrary volume Ω and by integration of (1.14) over Ω. Using Gauss' theorem, we see that the increase of the amount of $\mathbf{u}$ inside the volume Ω should be caused either by inflow over its boundary or by a source described in $\mathbf{s}(t, \mathbf{x}, \mathbf{v})$.

In natural coordinates, a physical system can often be described with $\mathbf{u} = \mathbf{u}(\mathbf{v})$, $\mathbf{J} = \mathbf{J}(\mathbf{v})$, $\mathbf{s} = \mathbf{s}(\mathbf{v})$. Then, the dependence on $(t, \mathbf{x})$ may be caused by a coordinate transformation.

Example 1.1.4 The Euler equations.
The *Euler equations* of gas dynamics describe the flow of an inviscid non-heat-conducting compressible fluid (a gas). They represent the conservation of mass, momentum and energy. With $\rho(t, \mathbf{x})$ density, $\mathbf{v}(t, \mathbf{x})$ velocity, $e(t, \mathbf{x})$ specific energy (temperature) and $p(t, \mathbf{x})$ the presure of the gas, these conservation laws read respectively

$$\frac{\partial}{\partial t}(\rho) + \text{div } (\rho\mathbf{v}) = 0, \tag{1.15}$$

$$\frac{\partial}{\partial t}(\rho v_i) + \text{div } (\rho v_i \mathbf{v}) = -\frac{\partial p}{\partial x_i}\,, \quad i = 1, 2, 3,$$

[3]Because we consider a system of equations, now each component of $\mathbf{J}$ is a vector.

$$\frac{\partial}{\partial t}(\rho e) + \text{div } (\rho e \mathbf{v}) = -\text{div } (p\mathbf{v}),$$

and $p = p(\rho, e, \mathbf{v})$. For a perfect gas we have the state equation

$$p = (\gamma - 1)\rho(e - \frac{1}{2}|\mathbf{v}|^2). \tag{1.16}$$

With suitable boundary conditions the variables $\rho, \mathbf{v}, e$ and p can be solved from (1.15) and (1.16). These variables play the role of the vector $\mathbf{v}$ in (1.14). Notice that they are different from the conserved quantities $\mathbf{u}$.

1.2 Hyperbolic equations

The second order equation (1.5) allows decomposition into a system of two first order equations. Similar to the factorisation of the characteristic polynomial $P(\tau, \xi) = \tau^2 - a^2\xi^2 = (\tau + a\xi)(\tau - a\xi)$ of (1.5), the first order operators are found to be $\frac{\partial}{\partial t} + a\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial t} - a\frac{\partial}{\partial x}$. The decomposition first can be written[4] as a system

$$\begin{cases} \frac{\partial u}{\partial t} = -a\frac{\partial p}{\partial x}, \\ \frac{\partial p}{\partial t} = -a\frac{\partial u}{\partial x}. \end{cases} \tag{1.17}$$

If we make the change of dependent variables $v = u + p$ and $w = u - p$, this equation (1.17) uncouples and transforms into

$$\begin{cases} \frac{\partial v}{\partial t} + a\frac{\partial v}{\partial x} = 0, \\ \frac{\partial w}{\partial t} - a\frac{\partial w}{\partial x} = 0. \end{cases} \tag{1.18}$$

The solution of the system reads $v(t, x) = v(x - at)$, $w(t, x) = w(x + at)$. These equations show that the quantities $u + p$ (respectively $u - p$) are constant along the line $x - at =$ constant ($x + at =$ constant). These lines are called the *characteristics* of the differential equation (or of the differential operator).

A typical first order "one-way" wave equation in $\mathbb{R}^d$, $d = 1, 2, 3$, is

$$\frac{\partial u}{\partial t} - \mathbf{v} \cdot \text{grad } u = 0. \tag{1.19}$$

It shares with (1.18) the property of having a characteristic direction, $\mathbf{v}$, along which the solution is constant. Therefore also the first order equations are called hyperbolic PDE. To illustrate further the importance of the characteristics, we consider the inhomogeneous equation

$$\frac{\partial u}{\partial t} - \mathbf{v}\,\text{grad } u = s(u). \tag{1.20}$$

We assume that the vector field $\mathbf{v}(\mathbf{x})$ is sufficiently smooth so that it determines a complete family of characteristics in Ω. Then we will show in Section 1.2.1 that the solution of this PDE is uniquely determined as soon as *one* value of the solution is given for each characteristic. Along the characteristic the solution is now determined by a simple ODE.

[4]The minus sign is by convention only.

1.2.1 Characteristics

We consider a general first order (quasi-)linear DE in more dimensions. We write it down for dimension $d = 3$; for the one-, two- or more- dimensional case the treatment is completely analogous. We consider the equation

$$Pp + Qq + Rr = S\,, \tag{1.21}$$

where $p = u_x$, $q = u_y$, $r = u_z$, and P, Q, R, S are functions of (x, y, z, u). I.e. (x, y, z) are the independent variables and u is the dependent variable.

Let $u(x, y, z)$ be the solution of (1.21), then, by definition of p, q, and r, (1.21) is satisfied as well as

$$du = pdx + qdy + rdz.$$

Or, if A is a point (x, y, z, u) on a solution, then at A we find

$$\begin{array}{lcl} (p, q, r, -1) & \perp & (P, Q, R, S), \\ (p, q, r, -1) & \perp & (dx, dy, dz, du), \\ (p, q, r, -1) & \perp & (\dot{x}, \dot{y}, \dot{z}, \dot{u}), \end{array}$$

with $\dot{} = \partial/\partial s$, and $A(s)$ is an (arbitrary) curve in a solution surface. Apparently, the vector field $(P, Q, R, S)(x, y, z, u)$ is tangential to the solution in each point of the solution. Streamlines in this vector field are characteristic solutions. The projections of these characteristic solutions on the (x, y, z)-space are characteristics (characteristic direction lines). For a solution along a characteristic solution line (parametrised by s) we have

$$(P, Q, R, S) \parallel (\dot{x}, \dot{y}, \dot{z}, \dot{u}), \tag{1.22}$$

i.e. the two vectors are linearly dependent. It follows that the characteristic direction can be computed from

$$\frac{P}{dx} = \frac{Q}{dy} = \frac{R}{dz} = \frac{S}{du}\cdot \tag{1.23}$$

Further, it follows that, if a value of (x, y, z, u) is given in one point of a characteristic, then the characteristic and the solution on it can be found by the solution of the simple ODE (1.23).

If the solution is prescribed on a suitable $(d-1)$-dimensional manifold (suitable means that the manifold intersects all characteristics once) then the solution is apparently uniquely determined.

Example 1.2.1 Inviscid Burgers' equation.
inviscid Burgers' equation A well-known model problem is the inviscid Burgers' equation

$$u_t + (\frac{1}{2}u^2)_x = 0\,.$$

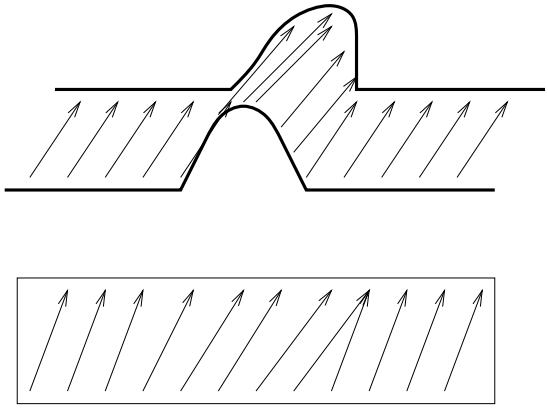


Figure 1.1: The generation of a shock for Burgers' equation

(See also example (1.2.2)). We notice that

$$\frac{1}{dt} = \frac{u}{dx} = \frac{0}{du}\,.$$

The characteristic satisfies $dx/dt = u$ and $du/dt = 0$, so that the solution is constant along the characteristics, and the characteristics are straight lines. If the initial solution is increasing in some area, the characteristics will intersect and no smooth solution will further be possible. This will be considered further in the next section.

We may study the above arguments more generally in d-dimensional space. Let a solution be prescribed along a $(d-1)$-dimensional manifold C. Under what conditions is a solution determined in a layer in the neighbourhood of C? We easily extend the question to the case of systems of equations. For simplicity we take $d = 3$. Let

$$\begin{cases} a_1 u_x + b_1 u_y + c_1 u_z + d_1 v_x + e_1 v_y + f_1 v_z + \cdots = g_1, \\ a_2 u_x + b_2 u_y + c_2 u_z + d_2 v_x + e_2 v_y + f_2 v_z + \cdots = g_2, \\ \cdots, \end{cases} \tag{1.24}$$

or, in short,

$$\sum_{\substack{j\,=\,1,\,2,\,3 \\ i\,=\,1,\,2,\,\cdots,\,m}} a_{kij} \frac{\partial u_i}{\partial x_j} = b_k, \quad k = 1, \cdots, m\,. \tag{1.25}$$

We introduce a function $\phi(x_1, x_2, x_3)$ such that a surface Φ is determined by $\phi(x_1, x_2, x_3) = 0$. (The surface Φ is our $(d-1)$-dimensional manifold, with d=3.) Moreover, we introduce a new set of independent variables (ϕ_1, ϕ_2, ϕ_3) such that the surface is determined by $\phi_1 =$ constant; e.g. $\phi_1 = 0$. (I.e. we introduce local coordinates around the surface Φ.) Now we have

$$\frac{\partial u_i}{\partial x_j} = \sum_{l=1,2,3} \frac{\partial u_i}{\partial \phi_l} \frac{\partial \phi_l}{\partial x_j}\,.$$

Equation (1.25) becomes

$$\sum_{\substack{l,\,j\,=\,1,\,2,\,3\\ i\,=\,1,\,2,\,\cdots,\,m}} a_{kij}\frac{\partial u_i}{\partial \phi_l}\frac{\partial \phi_l}{\partial x_j} = b_k, \quad k = 1, \cdots, m\,, \tag{1.26}$$

and we get

$$\sum_{i=1,\cdots,m}\left[\sum_{j=1,2,3} a_{kij}\frac{\partial \phi_1}{\partial x_j}\right]\frac{\partial u_i}{\partial \phi_1} = b_k - \sum_{\substack{l\,=\,2,\,3\\ j\,=\,1,\,2,\,3\\ i\,=\,1,\,2,\,\cdots,\,m}} a_{kij}\frac{\partial u_i}{\partial \phi_l}\frac{\partial \phi_l}{\partial x_j}. \tag{1.27}$$

This is a linear $m \times m$-system for the computation of $\partial u_i/\partial \phi_1$, i.e. the normal component of u_i on the surface Φ. For the computation we need (i) the coefficients a_{kij}; (ii) the right-hand-sides b_k; (iii) $\partial u_i/\partial \phi_l$, $l = 2, 3$, on the surface $\phi_1 = 0$.

If the solution $u(x, y, z)$ of the system of PDEs is known in the surface Φ, then we can also find the normal derivatives (and hence the analytic solution in a layer around Φ), provided that

$$\det\left(\sum_{j=1,2,3} a_{kij}\frac{\partial \phi_1}{\partial x_j}\right) \neq 0. \tag{1.28}$$

If, on the contrary, det = 0, then -in general- no smooth solution exists in the neighbourhood of Φ.

1.2.2 Discontinuous solutions

For simplicity, in this section we first consider the one-dimensional case. We have seen that the solution of

$$u_t(t, x) + a\, u_x(t, x) = 0, \tag{1.29}$$

with $a \in \mathbb{R}$, and the initial condition

$$u(0, x) = u_0(x)\,, \quad x \in \mathbb{R}, \tag{1.30}$$

is $u(t, x) = u_0(x - at)$. Also, with $a(x)$ a continuous function, the behaviour of $u(t, x)$ is easily understood for $t > 0$. The problem gets considerably more interesting if the problem is quasi-linear, i.e. if we consider

$$u_t(t, x) + a(u)\, u_x(t, x) = 0, \tag{1.31}$$

where $a(u)$ is a continuous function depending on the solution. Together with the initial condition (1.30), it now turns out that not in all cases a smooth solution exists for all $t > 0$, even if $u_0(x)$ is a smooth function. The reason is that characteristic lines for the equation (1.31) may intersect. As we have seen, a

classical solution of (1.31) should be constant on a characteristic line, and different characteristics can carry different values. At points where characteristics intersect a shock-line is formed and a discontinuity appears. The characteristic line ends. But this cannot be described by the differential equation.

There are essentially two ways out of this difficulty. The first approach is to generalise the concept of a solution. The class of possible solutions is extended from the class of differentiable functions to the class of integrable functions, and the equation is considered in a weak sense, i.e. it is written in an integral form. (We see this form later in (1.36).) However, the weak solutions turn out to be non-unique, (for a given set of initial data), and it remains to characterise the "physically relevant" weak solutions.

Before we consider this question, we present the second approach. Instead of (1.31) we now consider the slightly perturbed equation

$$u_t(t,x) + a(u)\,u_x(t,x) = \epsilon u_{xx}, \quad 0 < \epsilon \ll 1. \tag{1.32}$$

This corresponds with the transition from (1.1) to (1.2) and it means that we add a small diffusion term to the original equation without diffusion. Under reasonable conditions for the initial function $u_0(x)$ the parabolic equation (1.32) always has a unique smooth solution $u^\epsilon(t,x)$ for $\epsilon > 0$, and a (non-smooth) limit function $u^0(t,x) = \lim_{\epsilon\to 0} u^\epsilon(t,x)$ exists.

Now we return to the first approach, where we obtain weak solutions of (1.31). As the functions $u^\epsilon(t,x)$ converge to such a weak solution, then that solution is called the *physically relevant* weak solution of (1.31).

The integral or weak form[5] of the conservation law in divergence form[6]

$$u_t + \text{div } \mathbf{J}(u) = 0, \tag{1.34}$$

is found by integration over an area $R = [t_0, t_1] \times \Omega$. The integral form reads

$$\int_{[t_0,t_1]} \int_\Omega u_t + \text{div } \mathbf{J}(u)\, d\mathbf{x}\, dt = 0, \tag{1.35}$$

[5]The general principle of *weak solution* for a differential equation $N(u) = 0$ on $\mathbb{R}^d$ is actually to take an arbitrary $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R}^d)$, i.e. an arbitrary smooth function with compact support, and to consider the integral

$$\int_0^\infty \int_{\mathbb{R}^d} \phi(t,\mathbf{x})\, N(u) d\mathbf{x}\, dt = 0.$$

In our case, with $N(u) = u_t + \text{div } \mathbf{J}(u)$, on the half space $t \geq 0$, after partial integration this comes down to

$$-\int_0^\infty \int_{\mathbb{R}^d} u\phi_t + \mathbf{J}(u) \cdot \text{grad } \phi \; d\mathbf{x}\, dt = -\int_{\mathbb{R}^d} \phi(0,\mathbf{x})\, u_0(\mathbf{x}) \; d\mathbf{x} \tag{1.33}$$

for all $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R}^d)$. We see that here the requirement of a differentiable u has disappeared from the formulation of the equation. By taking a sequence of functions $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R}^d)$, that have the characteristic function on $[t_0, t_1] \times \Omega$ as its limit, it can be shown that (1.33) and (1.36) are equivalent.

[6]Notice that (1.19) is written as (1.34) with $\mathbf{J}(u) = -\mathbf{v}u$, with div $\mathbf{v} = 0$, because $\text{div}(-\mathbf{v}u) = -u\nabla_j(v_j u) = -u\nabla_j v_j - v_j \nabla_j u = -\mathbf{v}\text{grad}\, u$.

which can also be written as

$$\int_\Omega u(t_1)\,d\Omega - \int_\Omega u(t_0)\,d\Omega = -\int_{[t_0,t_1]} \oint_{\partial\Omega} \mathbf{n}\cdot\mathbf{J}(u)\,d\Gamma\,dt. \tag{1.36}$$

The integral form holds also if there is a $(d-1)$-dimensional manifold on which $u(t,x)$ is discontinuous. Thus, (1.35) is a generalisation of (1.34).

For $d=2$ we immediately see (cf. Figure ???) that at a discontinuity D we have

$$\left(\frac{dx}{dt}\right)_D [u] = [\mathbf{J}_\perp(u)]\,, \tag{1.37}$$

where $[u]$ denotes the jump of u over the discontinuity. Denoting the propagation speed of the discontinuity by s, we see[7]

$$s = \frac{\mathbf{J}_\perp(\mathbf{u}_R) - \mathbf{J}_\perp(\mathbf{u}_L)}{u_R - u_L}. \tag{1.38}$$

This expression for the shock speed is called the *Rankine-Hugoniot* relation.

Apparently two types of discontinuities may appear in first order hyperbolic equations: linear and nonlinear ones. The first type is present in the initial condition and is carried along a characteristic. It is convected with a speed $\mathbf{v}$ that does not essentially depend on the solution. The second type may appear even if the initial solution is smooth, and the speed essentially depends the solution.

Example 1.2.2 Burgers' equation.
The appearance of a nonlinear discontinuity is often shown for Burgers' equation. In one space dimension it reads

$$u_t + uu_x = \epsilon u_{xx}. \tag{1.39}$$

If the r.h.s. is neglected, i.e. if we take $\epsilon = 0$, in a positive solution the top of a wave moves faster than the bottom. Hence, after a sufficiently long period this would lead to a multiple valued function (if that would be possible). Because such solutions are not allowed, a discontinuity appears of which the speed s is determined by

$$\mathbf{J}(u) = \frac{u^2}{2}, \tag{1.40}$$

and it follows that the discontinuity moves with a speed $s = (u_{\text{left}} + u_{\text{right}})/2$.

[7] In more dimensions we get

$$\mathbf{J}_\perp(\mathbf{u}_R) - \mathbf{J}_\perp(\mathbf{u}_L) = s(\mathbf{u}_R - \mathbf{u}_L)$$

where $\mathbf{J}_\perp$ denotes the flux vector component in the direction perpendicular to the discontinuity.

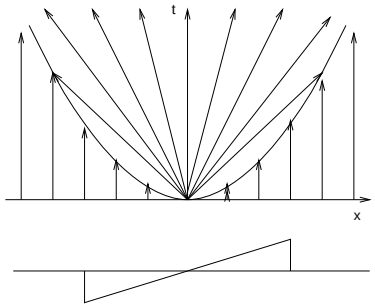


Figure 1.2: A curved shock solution for Burgers' equation

If we do allow $\epsilon > 0$, then the r.h.s. will always have a significant value in those regions the solution is steep and the dissipation will prevent the discontinuity to appear[8].

Example 1.2.3 *A discontinuous solution.*
As an example (from [17]), the following wave is also a solution to the inviscid *Burgers' equation*:

$$u(t,x) = \begin{cases} x/t, & -\sqrt{t} < x < \sqrt{t}, \\ 0, & \text{otherwise.} \end{cases} \tag{1.43}$$

This solution (see Figure 1.2) has two shocks, propagating with speeds $\pm 1/2\sqrt{t}$. The right shock has left- and right states $u_L = 1/\sqrt{t}$ and $u_R = 0$, so the Rankine-Hugoniot condition is satisfied. The other shock behaves similarly.

We can use the relation (1.38) to solve explicitly some initial value problems that are not classically solvable. However, now it appears that these solutions are not always uniquely determined. We take again the inviscid Burgers' equation as an example.

[8] For some initial conditions the shape of the dissipated shock is easy to determine for Burgers' equation. With

$$u(0,x) = \begin{cases} u_{-\infty} = q > 0, & x < 0, \\ u_{\infty} = -q < 0, & x > 0, \end{cases} \tag{1.41}$$

the steady solution of (1.39) is

$$u(\infty, x) = -q \tanh\left(\frac{qx}{2\epsilon}\right). \tag{1.42}$$

Notice that the initial condition is already a weak solution of the reduced Burgers' equation ((1.39) with $\epsilon = 0$). A non-trivial steady solution does not appear for $u_{-\infty} < 0 < u_{\infty}$. In that case, after a first dissipative phase, the nonlinear convection will level out all differences in u.

The dissipated shock moves at a speed $a = (u_L + u_R)/2$, and its shape is given by

$$u(t,x) = \frac{u_L + u_R}{2} - \frac{u_L - u_R}{2} \tanh\left(\frac{(u_L + u_R)(x - at)}{2\epsilon}\right).$$

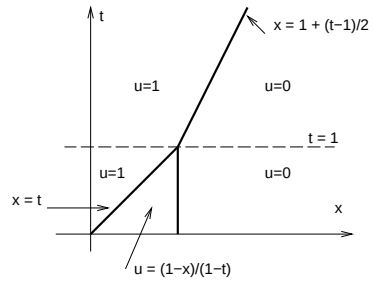


Figure 1.3: A continuous solution developing a discontinuity

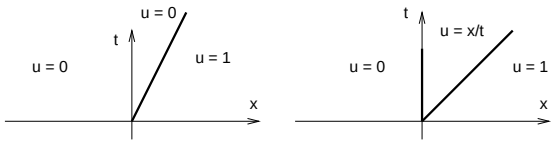


Figure 1.4: A discontinuous and a continuous solution solution

Example 1.2.4 *A unique solution.*
We take Burgers' equation with initial function

$$u_0(x) = \begin{cases} 1, & x < 0, \\ 1-x, & 0 \le x \le 1, \\ 0, & x > 1. \end{cases} \tag{1.44}$$

The solution we find reads (see Figure (1.3))

$$u(t,x) = \begin{cases} 1, & x < t, \text{ or} \\ & x < 1 + (t-1)/2, \\ (1-x)/(1-t), & t \le x \le 1, \\ 0, & x > 1, \text{ or} \\ & x > 1 + (t-1)/2. \end{cases} \tag{1.45}$$

Example 1.2.5 *A non-unique solution.*
We take Burgers' equation with initial function

$$u_0(x) = \begin{cases} 0, & x < 0, \\ 1, & x > 0. \end{cases} \tag{1.46}$$

The solutions we find are (see Figure (1.4)) are a family of discontinuous functions for $a \in (0,1)$

$$u(t,x) = \begin{cases} 0, & x < at/2, \\ a, & at/2 < x < (1+a)t/2, \\ 1, & (1+a)t/2 < x. \end{cases} \tag{1.47}$$

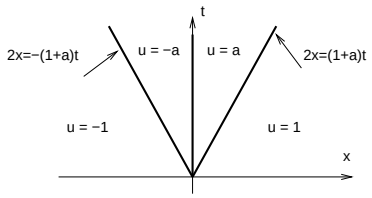


Figure 1.5: Another discontinuous solution

and a continuous function

$$u(t,x) = \begin{cases} 0, & x < 0, \\ x/t, & 0 < x < t, \\ 1, & x > t. \end{cases} \tag{1.48}$$

Example 1.2.6 *Another non-unique solution.*
We may extend the previous example to obtain a more complex non-unique solution. We again take Burgers' equation, now with initial function

$$u_0(x) = \begin{cases} -1, & x < 0, \\ 1, & x > 0. \end{cases} \tag{1.49}$$

The solution we find reads (see Figure (1.5)), with some $a \in (0,1)$,

$$u_a(t,x) = \begin{cases} -1, & 2x < -(1+a)t, \\ -a, & -(1+a)t < 2x < 0, \\ +a, & 0 < 2x < (1+a)t, \\ +1, & (1+a)t < 2x. \end{cases} \tag{1.50}$$

To pick out the "physically relevant solution" we have to find the solution $u(t,x) = \lim_{\epsilon \to 0} u^\epsilon(t,x)$, with u^ϵ the solution of

$$u_t + f(u)_x = \epsilon u_{xx}\,.$$

First, however, we formulate a condition that guarantees unicity of the discontinuous solution. Such conditions are generally called *entropy condition*s. Later we shall see a relationship between these entropy conditions and the physical relevance of a solution. In the case of a single equation $u_t + f(u)_x = 0$ with $f'' > 0$, one can show that there exists a unique solution that satisfies the "entropy" condition [*Oleinik*] (1.51),

$$\frac{u(t,x+h) - u(x,t)}{h} \leq \frac{E}{t}, \quad h > 0,\ t > 0, \tag{1.51}$$

where E is independent[9]of x, t, and h. This condition implies that if we fix $t > 0$, and we let x go from $-\infty$ to $+\infty$, then we can only jump down (in one direction) over the discontinuity.

[9]E depends on $\|u_0\|_\infty$, $\min\{f''(u)\}$ and $\max\{f'(u)\}$. We do not go into details. The interested reader is referred to Chapter 16 of [23].

This entropy condition of *Oleinik* guarantees uniqueness of the weak solution. Condition (1.51) says that, at discontinuities, we have $u_L > u_R$. From the Rankine-Hugoniot condition we have

$$s = \frac{f(u_L) - f(u_R)}{u_L - u_R} = f'(\zeta)\,,$$

for some $\zeta \in [u_R, u_L]$. We had assumed $f'' > 0$ (i.e. f is convex) and hence we finally find that Oleiniks entropy condition implies

$$f'(u_L) > s > f'(u_R)\,. \tag{1.52}$$

Sometimes this inequality is also called the *entropy condition*. The inequality (1.52) means that the characteristic lines on both sides of the discontinuity run into the discontinuity. The entropy condition (1.51) of Oleinik has not been extended to systems of conservation laws. However (1.52) has been extended by Lax [16].

Example 1.2.7 *Traffic flow (shocks and rarefaction waves).*
(This example is taken from [17].) Let $0 \le \rho(x,t) \le \rho_m$ denote car density on a road, and $u(x,t)$ car speed. Because cars don't disappear we have

$$\rho_t + (\rho u)_x = 0\,.$$

Let u be given as $u = u(\rho)$, e.g. $u(\rho) = u_m(1 - \rho/\rho_m)$. Combining both equations, gives

$$\rho_t + f(\rho)_x = 0\,.$$

with $f(\rho) = \rho u_m(1 - \rho/\rho_m)$. We easily derive the characteristic speed

$$f'(\rho) = u_m(1 - 2\rho/\rho_m)\,,$$

and the shock speed for a jump in the car density from ρ_L to ρ_R:

$$s = \frac{f(\rho_L) - f(\rho_R)}{\rho_L - \rho_R} = u_m(1 - (\rho_L + \rho_R)/\rho_m)\,. \tag{1.53}$$

The entropy condition implies that a shock must satisfy $f'(\rho_L) > f'(\rho_R)$, which implies $\rho_L < \rho_R$. Now, with initial data

$$\rho(x,0) = \begin{cases} \rho_L, & x < 0, \\ \rho_R, & x > 0, \end{cases}$$

several situations can be distinguished. With $0 < \rho_L < \rho_R < \rho_m$ a shock wave travels with speed (1.53). This shock speed can move in the positive or in the negative direction. With $0 < \rho_R < \rho_L < \rho_m$ we rather have a "rarefaction wave'.

Chapter 2

Discretisation Principles

In many cases it is important to obtain quantitative data about physical, technical or other real-life systems. One may obtain these through measurements. However, if we have sufficiently accurate mathematical models for a system, we can also try to derive quantitative data from these models. Models for technical problems often take the form of a PDE or a system of PDEs, and although much analytical theory is available for PDEs, most of these equations do not allow an explicit solution in closed form. Therefore numerical methods are used to get insight in their quantitative behaviour.

In practice it is clear that there are many restrictions in gathering reliable data from real-life measurements. In the numerical modelling of these systems, limitations of accuracy are caused by possible inaccuracies in the mathematical model, because the model is often a simplification of reality, and also because there is always a limitation of computing resources.

Numerical mathematics takes the mathematical equations as a starting point and strives for an efficient and accurate approximation of the the quantitative data.

2.1 Discrete representation of the solution

In many problems from practice, the solution of a PDE is a steady or time-dependent (vector-) function in d space variables ($d = 1, 2, 3$). Thus the solution is an element of an infinitely-dimensional space, and -in general- it cannot be described by a finite number of real numbers. Because computational capacity is always finite, the problem has to be discretised: i.e. the solution has to be approximated by a function that *can* be represented by a finite set of numbers.

Example 2.1.1
Let $d = 1$, and let the solution of the equation be a simple continuous function $u(x)$, $x \in [a, b] = \Omega$. A few possible ways to represent this function are:

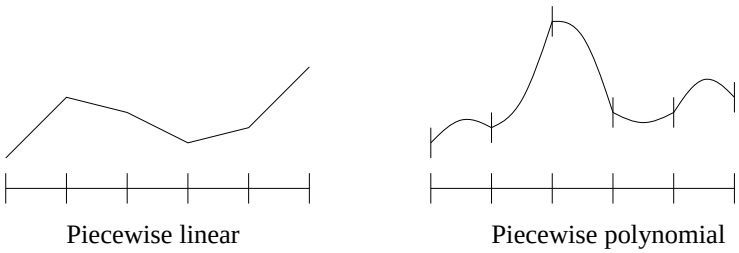


Figure 2.1: Piecewise polynomial approximations

i) a table of equidistant function values

$$u_i \approx u(x_i), \quad i = 0, 1, 2, \cdots, N, \quad x_i = a + i(b-a)/N;$$

ii) a table of non-equidistant function values

$$u_i \approx u(x_i), \quad i = 0, 1, 2, \cdots, N, \quad a \leq x_0 \leq x_1 \leq \cdots \leq \cdots \leq x_N \leq b;$$

iii) a table of function values and values of derivatives of that function (either equidistant or non-equidistant)

$$u_{i,k} = \left(\frac{d}{dx}\right)^k u(x_i); \quad k = 0, 1; \quad i = 0, 1, 2, \cdots, N.$$

In all these cases the solution at the intermediate values of x can be approximated by interpolation. This can be done e.g. by a choosing a polynomial, a spline or a piecewise polynomial as the interpolating function.

Another way to approximate the function is by choosing -a priory- a finite-dimensional function space, selecting a basis in this space, and determining the coefficients of the approximation with respect to that basis. Then we have to select a family of functions $\{\phi_i(x) | i = 1, \cdots, N\}$, which forms the basis for the approximating function space, and we compute coefficients $\{a_i | i = 1, \cdots, N\}$ so that the function

$$\phi(x) = \sum_{i=1}^{N} a_i \phi_i(x) \tag{2.1}$$

approximates the function that is to be represented.

Possible choices for $\{\phi_i\}$ are, e.g. (i) the polynomials $\phi_i(x) = x^i$; (ii) goniometric functions $\{\phi_i(x)\} = \{1, \sin x, \cos x, \sin(2x), \cos(2x), \cdots\}$; or (iii) piecewise polynomials on a partition of the interval $[a, b]$, e.g. piecewise linear or piecewise parabolic functions.

To obtain a more accurate approximation of the function $u(x)$, we need -in general- a denser set of function values or more coefficients to compute. If

a partition of nodal points is made, or a partition in subintervals, the largest distance between two neighbouring points, or the size of the largest interval in the partition is usually denoted by a parameter h. This parameter is called the mesh-size. For a smaller h the approximation gets (in general) (i) more accurate, and (ii) more laborious (time-consuming) ("more expensive") to compute.

We notice that the classification of approximation methods does not yield disjoint classes of methods. In the representation (2.1) the coefficients may coincide with function values or values of derivatives. Example (iii) can also be seen as an interpolation method for a set of function values, given at the sub-interval endpoints. If there is a set of nodal point $\{x_i\}$ and a set of basis function $\{\phi_i\}$, such that they satisfy the relation

$$\phi_i(x_j) = \delta_{i,j},$$

where $\delta_{i,j}$ denotes the Kronecker delta, the same method can be considered as a pointwise approximation and as a functional approximation as well.

2.1.1 Discretisation of the domain, solution and equation

In the transformation of the PDE into a finite set of equations, (i.e. in the discretisation process) it is convenient to distinguish a number of stages:

i) the discretisation of Ω, the domain of definition of the PDE;

ii) the discretisation of u, the solution (i.e. the selection of the representation for the solution); and

iii) the discretisation of the equation.

Before we describe the techniques how to discretise the equations, we first describe different methods for the discretisation of the domain and of the solution.

To find the discrete representation of the solution, one usually selects in the domain where the PDE is defined, a (regular or irregular) *mesh* of points, or a *partitioning* of the domain in triangles or quadrilaterals (See Figure 2.2 or 2.3).

2.1.2 Finite difference approximation

For the finite difference method (FDM), the domain Ω is represented by a finite subset of points $\{\mathbf{x}_i\} = \Omega_h \subset \Omega$. These points are the so called "nodal points" of the grid. This grid is almost always arranged in (uniform or non-uniform) rectangular manner, see Figure 2.2.

In the finite difference method (FDM) the function u is represented by a set of function values u_i that approximate $u(\mathbf{x}_i)$. The discrete equations are obtained by replacing the differentials in the PDE by finite differences in the discrete equations.

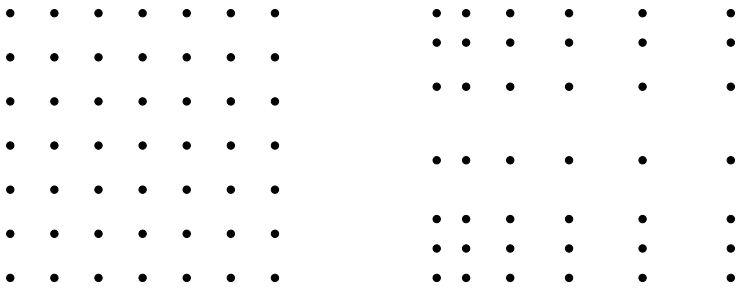


Figure 2.2: Sets of nodal points in rectangular regions

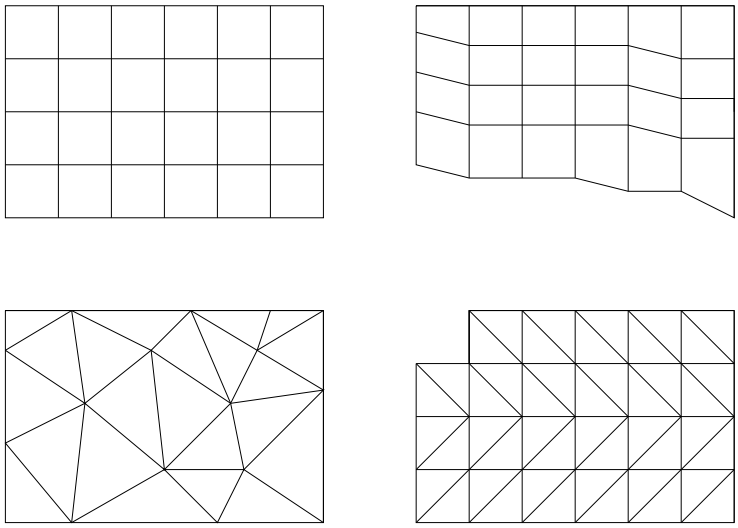


Figure 2.3: Partition of two-dimensional domains

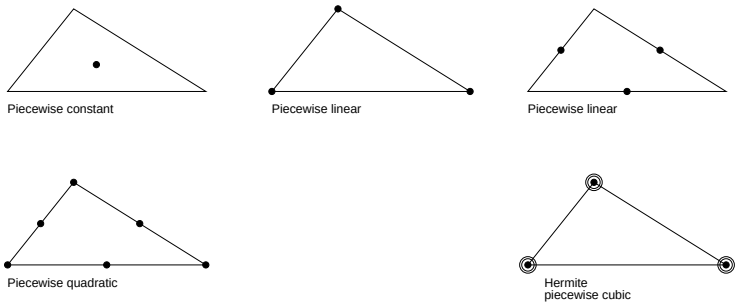


Figure 2.4: Examples of nodal points in a triangular domain

2.1.3 Finite element approximation

In the finite element method (FEM) the domain Ω is partioned in a a finite set of elements $\{\Omega_i\}$, so that $\Omega_i \cap \Omega_j = \emptyset$, for $i \neq j$, and $\cup\overline{\Omega_i} = \overline{\Omega}$. Usually one takes for Ω_i triangles or quadrangles. Then the function is approximated by

$$u_h = \sum a_i \phi_i(x)\,,$$

where ϕ_i are functions that are polynomials on each Ω_i (i.e. *piecewise polynomials*). Usually the functions ϕ_i are continuous polynomials of a low degree. Further they are constructed so that their support extends only over a small number of elements.

2.1.4 Finite volume approximation

In this case the domain Ω is also partitioned in a finite set of volumes (mostly triangles or quadrangles) so that $\Omega_i \cap \Omega_j = \emptyset$, for $i \neq j$, and $\cup\overline{\Omega_i} = \overline{\Omega}$. Now in A *cell centered* (or *block centered*) finite volume method (FVM), the function u is approximated by

$$u_h = \sum a_j \phi_j(x)\,, \tag{2.2}$$

where the (possibly discontinuous) functions $\phi_i(x)$ are defined only on a single volume. or on the boundaries of the volumes.

If there is a single coefficient a_i associated with each volume (so that a_i may represent $u(x)$ for $x \in \Omega_i$), then the grid is called a *block centered* or *cell centered*. If a (single) coefficient is associated with each vertex, then the method is called *cell vertex* (see figure 2.5). If the discretisation is constructed on cells (dashed lined in Figure 2.5) that are located around vertices of a grid (solid lines), then the mesh is called *vertex centered*.

2.1.5 Spectral methods

The function u is again approximated by

$$u_h = \sum a_i \phi_i(x)$$

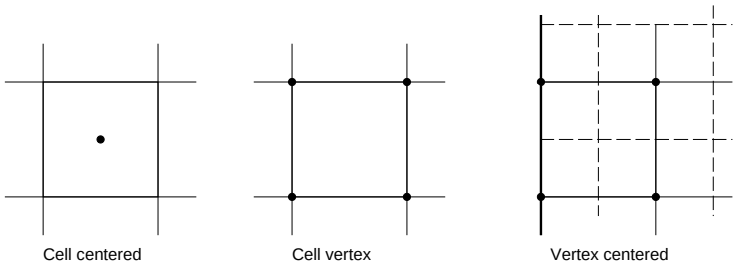


Figure 2.5: Examples of nodal points in a quadrangular domain

where the functions $\phi_i(x)$ are defined on all of Ω (they have a support that extends over more than only a small part of Ω) then the methods are called global or spectral methods. (We see that the different types of methods are not always clearly to distinguish.)

2.1.6 Staggered grid

For systems of equations all methods mentioned above are easy to generalise. All functions can simply be replaced by vector functions. However, it is also possible to approximate the different (dependent) variables by different representations. One way is to approximate the different components of the solution on different point sets. This is the case e.g. with "staggered grids". We explain this by two examples.

Example 2.1.2
We consider the *Cauchy Riemann* equations

$$\left.\begin{array}{ll} (i) & u_x + v_y = f_1 \\ (ii) & u_y - v_x = f_2 \end{array}\right\} \quad \text{on} \quad \Omega \subset \mathbb{R}^2\,, \tag{2.3}$$

Ω is a rectangular domain, and Ω is partitioned in a number of equal subrectangles. Approximations are computed for u and v. The values of u are calculated for the midpoints of the vertical edges, the values of v for the midpoints of the horizontal edges. The discrete equations corresponding with equation (2.3(i)) are related with the midpoints of the cells. The discrete equations corresponding with equation (2.3(ii)) are constructed at the cell vertices.

To understand the character of these equations, we write the Cauchy-Riemann equations as

$$L \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial y} & -\frac{\partial}{\partial x} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{2.4}$$

We see that

$$\det(L) = -\left(\frac{\partial}{\partial x}\right)^2 - \left(\frac{\partial}{\partial y}\right)^2 = -\Delta\,.$$

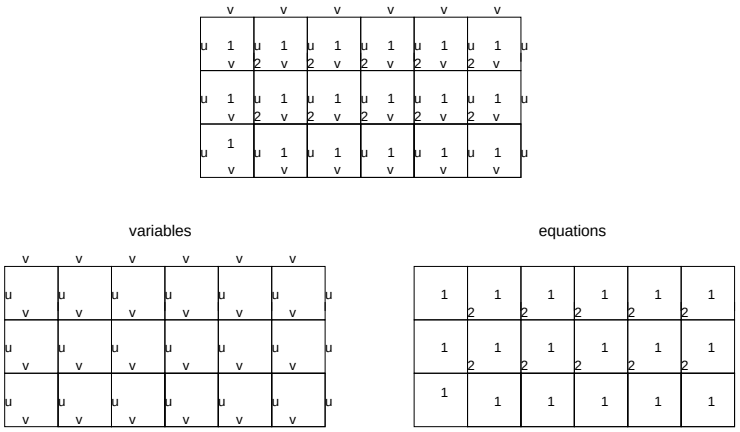


Figure 2.6: A staggered grid for the Cauchy-Riemann equations

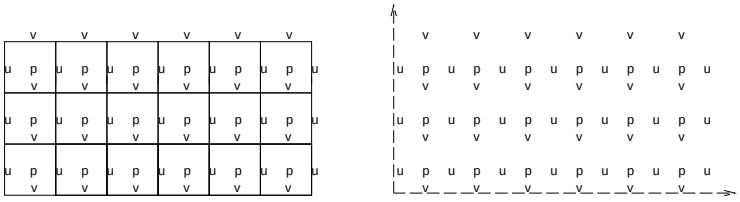


Figure 2.7: A staggered grid for Stokes equations

This shows that the system is elliptic and second order. It is also directly seen by elimination that [1] $\Delta u = \text{div } \mathbf{f}$, or $\Delta v = -\text{curl } \mathbf{f}$.

The system being second order elliptic, it needs *one* boundary condition along all the boundary. We take

$$u.n_x + v.n_y = \mathbf{v}.\mathbf{n} = g(x,y)\,.$$

In our discrete system, this condition fixes the values of u and v on the boundary. Now the number of unknowns is 27 (viz. 3×5 for u, and 2×6 for v), and the number of equations is 28 (3×6 for equation (i), and 2×5 for equation (ii)). Hence, the system seems to be overdetermined. However, it is easily seen that the boundary function $g(x,y)$ should satisfy a condition. From $u_x + v_y = \text{div } \mathbf{v} = f_1$, it follows that

$$\int_\Omega u_x + v_y \, d\Omega = \int_\Omega \text{div } \mathbf{v} \, d\Omega = \oint_{\partial\Omega} \mathbf{v}.\mathbf{n} \, d\Gamma = \oint_{\partial\Omega} g(x,y) \, d\Gamma\,,$$

[1] In three dimensions

$$\text{curl}\,\mathbf{v} = \begin{pmatrix} \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3} \\ \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1} \\ \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \end{pmatrix},$$

in two dimensions $\text{curl}\,\mathbf{v} = \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2}$.

i.e. the *compatibility condition*

$$\int_\Omega f_1(x,y)\,\Omega = \oint_{\partial\Omega} g(x,y)\,d\Gamma\,.$$

Similarly it follows from adding all discrete equations $(u_R - u_L)/h + (u_T - u_B)/h = f_1$ for each cell, that -in total-

$$h\sum_k g_k = h\sum_i u_{R,i} - h\sum_i u_{L,i} + h\sum_i v_{T,i} - h\sum_i v_{B,i} = h^2\sum_i f_{1,i}, \quad (2.5)$$

i.e.

$$\sum_{k\in\text{boundary}} g_k = h \sum_{i\in\text{interior}} f_{1,i} \quad (2.6)$$

Here $f_{1,i}$ denotes the value of $f_1(x,y)$ in the i-th cell in the interior of Ω, and g_k denotes the value of $g(x,y)$ in the k-th boundary edge element. If the boundary values satisfy condition (2.6), one (linear) dependence relation exists between the 28 equations and we get a solvable system of equations.

Example 2.1.3 *Stationary Stokes equations*
Analogous to the staggered grid for the Cauchy-Riemann equations, we can construct one for the stationary *Stokes equations*,

$$\begin{cases} \eta\Delta u - p_x = 0, \\ \eta\Delta v - p_y = 0, \\ u_x + v_y = 0. \end{cases}$$

This can be written as a system with linear differential operator L,

$$L\begin{pmatrix} u \\ v \\ p \end{pmatrix} = \begin{pmatrix} \eta\Delta & 0 & \nabla_1 \\ 0 & \eta\Delta & \nabla_2 \\ \nabla_1 & \nabla_2 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \\ p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The character of the system of equations is determined by the determinant of the principle part of the differential operator

$$\det L = \eta\Delta\cdot\nabla_1^2 + \eta\Delta\cdot\nabla_2^2 + 0 = \eta\Delta^2\,.$$

This is a 4-th order elliptic operator, for which we need two conditions over all of the boundary. The operator Δ is represented by the finite difference stencil $\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}$, the operator ∇_1 by $\begin{bmatrix} -1 & 1 \end{bmatrix}$ and ∇_2 by $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

2.2 Techniques for discretisation of PDEs

2.2.1 Finite difference methods

For this method, we have already seen examples. The discrete equation is constructed by replacing differential operators by difference operators. The

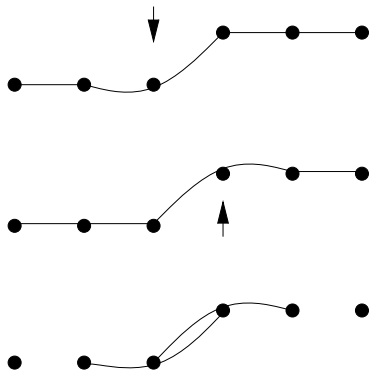


Figure 2.8: Difference approximation and polynomial interpretation

construction of the difference operators is as follows: in the neighbourhood of a point $\mathbf{x}\in\Omega$ the approximation of $u(x)$ is replaced by a truncated Taylor series:

$$u(\mathbf{x}) = a_0 + a_{10}x + a_{01}y + a_{11}x^2 + \cdots.$$

With the values u_{ij} in the nodal points x_{ij} around $\mathbf{x}$, the coefficients a are expressed in these function values u_{ij}. These coefficients a_{pq} are associated with the derivatives of the approximation in $\mathbf{x}$. Now, in the PDE at this point the differentials are replaced by difference approximations. In this way, for each unknown u_{ij} an equation is set up at the point x_{ij}.

For a regular grid, in the interior points of the domain the same difference approximation can be used. Different difference approximations appear near the boundary of the domain, or in parts of the domain where the nodal points are not regularly placed.

In the neighbourhood of $\mathbf{x}_{ij}$, the approximation of $u(\mathbf{x})$ is, thus, considered as a piecewise polynomial (of a low degree). That does *not* mean that the complete function u is considered as a polynomial. The interpretation of the polynomial representation depends on the nodal point where the interpretation is made.

The (linear or nonlinear) algebraic system of equations that results from the discretisation process has an important property. By the difference approximations the value of the unknown function at a particular point is only coupled with the values at neighbouring points. In this way a "sparse system" of algebraic equations appears. The corresponding matrix (of the linearisation) only contains a few non-zero elements in each row; most entries in the matrix vanish. Whereas the systems that appear as a result of discretisations can become very large, they mostly have this nice property, by which the solution can often be found in a relatively efficient[2] manner. As this property is shared by most discretisations of PDEs, special methods are sought (and found) for the efficient solution of such sparse systems of equations.

[2]Compared with a full system of equations of the same dimensions.

2.2.2 Variational method

Variational method for difference methods

An important class of PDEs, the symmetric elliptic equations, can be written as a minimisation problem. This gives another possibility to construct the discrete equations. We show this by means of the Poisson equation

$$\left.\begin{array}{ll} \Delta u = f, & \text{on } \Omega, \\ u = 0, & \text{on } \Gamma\,. \end{array}\right\}$$

This equation can be written as the minimisation problem: find a continuous function u such that $u = 0$ on Γ, and

$$J(u) = \int_\Omega \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + 2fu \; d\Omega$$

exists and is minimal. (Exists: $u_x^2 + u_y^2$ and fu should be integrable functions.)

Now we can set up a discrete analogue for this minimisation problem [5]. On a regular square grid (i.e. Ω a rectangle, and square meshes with a meshsize h) we can write: find $u_h = \{u_{ij}\}$ such that $u_{ij} = 0$ on Γ, and

$$J_h(u_h) = \sum_{ij} \left(\frac{u_{i+1,j} - u_{i,j}}{h}\right)^2 + \left(\frac{u_{i,j+1} - u_{i,j}}{h}\right)^2 + 2f_{ij}u_{i,j}$$

is minimal. The minimisation of this quadratic functional in the unknowns $\{u_{ij}\}$ leads to linear system in $\{u_{ij}\}$. This system also yields a difference equation for each internal point of the grid.

The variational method for coefficients

For problems that can be written as minimisation problems, we can use the *variational method* also if we seek an approximation in the form

$$u_h(x,y) = \sum_i a_i \phi_i(x,y)\,.$$

We take the same example (2.2.2), so that with $\phi_i(x,y) = 0$ on Γ. Now we compute the coefficients a_i such that

$$\int_\Omega \left\{\sum_i a_i \frac{\partial}{\partial x}\phi_i(x,y)\right\}^2 + \left\{\sum_i a_i \frac{\partial}{\partial y}\phi_i(x,y)\right\}^2 + 2f(x,y)\sum_i a_i\phi_i(x,y)\, d\Omega$$

attains its minimum. This minimisation problem also leads to a linear system, now in the unknown coefficients (a_i). The same technique is applicable on irregular grids.

2.2.3 Weighted residual methods

A quite general discretisation technique is based on the following consideration. Let the PDE be given by

$$\left.\begin{array}{ll} \mathcal{N}_\Omega(u) = s_\Omega, & \text{on } \Omega, \\ \mathcal{N}_\Gamma(u) = s_\Gamma, & \text{on } \Gamma, \end{array}\right\} \tag{2.7}$$

then, for any function w we have

$$\int_\Omega w_\Omega(\mathcal{N}_\Omega(u) - s_\Omega)\, d\Omega + \oint_\Gamma w_\Gamma(\mathcal{N}_\Gamma(u) - s_\Gamma)\, d\Gamma \;=\; 0. \tag{2.8}$$

Or, in an even more general notation, let the equation be $\mathcal{N}(u) = s$ and let $s, \mathcal{N}(u) \in V$, with V a Banach space, then for an arbitrary linear functional $l_w \in V'$ the equation[3]

$$l_w\left(\mathcal{N}(u) - s\right) = 0$$

is satisfied. Now, let S be a set of functions in which the solution u is sought: $\mathcal{N} : S \to V$, then the PDE can be formulated as: find $u \in S$ such that

$$l_w(\mathcal{N}(u) - s) = 0, \quad \text{for all } l_w \in V'.$$

In a weighted residual method an n-dimensional subspace $S_h \subset S$ is selected and an n-dimensional subspace $V_h^D \subset V'$. The discretisation now reads: find $u_h \in S_h$ such that

$$l_w(\mathcal{N}(u_h) - s) = 0 \quad \text{for all } l_w \in V_h^D.$$

In this way, the discretisation yields an $n \times n$-system of equations.

Of course, there are many different applications of the weighted residual principle. We will treat the most important.

2.2.4 Collocation methods

Collocation methods are weighted residual methods with the functionals l_w in the space V_h^D that is spanned by

$$\{l_{x_i} \,|\, x_i \in \Omega;\ i = 1, \cdots, N\}\ .$$

Here l_{x_i} is defined by

$$l_{x_i}(f) = f(x_i), \text{ for } f \in V.$$

This means that the discrete equations are

$$\mathcal{N}(u_h)(x_i) = s(x_i), \text{ for } i = 1, \cdots, N.$$

This means that the approximate solution u_h should satisfy the original PDE precisely in n given points, where $n = \dim(S_h)$.

[3]The dual space of a Banach space V (denoted by V') is the family of all bounded linear functionals defined on V, furnished with the obvious addition and scalar multiplication.

2.2.5 Galerkin methods

When the functionals l are defined by integrals with integrable weight functions, we obtain "Galerkin"-methods. Typically $l_w \in V_h^D$ is defined by

$$l_w(f) = \int_\Omega w(x)\, f(x)\, d\Omega\,.$$

The space V_h^D is now defined by n linearly independent functions $w_i(x)$.

If we choose $\{w_i(x)\}_{i=1,2,\cdots,N}$, such that $V_h := \text{span}\,(w_i) = S_h$ then the methods are called *Bubnov-Galerkin* methods. An obvious choice is the $w_i = \phi_i$. Then the discrete equations read

$$\int_\Omega \phi_j\, \mathcal{N}(\sum a_i \phi_i)\, d\Omega = \int_\Omega \phi_j\, s\, d\Omega.$$

If we take $V_h := \text{span}\,(w_i) \neq S_h$, the method is called a *Petrov-Galerkin* method.

Almost all usual Finite Elements methods are of the (Bubnov-) Galerkin type. When applied to symmetric positive definite elliptic PDEs, these methods can also be regarded as variational methods.

2.2.6 Box methods = Finite Volume methods

Box methods are a special kind of Galerkin methods. Here the domain Ω is partitioned in a number of volumes $\{\Omega_i\}_{i=1,\cdots,N}$ I.e. $\Omega_i \cap \Omega_j = \emptyset$, if $i \neq j$, and $\cup\overline{\Omega_i} = \overline{\Omega}$. The weight functions are now the characteristic functions on these Ω_i,

$$\left.\begin{array}{ll} w_i(x) = 1, & \text{if } x \in \Omega_i, \\ w_i(x) = 0, & \text{if } x \notin \Omega_i, \end{array}\right\}$$

and the discrete equations read

$$\int_{\Omega_i} \mathcal{N}\left(\sum a_i \phi_i\right)\, d\Omega_i = \int_{\Omega_i} s\, d\Omega_i.$$

The volumes Ω_i are called "boxes", "cells" or "control volumes".

Box methods are particularly popular for the discretisation of equations in divergence form

$$\text{div}\ \mathbf{J}(u) = s(u), \quad \text{on } \Omega.$$

Then the discrete equation gets the form

$$\oint_{\partial\Omega_i} \mathbf{J}(u_h) \cdot \mathbf{n}\, d\Gamma_i = \int_{\Omega_i} s(u_h)\, d\Omega_i, \quad \text{for} \quad i = 1, \cdots, N. \tag{2.9}$$

If the computation of $\mathbf{J}(u_h)$ at an interface between two cells is made consistently for both neighbouring cells, then the discretisation satisfies a discrete conservation law. This means that for an arbitrary union of cells $G = \cup\Omega_i$ hold:

$$\int_G \text{div}\ \mathbf{J}(u_h)\, d\Omega = \int_G s(u_h)\, d\Omega\,.$$

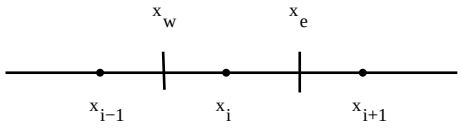


Figure 2.9: Nodal points and boxes on the real line

2.3 Examples

2.3.1 The diffusion equation

To show the principle of discretisation, in this section we discretise a steady one-dimensional diffusion equation, with a source-term s and a variable diffusion coefficient k,

$$-\frac{d}{dx}\left(k(x)\,\frac{d}{dx}u\right) = s(x,u), \quad \text{on} \quad \Omega \subset \mathbb{R}, \tag{2.10}$$

together with a boundary condition (e.g. $u = 0$, on $\partial\Omega$). This may be seen as the description of heat-conduction through a bar, where $u(x)$ denotes the temperature, $k(x)$ the heat diffusion coefficient, and $s(x,u)$ the heat generation per unit length, by some source.

We divide Ω in (unequal) boxes Ω_i and we choose for each Ω_i a nodal point x_i (see Figure 2.9).

We may choose the box-walls halfway the nodal points, or we may put the nodal points in the midpoint of the boxes. (Such choice is not relevant at the moment.)

The discrete approximation is represented by the function values at the nodal points $u_i \approx u(x_i)$. Integration of the equation over Ω_i yields

$$\int_{x_w}^{x_e} s(x,u)dx = -\int_{x_w}^{x_e} (ku_x)_x dx = -\,k(x)u_x(x)|_{x_w}^{x_e}\,. \tag{2.11}$$

To set up the discrete equations we need an approximation of $u(x)$ over Ω_i, and approximations $u_x(x_e)$ and $u_x(x_w)$. Starting from $u_i \approx u(x_i)$ we can approximate u, e.g. by
i) taking a constant value u_i on Ω_i;
ii) taking a piecewise linear approximation between the nodal points.
With (i) we cannot approximate the r.h.s. of (2.11), but with (ii) we can. Then we obtain

$$k(x_e)\frac{u_{i+1}-u_i}{x_{i+1}-x_i} - k(x_w)\frac{u_i-u_{i-1}}{x_i-x_{i-1}} = -\int_{x_w}^{x_e} s(x,u_h)\,dx = -\overline{s_i}(u_h)\,(x_e-x_w)\,, \tag{2.12}$$

where $\overline{s_i}$ is the mean value of $s(x,u)$ over the volume Ω_i.

2.3.2 The source term and solution of the system

If s is independent of u, or if s depends linearly on u, then (2.12) is a linear system of equations that can be solved immediately (as soon as the values of the endpoints u_0 and u_n are given) if we take $s_i(u_h) = s(u_i)$.

On the other hand, if $s = s(u)$ is a nonlinear function of u, then (i) $s_i(u_h)$ should be approximated, and (ii) the equation has to be solved iteratively. The source term $\overline{s_i}$ can **either** be approximated by a constant $\overline{s_i} := s_i(u_h^{(n)})$, where the previous approximation $u_h^{(n)}$ is used directly to obtain a (hopefully better) approximation $u_h^{(n+1)}$ by (2.12), **or** by *linearisation*, i.e. taking a linear approximation of s as $\overline{s_i} = s_i^C(u_h^{(n)}) + s_i^L(u_h^{(n)})(u_i^{(n+1)} - u_i^{(n)}$. In this case we have to adapt (2.12) slightly.

The difference between (i) and (ii) is that in (i) $\overline{s_i}$ depends completely on the previous approximation, whereas in (ii) linear dependence of s on u is taken into account; s_i^C and s_i^L are constants, that have to be determined for each nodal point. This makes that the linear system to solve is changed:

$$\begin{aligned} \tfrac{k(x_e)}{x_{i+1}-x_i}u_{i+1} - \left(\tfrac{k(x_e)}{x_{i+1}-x_i} + \tfrac{k(x_w)}{x_i-x_{i-1}} - (x_e-x_w)s_i^L(u_h^{(n)})\right)u_i + \\ + \tfrac{k(x_w)}{x_i-x_{i-1}}u_{i-1} = -(x_e-x_w)s_i^C(u_h^{(n)}) \end{aligned} \tag{2.13}$$

where s_i^L and s_i^C are such that $s_i^L(u_h)u_i + s_i^C(u_h) = \overline{s_i}(u_h)$. Here, the process reads: first make an initial estimate for u_h, then determine $s_i^C(u_h)$, and $s_i^L(u_h)$, and solve the linear system (2.13). This process is repeated until convergence is reached.

The final solution of the two processes (i) and (ii) is the same if

$$s_i(u_h) = s_i^C(u_h) + s_i^L(u_h)\, u_i\,.$$

An advantage of the alternative (ii) is that, by a judicious choice of s_i^L we may expect that the iterative process (ii) converges faster than process (i). The processes are identical if we take $s_i^L = 0$ for all i. However, if s is a linear function of u, and if we know $s^L(u) = ds/du$, then process (ii) may converge in one step.

We see that by (2.12) or (2.13) u_i is only coupled with u_{i-1} nd u_{i+1}, the values in the neighbouring cells. Hence the linear system is tridiagonal. This band structure of the matrix makes that the linear system can be solved very efficiently.

In the construction of (2.12) we have used the simplest possible representation for u_h that still allows an approximation of (ku_x). Other representations are possible as well. Apparently, it is not necessary to be consistent in the choice for u_h, (i.e. in the assumptions for the approximation of u) when the different parts in the differential operator are discretised. E.g. we could take $\overline{s_i} = s(u_i)$ piecewise constant, whereas a piecewise linear approximation was used to determine u_x.

2.3.3 The convection equation

Now we consider the problem

$$\frac{\partial}{\partial x}(v(x)u(x)) = s(u,x), \tag{2.14}$$

where v(x) is a given function. As a boundary condition we use e.g. $u = 0$ at the left boundary of the interval. Integration, analogous to (2.11) gives

$$\int_{x_w}^{x_e} s(u,x)dx = \int_{x_w}^{x_e} (vu)_x dx = v(x)u(x)|_{x_w}^{x_e}\,. \tag{2.15}$$

We take the same grid as in (2.3.1) and, again, a piecewise linear approximation for u_h. This yields the discrete equation

$$v(x_e)\frac{u_{i+1}+u_i}{2} - v(x_w)\frac{u_i+u_{i-1}}{2} = \overline{s_i}(u_h)\,(x_e - x_w)\,. \tag{2.16}$$

For a constant $v(x) = v$, this leads to

$$\frac{v}{2}(u_{i+1} - u_{i-1}) = \overline{s_i}(u_h)\,(x_e - x_w)\,. \tag{2.17}$$

Here we see that an important problem appears. In the simplest case of a constant v and s, we see that in equation (2.16) u_{i+1} is coupled with u_{i-1} only. The approximations of u at the even nodes are coupled to each other, and also the values at the odd nodes, but there is no coupling between the odd and the even nodes. If the problem is defined for $0 \le i \le N$, and the boundary condition is given at the inflow boundary x_0 ($i = 0$), then there is no boundary condition to determine the values at the odd nodes. Because the linear system is singular there is a family of possible solutions. The homogeneous equation allows nontrivial solutions. This is also called instability: the discrete solution is not bounded by the rhs of the discrete equation.

Notice that the technique for the discretisation of the equations in the interior domain cannot be used at the outflow boundary. There a different discretisation method should be used, because no value u_n is available. In fact it is this particular discretisation at the outflow boundary that determines the behaviour of the discrete solution at the odd points. We consider this completely intolerable.

2.4 Techniques for time-discretisation

2.4.1 Time-space discretisation or semi-discretisation

We have seen that (because the time-direction is unique) it makes sense to treat the time-dependence of a problem different from the space-dependence. The discrete solution is determined in a number of subsequent time-steps. Usually one uses the same time-steps for all nodal points in a grid. Thus, in a 2-dimensional

finite difference discretisation, the solutions are represented by $\{u_{ij}^n\}$, where (i,j) denotes the nodal point x_{ij} and n the time-coordinate t_n.

By the special character of the time-variable (the future depends on the past, and not the past on the future) the values $\{u_{ij}^n\}$ are computed on basis of the known values $\{u_{ij}^k\}$, $k = n-1, n-2, \cdots$.

Another possibility for discretisation of time-dependent PDEs is to defer the time-discretisation and to apply semi-discretisation: the solution of the equations is represented by $\{u_{ij}(t)\}$. The value of the solution at the point x_{ij} (or another coefficient used for the space-discretisation) is considered as a function in time. Then, the semi-discretised PDE leads to a very large system of ODEs for the unknown functions $\{u_{ij}(t)\}$, as e.g. in

$$\frac{d}{dt}u_{ij}(t) = F_{ij}(\{u_{kl}(t)\}_{kl}), \quad \forall i,j.$$

For the solution of these equations any suitable technique for the solution of ODEs can be used.

2.4.2 FDM for a linear hyperbolic problem

In this section we will show and analyse a number of characteristic problems that are encountered in the treatment of time-dependent problems. We do this by means of the simple convection equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}f(u) = 0. \tag{2.18}$$

As a grid we simply take a set of equidistant nodal point $\{x_i\}$. Halfway the nodal points we situate the cell interfaces of the cell Ω_i at $x_{i\pm\frac{1}{2}}$. Integration of (2.18) over Ω_i yields

$$\frac{\partial}{\partial t}\int_{\Omega_i} u\,dx + f(u(x))\,\Big|_{x=x_{i-\frac{1}{2}}}^{x=x_{i+\frac{1}{2}}} = 0\,. \tag{2.19}$$

We represent the discrete approximation by $\{u_i\}$, with u_i the mean value of u over the cell Ω_i. Then we arrive at the equation

$$(x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}})\,\frac{\partial}{\partial t}u_i(t) + f_{x_{i+\frac{1}{2}}} - f_{x_{i-\frac{1}{2}}} = 0\,. \tag{2.20}$$

Here, $f_{x_{i\pm\frac{1}{2}}}$ denotes an approximation for $f(u(x_{i\pm\frac{1}{2}}))$.

For the space discretisation we use *upwind, downwind* and *central* flux computations. To construct a system of equations for the variables $\{u_i\}$ we still have to decide how we relate $f(u(x_{i\pm\frac{1}{2}}))$ to the values of $\{u_i\}$. As in section 2.3.3, one possible choice is

$$f_{x_{i+\frac{1}{2}}} = \frac{1}{2}(f(u_i) + f(u_{i+1})),$$

the central approximation. Other choices are e.g.

$$f_{x_{i+\frac{1}{2}}} = f(u_i) \quad \text{or} \quad f_{x_{i+\frac{1}{2}}} = f(u_{i+1}),$$

respectively the backward and the forward discretisation[4].

For convenience we will combine the three choices in one formula, with a parameter w. After some analysis we make a choice for w. We write

$$f_{i+\frac{1}{2}} = wf(u_i) + (1-w)f(u_{i+1}), \quad i = 1, 2, 3, \cdots. \tag{2.21}$$

Now we have completed our semi-discretisation:

$$\frac{\partial}{\partial t}u_i(t) = \frac{w}{\Delta x}\, f(u_{i-1}) + \frac{1-2w}{\Delta x}\, f(u_i) - \frac{1-w}{\Delta x}\, f(u_{i+1}), \tag{2.22}$$

where $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$. We write this equation briefly as

$$\frac{\partial}{\partial t}u_h(t) = g_h(u_h(t)). \tag{2.23}$$

We realise the time discretisation of (2.23) simply by time-stepping with steps Δt. This can be done e.g. by the *forward Euler method*:

$$\frac{1}{\Delta t}\left[u_h(t^{n+1}) - u_h(t^n)\right] = g_h(u_h(t^n)). \tag{2.24}$$

In this case the new values $u_h(t^{n+1})$ are expressed explicitly in the old values $u_h(t^n)$.

An alternative possibility is the *backward Euler method*:

$$\frac{1}{\Delta t}\left[u_h(t^{n+1}) - u_h(t^n)\right] = g_h(u_h(t^{n+1})), \tag{2.25}$$

or also the *Cranck-Nicolson method*

$$\frac{1}{\Delta t}\left[u_h(t^{n+1}) - u_h(t^n)\right] = (\ g_h(u_h(t^n)) + g_h(u_h(t^{n+1}))\ )/2. \tag{2.26}$$

These two methods are "implicit" time-discretisation methods. Here we have to solve a system of equations to compute the new values $u_h(t^{n+1})$.

We summarise the three possibilities, again, in one parametrised formula:

$$\frac{1}{\Delta t}\left[u_h(t^{n+1}) - u_h(t^n)\right] = (1-\theta)g_h(u_h(t^n)) + \theta g_h(u_h(t^{n+1}))\,. \tag{2.27}$$

The fully discretised system reads:

$$\begin{array}{lll} u_i^{n+1} - u_i^n = & \frac{\Delta t}{\Delta x}\theta & \{wf_{i-1}^{n+1} + (1-2w)f_i^{n+1} - (1-w)f_{i+1}^{n+1}\}+ \\ & \frac{\Delta t}{\Delta x}(1-\theta) & \{wf_{i-1}^{n} + (1-2w)f_i^{n} - (1-w)f_{i+1}^{n}\}, \end{array} \tag{2.28}$$

[4]With $\partial f/\partial u > 0$, the approximation $f_{x_{i+\frac{1}{2}}} = f(u_i)$ is called upwind approximation, because the flow of information is from left to right and for the approximation of $f_{x_{i+\frac{1}{2}}}$ the upstream value u_i is used. The other approximation is called downstream approximation.

where f_j^k represents $f(u_j^k)$.

In general, to compute u_i^{n+1} from u_i^n, we have to solve a system of nonlinear equations. Because it is difficult to analyse nonlinear equations we simplify the problem by linearising it. We set $\frac{df}{du} = f'$ constant, and we combine a few parameters in a single: $\lambda = f'\Delta t/\Delta x$. Now equation (2.28) becomes

$$u_i^{n+1} - u_i^n = \begin{array}{ll} \lambda\theta & \{wu_{i-1}^{n+1} + (1-2w)u_i^{n+1} - (1-w)u_{i+1}^{n+1}\} + \\ \lambda(1-\theta) & \{wu_{i-1}^{n} + (1-2w)u_i^{n} - (1-w)u_{i+1}^{n}\}. \end{array} \tag{2.29}$$

Having simplified the equation so far, it is simple to solve it also analytically. We will use this knowledge about the analytic solution to check if our numerical techniques yield reliable results.

The equation now is $u_t + f'u_x = 0$, and the solution is $\psi(x - f't)$ for an arbitrary initial solution $\psi(x) = u(0, x)$. In time the solution doesn't really change, it doesn't grow or shrink, it only moves with the velocity f'. With the different possible choices for the numerical method, we can see which properties of the true solution are inherited by the numerical counterparts.

In order to study the behaviour of the solution of the discrete system, we consider the behaviour in time of the approximate solution in the domain $\Omega = (-\infty, +\infty)$, for an initial function $u(t_0, x) = e^{j\omega x}$. We use the notation $j := \sqrt{-1}$ in order not to interfere with the index i previously used. (We refer to a textbook in Fourier analysis to understand the relevance of this choice.) The discrete initial function, then, is

$$u_i^0 = e^{j\omega hi}.$$

Using equation (2.29) we see immediately that the solution at a later stage is given by

$$u_i^n = \gamma^n\, e^{j\omega hi}, \tag{2.30}$$

where γ is a complex number that satisfies

$$\gamma - 1 = \begin{array}{ll} \lambda\theta & \{w\gamma e^{-j\omega h} + (1-2w)\gamma - (1-w)\gamma e^{+j\omega h}\} + \\ \lambda(1-\theta) & \{we^{-j\omega h} + (1-2w) - (1-w)e^{+j\omega h}\}. \end{array} \tag{2.31}$$

This can also be written as

$$\begin{aligned} \gamma &= \frac{1 + \lambda(1-\theta)\{(1-2w) + we^{-j\omega h} - (1-w)e^{+j\omega h}\}}{1 - \lambda\theta\{(1-2w) + we^{-j\omega h} - (1-w)e^{+j\omega h}\}} \\ &= \frac{1 + \lambda(1-\theta)\{(1-2w)(1-\cos(\omega h)) - j\sin(\omega h)\}}{1 - \lambda\theta\{(1-2w)(1-\cos(\omega h)) - j\sin(\omega h)\}}; \end{aligned} \tag{2.32}$$

γ is the *amplification factor* for the time-integration: $|u_i^n| = |\gamma|^n$. The solutions increases for $t \to \infty$ if $|\gamma| > 1$, and it decreases for $|\gamma| < 1$. We see that

$$|\gamma|^2 = \frac{\{1 + \lambda(1-\theta)(1-2w)(1-\cos(\omega h))\}^2 + \lambda^2(1-\theta)^2\sin^2(\omega h)}{\{1 - \lambda\theta(1-2w)(1-\cos(\omega h))\}^2 + \lambda^2\theta^2\sin^2(\omega h)}. \tag{2.33}$$

Now we want to study the effect of the parameter w, that was introduced in the space discretisation, and of θ that was introduced in the time discretisation. Further we can see what is the effect of the parameter λ.

- If $w = \frac{1}{2}$ (i.e. for central discretisation in space) equation(2.33) reduces to

$$|\gamma|^2 = \frac{1 + \lambda^2(1-\theta)^2 \sin^2(\omega h)}{1 + \lambda^2\theta^2 \sin^2(\omega h)}\,, \tag{2.34}$$

  and we find

$$|\gamma| < 1 \Leftrightarrow (1-\theta)^2 \leq \theta^2 \Leftrightarrow \theta \geq \frac{1}{2}.$$

  Depending on θ we can draw the following conclusion:
  i) $\theta = \frac{1}{2} \Leftrightarrow |\gamma| = 1$. This implies neutral stability (Crank-Nicolson).
  ii) $\theta = 0 \Leftrightarrow |\gamma| > 1$. This means unconditional instability (forward Euler).
  iii) $\theta = 1 \Leftrightarrow |\gamma| < 1$. This means unconditional stability (backward Euler).

- If $w = 1$ (i.e. for "backward" discretisation in space)
  stability $\Leftrightarrow |\gamma| \leq 1 \Leftrightarrow \lambda(\lambda(1-2\theta) - 1) \leq 0$. Depending on the sign of f' we find:

  - $f' > 0 \Leftrightarrow \lambda > 0 \Leftrightarrow$ stability for: $\lambda(1-2\theta) \leq 1$.
    This implies stability for $\theta \geq \frac{1}{2}$. If $\theta < \frac{1}{2}$ stability may depend on λ; stability is obtained for $\lambda \leq \frac{1}{1-2\theta}$ and we obtain stability only for λ (i.e. the time-step) small enough.
  - $f' < 0 \Leftrightarrow \lambda < 0 \Leftrightarrow$ stability for: $\lambda(1-2\theta) \geq 1$.
    This implies instability for $\theta \in [0, \frac{1}{2}]$. If $\theta > \frac{1}{2}$ then stability is guaranteed for $|\lambda| \geq \frac{1}{2\theta-1}$ and we obtain stability only for λ large (!) enough.

- If $w = 0$ (i.e. for "forward" discretisation in space)
  stability $\Leftrightarrow |\gamma| \leq 1 \Leftrightarrow \lambda(\lambda(1-2\theta) + 1) \leq 0$

  - $f' > 0 \Leftrightarrow \lambda > 0 \Leftrightarrow$ stability for $\lambda(1-2\theta) \leq 1$.
    This implies instability for $\theta \in [0, \frac{1}{2}]$; if $\theta > \frac{1}{2}$ then get stability for $|\lambda| \geq \frac{1}{2\theta-1}$, i.e. only for λ large(!) enough.
  - $f' < 0 \Leftrightarrow \lambda < 0$; this implies stability for $\theta \geq \frac{1}{2}$; if $\theta \in [0, \frac{1}{2}]$ we get stability for $|\lambda| \leq \frac{1}{1-2\theta}$, i.e. we obtain stability for λ small enough.

We recognise a symmetry If $f' > 0$ then $w = 1$ corresponds with upwind discretisation and $w = 0$ with downwind; with $f' < 0$ it is the other way. We notice that the upwind discretisation is unconditionally stable when $\theta \geq \frac{1}{2}$. For $\theta < \frac{1}{2}$ it is stable if the time-step is small enough. The downwind discretisation is unstable for all reasonable cases.

2.4.3 The equivalent differential equation

In this section we study the local discretisation error (i.e. the amount to which the true solution -at the nodal points- does not satisfy the discrete equation [5])

[5] For the differential operator $L : X \to Y$ and its discrete operator $L_h : X_h \to Y_h$, and the restriction operators $R_h : X \to X_h$ and $\overline{R}_h : Y \to Y_h$, we remember the definitions for *global*

for (2.28). We consider again the linear problem

$$\frac{\partial}{\partial t}u + f'\frac{\partial}{\partial x}u = 0, \tag{2.35}$$

and its discretisation equation (2.28). We can write (2.29) also as

$$\begin{aligned} u_i^{n+1} - u_i^n = \quad & -\lambda\theta\{\tfrac{1}{2}(u_{i+1}^{n+1} - u_{i-1}^{n+1}) + (\tfrac{1}{2} - w)(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1})\} + \\ & -\lambda(1-\theta)\{\tfrac{1}{2}(u_{i+1}^{n} - u_{i-1}^{n}) + (\tfrac{1}{2} - w)(u_{i+1}^{n} - 2u_i^{n} + u_{i-1}^{n})\}. \end{aligned} \tag{2.36}$$

In order to compute the local truncation error, we expand the function $u(x,t)$ in a Taylor series around the point $x = x_i$, and $t_\star = t_n + \theta\Delta t$. We denote $u_\star := u(x_i, t_\star)$. We obtain

$$\begin{aligned}
&(i) && u_i^n = u_\star - \theta\Delta t\left(\tfrac{\partial u}{\partial t}\right)_\star + \tfrac{1}{2}\theta^2\Delta t^2\left(\tfrac{\partial^2 u}{\partial t^2}\right)_\star + \mathcal{O}(\Delta t^3);\\
&(ii) && u_i^{n+1} = u_\star + (1-\theta)\Delta t\left(\tfrac{\partial u}{\partial t}\right)_\star + \tfrac{1}{2}(1-\theta)^2\Delta t^2\left(\tfrac{\partial^2 u}{\partial t^2}\right)_\star + \mathcal{O}(\Delta t^3);\\
&(iii) && u_{i-1} = u_i - \Delta x\left(\tfrac{\partial u}{\partial x}\right)_i + \tfrac{1}{2}\Delta x^2\left(\tfrac{\partial^2 u}{\partial x^2}\right)_i + \mathcal{O}(\Delta x^3);\\
&(iv) && u_{i+1} = u_i + \Delta x\left(\tfrac{\partial u}{\partial x}\right)_i + \tfrac{1}{2}\Delta x^2\left(\tfrac{\partial^2 u}{\partial x^2}\right)_i + \mathcal{O}(\Delta x^3);\\
&(v) && \tfrac{1}{2}(u_{i+1} - u_{i-1}) = +\Delta x\left(\tfrac{\partial u}{\partial x}\right)_i + \mathcal{O}(\Delta x^3);\\
&(vi) && (u_{i+1} - 2u_i + u_{i-1}) = \Delta x^2\left(\tfrac{\partial^2 u}{\partial x^2}\right)_i + \mathcal{O}(\Delta x^3);\\
&(vii) && \left(\tfrac{\partial u}{\partial x}\right)_i^n = \left(\tfrac{\partial u}{\partial x}\right)_\star - \theta\,\Delta t\left(\tfrac{\partial^2 u}{\partial x\partial t}\right)_\star + \mathcal{O}(\Delta t^2);\\
&(viii) && \left(\tfrac{\partial u}{\partial x}\right)_i^{n+1} = \left(\tfrac{\partial u}{\partial x}\right)_\star + (1-\theta)\,\Delta t\left(\tfrac{\partial^2 u}{\partial x\partial t}\right)_\star + \mathcal{O}(\Delta t^2);\\
&(ix) && (1-\theta)\left(\tfrac{\partial u}{\partial x}\right)_i^n + \theta\left(\tfrac{\partial u}{\partial x}\right)_i^{n+1} = \left(\tfrac{\partial u}{\partial x}\right)_\star + \mathcal{O}(\Delta t^2);\\
&(x) && (1-\theta)\left(\tfrac{\partial^2 u}{\partial x^2}\right)_i^n + \theta\left(\tfrac{\partial^2 u}{\partial x^2}\right)_i^{n+1} = \left(\tfrac{\partial^2 u}{\partial x^2}\right)_\star + \mathcal{O}(\Delta t^2).
\end{aligned} \tag{2.37}$$

Substitution of this in (2.36) yields for the left-hand-side

$$\begin{aligned} & u_i^{n+1} - u_i^n \\ & = \Delta t\left(\tfrac{\partial u}{\partial t}\right)_\star + \tfrac{1}{2}(1-2\theta)\Delta t^2\left(\tfrac{\partial^2 u}{\partial t^2}\right)_\star + \mathcal{O}(\Delta t^3) = \end{aligned}$$

and for the right-hand-side

$$\begin{aligned} & -\lambda\theta\left\{\Delta x\left(\tfrac{\partial u}{\partial x}\right)_i^{n+1} + (\tfrac{1}{2} - w)\Delta x^2\left(\tfrac{\partial^2 u}{\partial x^2}\right)_i^{n+1} + \mathcal{O}(\Delta x^3)\right\} \\ & -\lambda(1-\theta)\left\{\Delta x\left(\tfrac{\partial u}{\partial x}\right)_i^{n} + (\tfrac{1}{2} - w)\Delta x^2\left(\tfrac{\partial^2 u}{\partial x^2}\right)_i^{n} + \mathcal{O}(\Delta x^3)\right\} \\ & = -\lambda\left\{\Delta x\left[\left(\tfrac{\partial u}{\partial x}\right)_\star + \mathcal{O}(\Delta t^2)\right] + (\tfrac{1}{2} - w)\Delta x^2\left[\left(\tfrac{\partial^2 u}{\partial x^2}\right)_\star + \mathcal{O}(\Delta t^2)\right]\right\} + \mathcal{O}(\Delta x^3)\,. \end{aligned}$$

discretisation error

$$|u(x_i) - u_i| = |(R_h u)_i - u_i|,$$

and for the *local discretisation error*

$$|(L_h R_h u - \overline{R}_h L u)_i|.$$

Making use of the parameter $\lambda = f'\Delta t/\Delta x$ we get the expression for the truncation error (in equation (2.36) $\tau_\star(u) = \text{rhs} - \text{lhs}$)

$$\begin{aligned}\frac{\tau_\star(u)}{\Delta t} &= -\left(\frac{\partial u}{\partial t}\right)_\star - (\frac{1}{2}-\theta)\Delta t\left(\frac{\partial^2 u}{\partial t^2}\right)_\star + \mathcal{O}(\Delta t^2)\\ &\quad - f'\left\{\left(\frac{\partial u}{\partial x}\right)_\star + \mathcal{O}(\Delta t^2) + (\frac{1}{2}-w)\Delta x\left(\frac{\partial^2 u}{\partial x^2}\right)_\star + \mathcal{O}(\Delta x^2)\right\}\\ &= (\theta-\frac{1}{2})\Delta t\left(\frac{\partial^2 u}{\partial t^2}\right)_\star + f'(w-\frac{1}{2})\Delta x\left(\frac{\partial^2 u}{\partial x^2}\right)_\star + \mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta t^2)\,.\end{aligned}$$

This means that the method is first order consistent in time and space. If $\theta = \frac{1}{2}$ it is second order in time, and if $w = \frac{1}{2}$ it is second order in space.

Differentiating the PDE (2.35) we find that $u_{tt} = (f')^2 u_{xx}$, so that the principal term in the truncation error can be written as

$$f'\left[(\theta-\frac{1}{2})f'\Delta t + (w-\frac{1}{2})\Delta x\right]u_{xx} = \tag{2.38}$$

$$f'\Delta x\left[(\theta-\frac{1}{2})\lambda + (w-\frac{1}{2})\right]u_{xx} \;=\; \epsilon_{\text{num}}u_{xx}\,. \tag{2.39}$$

So, we may see the difference scheme as a second order discretisation of a *modified equation* (or the *equivalent differential equation*)

$$u_t = f'\Delta x\left[(\theta-\frac{1}{2})\lambda + (w-\frac{1}{2})\right]u_{xx} - f'u_x\,. \tag{2.40}$$

For this reason ϵ_{num} is also called the *numerical diffusion* of the difference scheme.

Remark:
Notice that the stability of the difference scheme as studied in Section 2.4.2 corresponds exactly with a *positive* numerical diffusion!

The numerical diffusion, however, can cause problems in a convection diffusion problem if the diffusion coefficient in the equation is (much) smaller than the numerical diffusion. The numerical diffusion may then overrule the 'physical' diffusion, and the discrete solutions are much more smeared than the true solution of the differential equation.

If we make computations with one of the above schemes, we notice:
(i) the numerical diffusion can have a significant effect;
(ii) the second order scheme (without numerical diffusion) shows an "overshoot", i.e in the numerical solutions maxima or minima appear that do not exist in the true solution;
(iii) unstable scheme are useless.

Example 2.4.1
We solve the equation

$$u_t = 0.01\, u_{xx} - u_x$$

on the domain $t \geq 0$, $x \geq 0$, with initial condition $u = 0$ at $t = 0$, and boundary condition $u = 1$ at $x = 0$ (at the "inflow" boundary). We discretise the term u_{xx} by

$$u_{xx} \approx \frac{1}{2}(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}) + \frac{1}{2}(u_{i+1}^{n} - 2u_i^{n} + u_{i-1}^{n})\,.$$

The other terms are discretised by (2.29), and we solve the problem for different values of w and θ. In the Figures 2.10 and 2.11 we show the results for $t = 1$ (and in Figure 2.12 for $t = 0.5$). We take a fixed $\Delta x = 0.01$ and either $\Delta t = 0.05$ ($\lambda = 0.5$) or $\Delta t = 0.1$ ($\lambda = 1.0$). For comparison the true solution is also shown, as well as the solution of the modified equation

$$u_t = (0.01 + \epsilon_{\text{num}})u_{xx} - u_x\,.$$

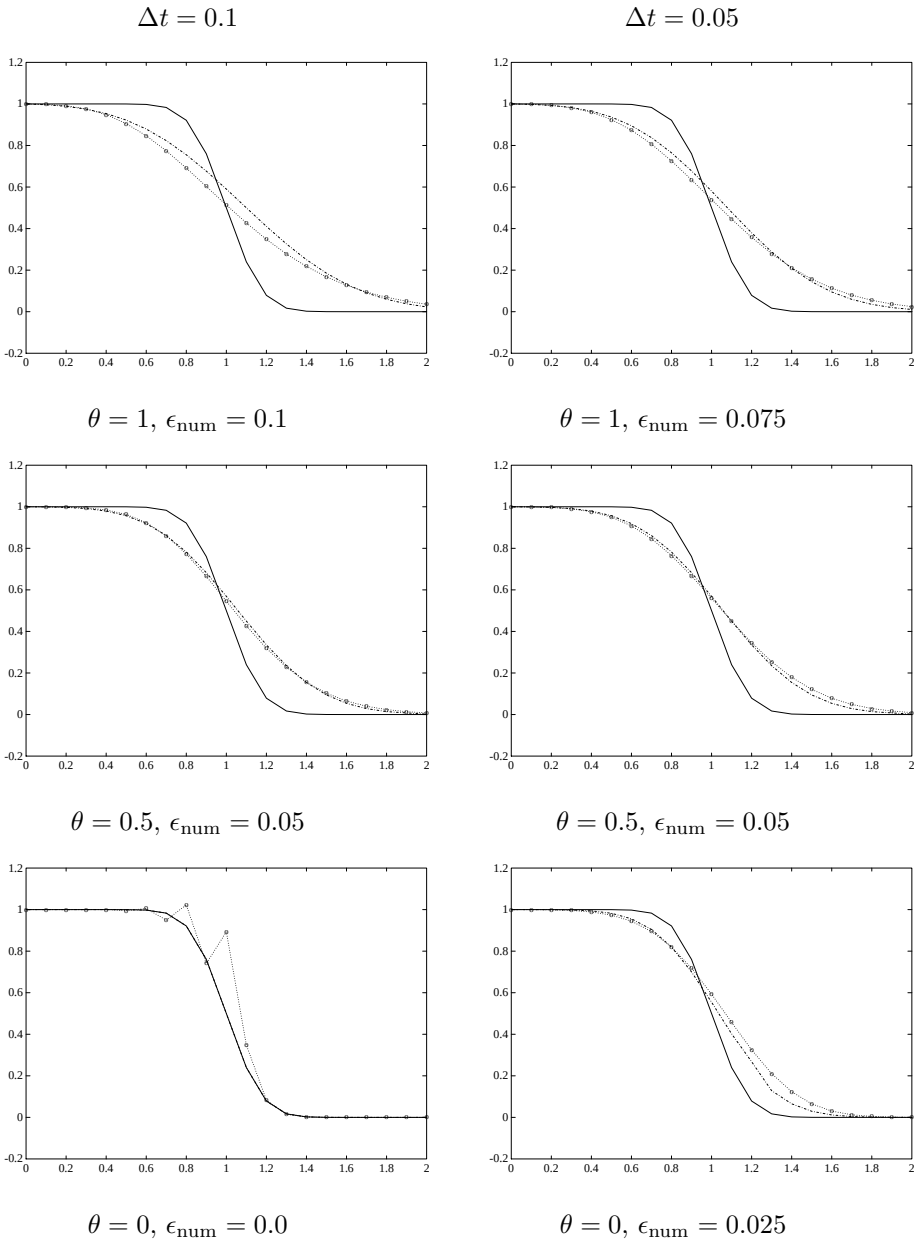


Figure 2.10: Numerical solution of the convection diffusion equation
In all figures the numerical solution is given for $t = 1$; The true sulution is given by the solid line; the dashed line represents the solution of the modified equation. The dotted line is the numerical approximation; obtained with upwind space discretisation: $w = 1.0$, $\Delta x = 0.1$.

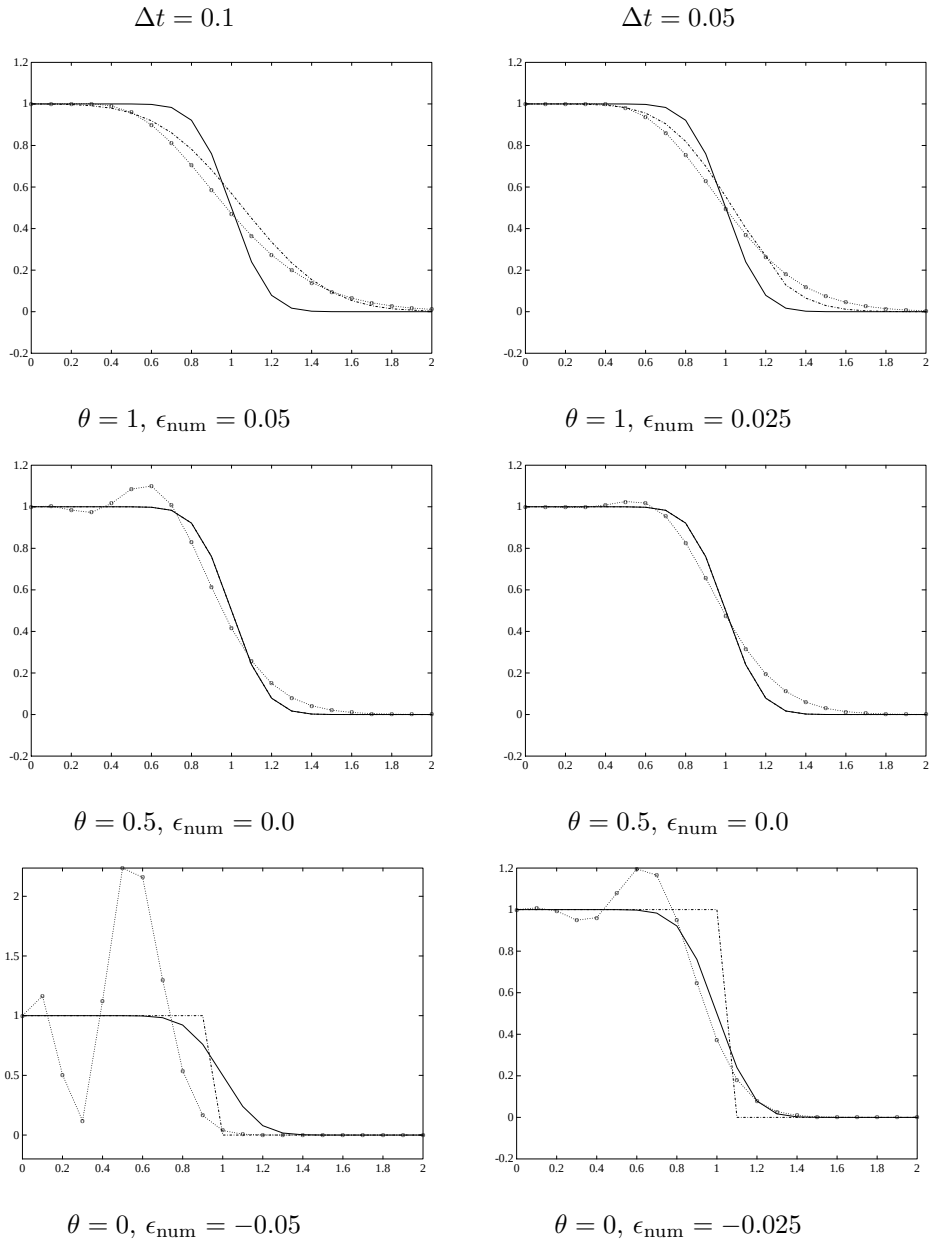
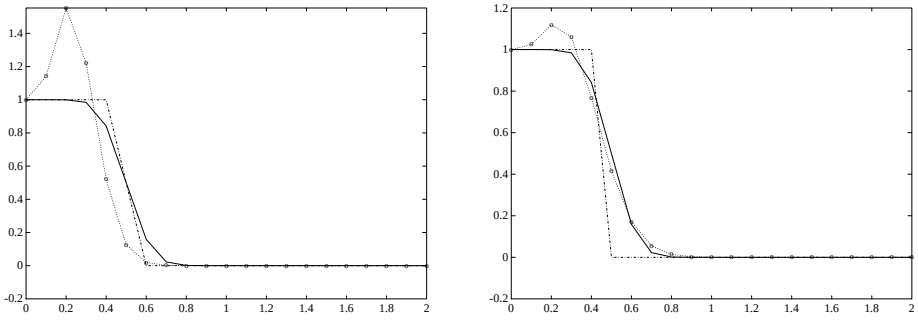


Figure 2.11: Numerical solution of the convection diffusion equation
In all figures the numerical solution is given for $t = 1$; The true sulution is given by the solid line; the dashed line represents the solution of the modified equation. The dotted line is the numerical approximation; obtained with central space discretisation: $w = 0.5$, $\Delta x = 0.1$.



$\Delta t = 0.1$, $\epsilon_{\text{num}} = -0.05$ $\Delta t = 0.05$, $\epsilon_{\text{num}} = -0.025$

Figure 2.12: Numerical solution of the convection diffusion equation
In these figures the numerical solution is given for $t = 0.5$; The true sulution is given by the solid line; the dashed line represents the solution of the modified equation. The dotted line is the numerical approximation; obtained with central space discretisation ($w = 0.5$), and forward Euler time-discretisation ($\theta = 0$), $\Delta x = 0.1$.

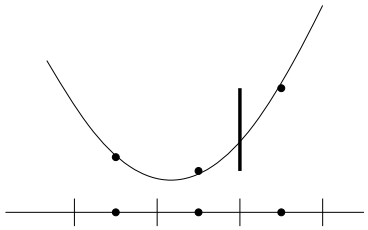


Figure 2.13: Non-conservative interpolation

2.5 Criteria for good discretisation methods

In this section we want to summarise a set of rules and give criteria that can help us to find good discretisation methods. The last three rules are taken from [19].

Rule 1. *If the number of nodal points (coefficients for the representation) increases we may expect (and we should require) that the discrete solutions converge to a true solution of the continuous problem.*

This property of a discretisation is usually mathematically formulated by an asymptotic statement: if the mesh-width h of the discretisation is sufficiently small, the discretisation error (in the solution) (in some norm) should vanish (at some rate) as h tends to zero.

Rule 2 *Often we wish more than only an asymptotic validity. We wish that a reasonably accurate solution is already obtained for a reasonably fine (= reasonably coarse) grid. We do not want to construct an extremely fine grid before we can be sure about the value of the approximation.*

This requirement is more difficult to formulate mathematically. Often it is expressed that we want to obtain a "physically realistic" solution. I.e. one requires that the discrete solution -already for a coarse grid- shows the same "global character" as the continuous solution.

For example, if it is known that the continuous solution is positive or monotone, one often requires that also the discrete solution is positive or monotone. E.g. (i) if concentrations of a chemical are computed, one doesn't like to find negative concentrations by approximation errors; (ii) for the stationary solution of a heat-conduction problem without sources or sinks, one requires that the temperature as computed does not show local extrema; (iii) if the continuous equation describes a conservation law, one requires that also the computed quantity is not created or does not disappear in the computation.

For the class of (hyperbolic) problems in conservation form it is relatively simple to satisfy the requirement that the discrete equations *satisfy the conservation law*, not only globally, for all Ω, but also locally, for each cell Ω_e of the discretisation of Ω.

A class of (elliptic) problems for which the requirement of "physical relevance" is usually not too hard to satisfy, are those problems that can be formulated as a minimisation problem. For such *symmetric elliptic problems* a convex functional exists that is minimised by the solution of the PDE.

Often the functional has a physical meaning (e.g. the energy left in the system) and, therefore, the minimising element in a subspace satisfies automatically the requirement of physical relevance. Some discretisation methods (finite element methods) find this minimising approximate solution.

Rule 3: Obeying the conservation law. A conservation law is most reliably discretised if the discrete system satisfies the law on each sub-element of the discretisation separately. In order to satisfy the conservation law locally in a box method, the flux over box-interfaces should be computed consistently. I.e. the computed flux $\mathbf{J}(u) \cdot \mathbf{n}$ over the box interface should be the same when it represents the flow out of the one cell or when it represents the flow into the

other. The explanation is simple: the amount of the conserved quantity that leaves the one cell should reappear (exactly) in the other cell.

Example 2.5.1
If we use a three-point, parabolic approximation to compute the flux over the right boundary of cell Ω_i (fig.2.13) then the requirement of conservation is not fulfilled. If we take the two-point linear approximation the conservation requirement is satisfied.

Example 2.5.2
See Section 2.3.1. In equation (2.12) we have (correctly) approximated $\mathbf{J}(u)$ by

$$\mathbf{J}(u) = k(x_{i+\frac{1}{2}})\frac{u_{i+1} - u_i}{x_{i+1} - x_i} \ .$$

If in equation (2.12) we had approximated $k(x)$ by $k(x_i)$, then the discretisation would not have been conservative.

Rule 4: Trivial (constant) solutions should satisfy the discrete solution.

If a differential operator only contains derivatives of a dependent variable, then with an approximate solution u also $u +$ constant should be a solution of the discrete equations. I.e. the constant function should be in the kernel of the discrete operator. (The row-sum of the linearisation of the discrete operator should vanish.)

This rule can be seen as an instance for the more *general principle:* Seeking a good discretisation for a complex (intricate) operator, it makes sense to find a discretisation that -at least- satisfies basic requirements for most simplifications of the problem. (A good method for a problem is also a good method for any simplification of that problem.)

Rule 5: Guaranteed positivity of the solution. In order to guarantee the positivity of the solution (in the absence of sources and sinks) for the linearisation of the discrete operator, the diagonal and the off-diagonal elements of the Jacobian matrix should have different signs.

Explanation: If we write the linearised discrete equation for one space dimension as

$$a_i u_i = a_{i-1}u_{i-1} + a_{i+1}u_{i+1} + s_i \tag{2.41}$$

the coefficients a_i, a_{i-1} and a_{i+1} should have the same sign to guarantee that for all possible u_{i-1} and u_{i+1}. The same argument is simple to generalise for two or three dimensions.

For discretisations with a higher order of accuracy (than one) it is difficult (often impossible) to satisfy this requirement. For higher order discretisations more non-zero diagonals appear in the linearised operator.

The fifth rule allows an extension for the case that a source term is present. First, if the source term is neglected the discretisation should satisfy rule 5. Further, (ii) if there is a source term, it should not disturb the positivity (monotonicity).

Let $s(u)$ the source term in the equation $Lu = s(u)$ allow the linearisation $s(u) \approx s^C + s^L + u$, then the discretisation of $Lu = s(u)$ leads to (see sect (2.3.2))

$$(a_i - s_i^L)u_i = a_{i-1}u_{i-1} + a_{i+1}u_{i+1} + s_i^C \,,$$

then s_i^L and a_i should have opposite signs.

Chapter 3

Finite Element Methods

3.1 Introduction

This chapter is devoted entirely to the subject of Finite Elements. First we give a number of examples that also will be treated later in applications of the method. Then we give the theoretical foundation of the method. Subsequently we discuss how the method can be used in practice and finally we show how accurate the method is.

Excellent reference books on the subject are, e.g., [2] and [3].

3.2 Examples of boundary value problems

The finite element method is best applied to (partial) differential equations in variational form. This is best illustrated by means of elliptic equations, which are always of even order. In this section we discuss a number of typical examples we will frequently use. Although we will restrict ourselves in these lectures to the examples, in general we can apply the FEM also to systems of PDEs, nonlinear equations or problems that involve differentiation in time.

3.2.1 One-dimensional second order

The simplest problem we meet is the homogeneous Dirichlet two-point boundary-value problem (TPBVP) on an interval $\Omega = [a, b] \subset \mathbb{R}$:

$$-(a_2 u_x)_x + a_1 u_x + a_0 u = f,$$

$$u(a) = u(b) = 0.$$

For smooth functions a_2, a_1, a_0 and f, and $a_2 > 0$, it is known that this problem has a unique solution.

3.2.2 Two-dimensional second order

An example of a two-dimensional, second-order, elliptic equation is the steady-state convection-diffusion equation

$$\nabla \cdot (\mathbf{D}\nabla\psi) - \nabla \cdot (\psi \mathbf{v}) = f \ \text{ on } \ \Omega \subset \mathbb{R}^2, \tag{3.1}$$

where $\psi(x, y)$ is the unknown function (for instance the temperature or the density of some matter), $\mathbf{v}(x, y)$ is a (given) velocity field, $\mathbf{D}(x, y)$ a diffusion tensor, and $f(x, y)$ is the source function. Special examples are the *diffusion equation* ($\mathbf{v} = 0$), the *potential equation* ($\mathbf{v} = 0$, $\mathbf{D}$ is is the identity matrix) and the *Laplace equation* ($\mathbf{v} = 0$, $\mathbf{D}$ is is the identity matrix, $f = 0$). Because of the special difficulties in the numerical treatment of the convection diffusion equation, we will discuss it in more detail in a later chapter.

We write the linear two-dimensional second-order equation in its most general form:

$$a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy} + a_1u_x + a_2u_y + a_0u = f, \tag{3.2}$$

on a domain $\Omega \subset \mathbb{R}^2$, where the coefficients a_{ij}, a_i and f are functions of x and y.

We assume the equation to be *elliptic* on Ω, which means the coefficients of the *principle part* of (3.2) satisfy

$$a_{11}\xi_1^2 + 2a_{12}\xi_1\xi_2 + a_{22}\xi_2^2 \neq 0, \quad \forall \xi = (\xi_1, \xi_2) \neq (0, 0) \in \Omega. \tag{3.3}$$

It is easy to see that (3.3) is equivalent with $a_{12}^2 < a_{22}a_{11}$. The sign of a_{11} completely determines the sign of the left hand side of (3.3).

Define the coefficient matrix by

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix},$$

then (3.3) is equivalent with

$$\xi^T A\, \xi > 0, \quad \forall \xi \neq 0,$$

i.e. the matrix A is *positive definite*.

The elliptic equation (3.2) is defined on an open connected domain $\Omega \subset \mathbb{R}^2$, and to obtain a unique solution we have to provide *boundary conditions*. It appears that we need one condition along all the boundary $\Gamma := \partial\Omega$. This boundary condition can be of the form

$$u = g \text{ on } \Gamma, \tag{3.4}$$

or, more generally,

$$\frac{\partial u}{\partial \mathbf{n}} + b_1 \frac{\partial u}{\partial \mathbf{s}} + b_0 u = h \text{ on } \Gamma, \tag{3.5}$$

where $\mathbf{n}$ is the outward unit normal on Γ and $\mathbf{s}$ is the vector tangent to the boundary. Boundary conditions of type (3.4) are called *Dirichlet* conditions, boundary conditions of the form (3.5) *mixed* conditions. The special case when $b_0 = b_1 = 0$ is called *Neumann* condition.

Theorem 3.2.1 If the coefficients in equation (3.2) are taken constant, the principle part

$$a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy}$$

can be reduced to the form

$$\Delta u$$

by a linear coordinate transformation.

Proof: First we notice that the principle part of (3.2) can be written as $(\nabla, A\nabla u)$, where we use the notation $\nabla = (\partial/\partial x, \partial/\partial y)^T$. We also define $\nabla' = (\partial/\partial\xi, \partial/\partial\eta)^T$. The transformation we need is of the form:

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = T \begin{pmatrix} x \\ y \end{pmatrix}.$$

A direct consequence is $(\xi, \eta) = (x, y)\ T^T$. Also easy to check are

$$\begin{pmatrix} \partial/\partial\xi \\ \partial/\partial\eta \end{pmatrix} (\xi, \eta) = I,$$

and a similar expression for the other coordinates. Assuming the existence of a W so that $\nabla = W\nabla'$ we see

$$\begin{aligned} T^T &= \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \end{pmatrix} (x, y)\ T^T = \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \end{pmatrix} (\xi, \eta) \\ &= W \begin{pmatrix} \partial/\partial\xi \\ \partial/\partial\eta \end{pmatrix} (\xi, \eta) = W. \end{aligned}$$

Hence $\nabla = T^T\ \nabla'$, and we write the principle part as:

$$\begin{aligned} (\nabla, A\nabla u) &= (T^T\nabla', AT^T\nabla' u) \\ &= (\nabla', TAT^T\nabla' u), \end{aligned}$$

by the definition of adjoint. Because A is positive definite we can choose T such that $TAT^T = D$, with $D = \text{diag}\{\alpha_{11}, \alpha_{22}\}$, where α_{11}, α_{22} are positive. Because both α_{11} and α_{22} are positive we can reduce the principle part of the elliptic equation by a simple scaling to

$$\Delta u.$$

This completes the proof. ■

A transformation of equation (3.2) along the lines of the above theorem reduces the equation to

$$-\Delta u + \beta_1 u_x + \beta_2 u_y + \beta_0 u = f. \tag{3.6}$$

By an additional transformation of the dependent variable u we can remove the first order terms from (3.6). A substitution

$$u(x, y) = v(x, y) \cdot \exp\{\frac{1}{2}(\beta_1 x + \beta_2 y)\}$$

reduces (3.6) to

$$-\Delta v + [\beta_0 + \frac{1}{4}(\beta_1^2 + \beta_2^2)]v = f(x,y)\exp\{-\frac{1}{2}(\beta_1 x + \beta_2 y)\}.$$

Note that the right hand side of the equation is weighted by an exponential factor. In case of variable coefficients β_j, if the field $\vec{\beta}$ is rotation free (.e., if $\nabla \times \vec{\beta} = 0$, so that there exists a $B(x,y)$ such that $\vec{\beta} = \nabla B \equiv \text{grad } B$), we use the transformation

$$u(x,y) = v(x,y) \cdot \exp\{\frac{1}{2}\int \vec{\beta} \cdot d\vec{x}\}. \tag{3.7}$$

Inserting this in (3.6) gives

$$-\Delta v + [\beta_0 + \frac{1}{4}(\beta_1^2 + \beta_2^2) - \frac{1}{2}((\beta_1)_x + (\beta_2)_y)]v = f \cdot \exp\{-\frac{1}{2}\int \vec{\beta} \cdot d\vec{x}\}.$$

For constant β's this is the same as the previous reduced equation.

These computations show that any uniqueness/existence result we obtain for $-\Delta u + au = f$ will also be valid for the general case. To understand the existence and uniqueness of such problem, in Section **3.3** we will discuss parts of the theory on elliptic partial differential equations.

3.2.3 Fourth order problems

The results on existence and uniqueness of elliptic problems are also applicable to fourth order elliptic problems like the *biharmonic problem*

$$\begin{aligned} \Delta^2 u &= f \quad \text{in} \quad \Omega, \\ u &= 0 \quad \text{on} \quad \Gamma, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on} \quad \Gamma. \end{aligned}$$

To this form we may reduce the Stokes problem in fluid dynamics and it also models the displacement of a thin elastic plate, clamped at its boundary, under a transversal load.

We may consider the one-dimensional homogeneous biharmonic problem on $[a,b] \subset \mathbb{R}$, e.g.,

$$\begin{aligned} a_2 u_{xxxx} - a_1 u_{xx} + a_0 u &= f, \\ u(a) = u(b) = u'(a) = u'(b) &= 0. \end{aligned}$$

3.3 Abstract Elliptic Theory

A powerful tool to prove the existence and uniqueness of solutions of PDEs of elliptic type is the (generalised) Theorem of Lax-Milgram **3.3.18**, which can be seen as a variant of the Riesz Representation Theorem **3.3.13**. To formulate this theory properly we first introduce some basic notions. After this we state the theorem and give a number of applications.

3.3.1 Introduction to Functional Analysis

Vector-spaces

A *linear space* or *vector space* X is a non-empty set of elements, provided with an addition and a scalar multiplication over a scalar field K. These operations satisfy (for all $x, y \in X$ and scalars $\alpha, \beta \in K$) the following properties

1. $x + y = y + x$,
2. $x + (y + z) = (x + y) + z$,
3. $\exists! \, 0 \in X$, such that $0 + x = x + 0 = x$,
4. $\forall x \in X$, $\exists (-x) \in X$ such that $x + (-x) = 0$,
5. $\alpha(x + y) = \alpha x + \alpha y$,
6. $(\alpha + \beta)x = \alpha x + \beta x$,
7. $(\alpha\beta)x = \alpha(\beta x)$,
8. $1 \cdot x = x$. where $1 \in K$.

The usual fields are the fields of the real ($K = \mathbb{R}$) or the complex ($K = \mathbb{C}$) numbers.

A *normed linear space* is a linear space X provided with a *norm* $\| \cdot \|$, i.e. a mapping $X \to \mathbb{R}$, that satisfies for all $x \in X$ and every scalar $\alpha \in K$ the properties

1. $\|x\| \geq 0$,
2. $\|x\| = 0 \iff x = 0$,
3. $\|x + y\| \leq \|x\| + \|y\|$,
4. $\|\alpha x\| = |\alpha| \, \|x\|$.

Property 3 is called the *triangle inequality*. A *semi-norm* is a similar mapping satisfying 1, 3 and 4, but not necessarily 2. Two norms $\|.\|$ and $\|.\|'$ are called *equivalent* if there exist $c, C > 0$, such that for all $u \in X$

$$c\|u\| \leq \|u\|' \leq C\|u\|.$$

An *inner product space* is a linear space X provided with a mapping $(\cdot, \cdot) : X \times X \to K$ (the *inner product* or *scalar product*), such that for all $x, y \in X$ and scalars α, $\beta \in K$

1. $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$;
2. $(x, y) = \overline{(y, x)}$;
3. $(x, x) \geq 0$, and $(x, x) = 0 \iff x = 0$.

Here $\overline{z}$ denotes the complex conjugate of $z \in \mathbb{C}$.

We say that two elements x and y of an inner product space X are *perpendicular* to each other (notation $x \perp y$) if $(x, y) = 0$.

Every inner product space is a normed linear space because we can define a norm:

$$\|x\| := \sqrt{(x, x)}, \quad \forall x \in X.$$

Quite useful is the *Cauchy-Schwarz inequality*:

$$|(x, y)| \leq \|x\|\|y\|. \tag{3.8}$$

Equality holds in (3.8) if and only if x and y are linearly dependent.

With the norm we have introduced a *metric* or *distance* $\rho(x, y) = \|x - y\|$. So we can discuss the convergence of Cauchy sequences in X and the closure $\overline{X}$. A *Banach space* is a normed linear space which is *closed* in the given norm. This means that every Cauchy sequence converges in the norm given. Such spaces for which $X = \overline{X}$ are called *complete*. A *Hilbert space* is a inner product space which is closed in the norm induced by the inner product. Note that different norms can cause different closures and that a space can be a Banach space when equipped with one norm but will not be a Banach space in another norm. A subset $M \subset X$ is said to be *dense* in X if $\overline{M} = X$. The set X is *separable* if it contains a countable dense subset, or equivalently if it contains a dense sequence.

We conclude this section with some elementary results on Banach and Hilbert spaces.

Lemma 3.3.1 A closed linear subspace Y of a Banach space X is also a Banach space.

Lemma 3.3.2 Let V be a closed linear subspace of a Hilbert space H and $V \neq H$. Then there is a space $V^{\perp}$ (the *orthogonal complement*) with $V \perp V^{\perp}$ such that each element $e \in H$ can be uniquely written as $e = u + v$, where $u \in V$ and $v \in V^{\perp}$

Corollary 3.3.3
Let V be a closed linear subspace of a Hilbert space H ($V \neq H$), then there exists an element $e \neq 0$ in H such that for all $v \in V$:

$$(v, e) = 0.$$

Corollary 3.3.4
Let V be a closed linear subspace of a Hilbert space H and let $u \in H$. If $P : H \to V$ is the orthogonal projection on V, then $Pu \perp u - Pu$. Further

$$(Pv, w) = (v, Pw)$$

for all $v,\ w \in H$, $P^2 = P$ and $\|P\| = 1$.

Function spaces

Functions on Ω (mappings $\Omega \to \mathbb{R}$ or $\Omega \to \mathbb{C}$) can be considered as elements of linear spaces. Such linear spaces are called *function spaces*. Here and in what follows we denote by Ω an open set $\Omega \subset \mathbb{R}^d$. The boundary of Ω is denoted by $\Gamma = \partial\Omega$. This boundary is supposed to be *Lipschitz continuous* [1].

Before we give examples of function spaces we first introduce the *multi-integer* or *multi-index* notation for higher derivatives of functions in more variables. Let $\Omega \subset \mathbb{R}^d$ then we will denote the i-th partial derivative $D_i u$ of the function u by:

$$D_i u = (\frac{\partial}{\partial x_i})u, \quad i = 1, 2, \ldots d.$$

Further we introduce a multi-index $\alpha = (\alpha_1, \alpha_2, \ldots \alpha_d) \in \mathbb{N}^d$ and we write

$$|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_d.$$

Then the α-th derivative of u is

$$D^\alpha u = D_1^{\alpha_1} D_2^{\alpha_2} \cdots D_d^{\alpha_d} u = (\frac{\partial}{\partial x_1})^{\alpha_1} (\frac{\partial}{\partial x_2})^{\alpha_2} \cdots (\frac{\partial}{\partial x_d})^{\alpha_d} u. \tag{3.9}$$

We will now give a number of examples of function spaces as well as norms and inner products we can define on them.

$\mathcal{C}^0(\Omega)$ is the space of all continuous functions defined on Ω and

$$\mathcal{C}^m(\Omega) = \{u \in \mathcal{C}^0(\Omega);\ D^\alpha u \in \mathcal{C}^0(\Omega),\ \ \forall\ |\alpha| \leq m\}.$$

As the elements of these spaces are not necessarily bounded functions we also introduce the space

$$\mathcal{C}^m(\overline{\Omega}) = \{u \in \mathcal{C}^m(\Omega);\ D^\alpha u \text{ are bounded and uniformly continuous on } \overline{\Omega}, \quad \forall\ 0 \leq |\alpha| \leq m\}.$$

This is a Banach space with norm

$$\|u\|_{\mathcal{C}^m(\overline{\Omega})} = \max_{|\alpha| \leq m} \sup_{x \in \Omega} |D^\alpha u(x)|, \tag{3.10}$$

in which α is a multi-integer. We will also use the associated semi-norms:

$$|u|_{\mathcal{C}^m(\overline{\Omega})} = \max_{|\alpha| = m} \sup_{x \in \Omega} |D^\alpha u(x)|. \tag{3.11}$$

The space $\mathcal{C}_0^\infty(\Omega)$ is the subset of all functions in $C^\infty(\overline{\Omega})$ with *compact support*: the function values and the derivatives vanish on the boundary. Analogously we define the subsets $\mathcal{C}_0^m(\Omega)$ of $\mathcal{C}^m(\overline{\Omega})$.

[1] A boundary Γ of a subset Ω of a d-dimensional space is *Lipschitz continuous* if it is locally a sufficiently smooth $(d-1)$-dimensional manifold. In effect we can define an outward unit normal $\mathbf{n}$ a.e. on Γ. See [27].

Example 3.3.5
Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \exp(\frac{1}{|x|^2-1}) & |x| < 1, \\ 0 & |x| \geq 1. \end{cases}$$

This function is arbitrarily often differentiable and its support is the unit ball $\mathbb{T}^d = \{x \in \mathbb{R}^d \mid |x| \leq 1\}$. Then $f \in \mathcal{C}_0^\infty(\mathbb{R}^d)$.

Further we define the Lebesgue-spaces of integrable functions [30]. First, $L^p(\Omega)$, $1 \leq p < \infty$, is the set of functions u on Ω, such that $|u|^p$ is Lebesgue integrable. On $L^p(\Omega)$ we define the (obvious) norm:

$$\|u\|_{L^p(\Omega)} := \left\{ \int_\Omega |u(x)|^p dx \right\}^{\frac{1}{p}}. \tag{3.12}$$

As a particular case , $L^2(\Omega)$ is a Hilbert space with inner product:

$$(u, v)_{L^2(\Omega)} := \int_\Omega u(x)\overline{v(x)}dx. \tag{3.13}$$

By $L^p(\Omega)$ with $p = \infty$ we denote the space $L^\infty(\Omega)$ of essentially bounded functions, with norm

$$\|f\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{x\in\Omega} |f| = \inf\{\lambda \in \mathbb{R} \mid |f(x)| \leq \lambda \text{ a.e.}\}. \tag{3.14}$$

Lemma 3.3.6 $\mathcal{C}_0^\infty(\Omega) \subset L^2(\Omega)$ and $\mathcal{C}_0^\infty(\Omega)$ is densely embedded in $L^2(\Omega)$.
Proof: See [1]. ■

Remark:
As $\mathcal{C}_0^\infty(\Omega)$ is densely embedded in $L^2(\Omega)$ we can say that $L^2(\Omega)$ is the completion of $\mathcal{C}_0^\infty(\Omega)$ in the norm

$$\|u\|_{L^2(\Omega)} = \left\{ \int_\Omega |u(x)|^2 dx \right\}^{1/2}.$$

It is important to note the following inequality due to Hölder.

Lemma 3.3.7
Let p, q be two real numbers such that $q \geq 1$, $p \geq 1$, $\frac{1}{q} + \frac{1}{p} = 1$ and $f \in L^p(\Omega)$, $g \in L^q(\Omega)$. Then $fg \in L^1(\Omega)$ and

$$\|fg\|_{L^1(\Omega)} \leq \|f\|_{L^p(\Omega)} \cdot \|g\|_{L^q(\Omega)}. \tag{3.15}$$

Proof: [30, page 33]. ■

The Hölder inequality (3.15) is a generalisation of the Cauchy-Schwarz inequality (3.8).

Linear operators and functionals

Let X and Y be normed linear spaces over a field K. A *linear operator* $T : X \to Y$ is a mapping such that $T(u+v) = T(u) + T(v)$ and $T(\lambda u) = \lambda T(u)$ for all $u, v \in X$ and scalar $\lambda \in K$. A *conjugate-linear operator* $S : X \to Y$ is a mapping such that $S(u+v) = S(u) + S(v)$ and $S(\lambda u) = \bar{\lambda} S(u)$.

The *norm* of a linear operator T is defined by

$$\|T\| = \|T\|_{Y \leftarrow X} := \sup_{x \in X,\ x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X}. \tag{3.16}$$

A linear operator is *bounded* iff $\|T\| < \infty$.

Lemma 3.3.8 A linear operator T is continuous iff it is bounded.
Proof:
Assume T is bounded: $\|T\| < \infty$. Let $\varepsilon > 0$, then choose $\delta < \frac{\varepsilon}{\|T\|}$. So for all x, y such that $\|x - y\| < \delta$ we have (because T is linear): $\|Tx - Ty\| = \|T(x-y)\| \leq \|T\|\|x - y\| < \|T\| \cdot \frac{\varepsilon}{\|T\|} = \varepsilon$. So T is continuous.
Assume T is continuous, so in 0 we have: $\forall \varepsilon > 0 \ \ \exists \delta > 0 : \ \ \forall x \ : \|x\| < \delta \Rightarrow \|Tx\| < \varepsilon$. Take an y and consider $y^* = \frac{\delta}{2} \cdot \frac{y}{\|y\|}$. We have $\|y^*\| = \delta/2 < \delta$, so $\|Ty^*\| = \frac{\delta}{2} \cdot \frac{\|Ty\|}{\|y\|} < \varepsilon$, so $\frac{\|Ty\|}{\|y\|} < \frac{2\varepsilon}{\delta}$, or $\|T\| \leq \frac{2\varepsilon}{\delta} < \infty$, which means that T is bounded. ■

Definition 3.3.9 A *bilinear operator* $F : X_1 \times X_2 \to Y$ is an operator that is linear in each of its arguments. A *sesquilinear operator* is a mapping $F : X \times X \to Y$ that is linear in its first argument and conjugate-linear in its second.

A bilinear operator $F : X_1 \times X_2 \to Y$ is called *bounded* with bound $\|F\|$ if

$$\|F\| := \sup_{u,v \neq 0} \frac{\|F(u,v)\|_Y}{\|u\|_{X_1}\|v\|_{X_2}} < \infty.$$

A bilinear operator $F : X \times X \to \mathbb{C}$ is called *symmetric* if $F(u,v) = \overline{F(v,u)}$.

Remark:
An inner product is an example of a sesquilinear operator. In fact every strictly positive, symmetric, sesquilinear operator can be considered as an inner product.

A special class of operators are *functionals*, i.e. operators that map elements of a Banach space to the field, K, of real or complex numbers. So we obtain the following definitions.

Definition 3.3.10
(1) A *linear functional* T is a linear operator $T : X \to K$.

(2) A *bilinear functional* F is a bilinear operator $F : X_1 \times X_2 \to K$.
(3) A *sesquilinear functional* F is a sesquilinear operator $F : X \times X \to K$.

Examples: *Functionals*

- Let $X = \mathcal{C}(\overline{\Omega})$ and $f \in X$, then for every $x_0 \in \Omega$ is $T : f \to T(f) = f(x_0)$ a bounded linear functional.
- Let $X = L^p(\Omega)$, $f \in X$ and $w \in \mathcal{C}_0^\infty(\Omega)$ a given function. Then

$$f \mapsto T(f) = \int_\Omega f(x)w(x)dx \tag{3.17}$$

  is a bounded linear functional. The function w is called the *weighting function*.
- Equation (3.17) also defines a bounded linear functional if w is an element in $L^q(\Omega)$ under condition that p and q are *conjugate exponents*, that is $p, q \geq 1$ and

$$\frac{1}{p} + \frac{1}{q} = 1.$$

  This follows from Hölders inequality (3.15).

Dual spaces

Definition 3.3.11 The space of all bounded linear functionals $X \to K$ is called the *dual space* X' of X. It is equipped with the norm

$$\|x'\|_{X'} := \sup_{x \in X,\ x \neq 0} \frac{|x'(x)|}{\|x\|_X} \tag{3.18}$$

We sometimes write $< x', x >$ instead of $x'(x)$ and we will call $< ., . >$ the *duality paring* between X and X'.

We want to characterise the dual of some spaces. First the following elementary theorem.

Theorem 3.3.12 The dual space X' of a normed linear space X is a Banach space.
Proof: Let x'_n be a Cauchy-sequence in X'. This means that

$$\forall \epsilon > 0\ \exists N \geq 0\ \forall n, m \geq N :\ \|x'_n - x'_m\|_{X'} < \epsilon,$$

or

$$\forall \epsilon > 0\ \exists N \geq 0\ \forall n, m \geq N :\ \forall x \in X\ \ |x'_n(x) - x'_m(x)| < \epsilon.$$

Now $\mathbb{R}$ and $\mathbb{C}$ are complete, so for every x the sequence $x'_n(x)$ has a limit, say $x'(x)$. So,

$$\forall \epsilon > 0\ \exists N \geq 0\ \forall n \geq N :\ \forall x \in X\ \ |x'_n(x) - x'(x)| < \epsilon.$$

Now define the function x' as the function that maps every x to the number $x'(x)$ then we see

$$\forall \epsilon > 0 \; \exists N \geq 0 \; \forall n \geq N : \; \|x'_n - x'\|_{X'} < \epsilon,$$

or rather that the Cauchy-sequence x'_n has limit x'. ■

To characterise the dual of a Hilbert space we have the following important theorem. It says that the dual space can be identified with the space itself.

Theorem 3.3.13 (The Riesz Representation Theorem)
Let $u : X \to K$ be a bounded linear functional on a Hilbert space X, then there exists one and only one $y_u \in X$ such that [2]

$$u(x) = (x, y_u)_X \quad \forall x \in X. \tag{3.19}$$

Proof: Consider the null space $N(u) = \{x | u(x) = 0\}$. Due to continuity of u and completeness of X, this is a closed linear subspace of X. Let P be the projection of X to $N(u)$. Assuming $u \neq 0$ (or we could take $y_u = 0$) there is a y_0 such that $u(y_0) \neq 0$. Define $y_1 = y_0 - Py_0$, then $y_1 \perp N(u)$. Further $u(y_1) = u(y_0) - u(Py_0) = u(y_0) \neq 0$. Normalisation yields $y_2 = \left(\frac{1}{u(y_1)}\right) y_1$, so that $u(y_2) = \left(\frac{1}{u(y_1)}\right) u(y_1) = 1$ and still $y_2 \perp N(u)$. Now define $y_u := \frac{y_2}{\|y_2\|^2}$.

Take $x \in X$ arbitrarily, then $z := x - u(x)y_2 \in N(u)$, because $u(z) = u(x) - u(x)u(y_2) = 0$. So

$$\begin{aligned}
& (z, y_2) = 0 \\
\iff \quad & (x - u(x)y_2, y_2) = 0 \\
\iff \quad & (x, y_2) = u(x)(y_2, y_2) \\
\iff \quad & u(x) = \frac{(x, y_2)}{(y_2, y_2)} = (x, \frac{y_2}{\|y_2\|^2}) = (x, y_u).
\end{aligned}$$

Uniqueness is clear, since $(x, z) = 0$ for all x implies $z = 0$. ■

Corollary 3.3.14
The norms of $u \in X'$ and $z \in X$ with $u(x) = (x, z)_X$ are identical:

$$\|u\|_{X'} = \sup_{x \in X} \frac{|u(x)|}{\|x\|_X} = \sup_{x \in X} \frac{|(x, z)|_X}{\|x\|_X} = \|z\|_X,$$

so the mapping $R : X' \to X$ of u to its representation $Ru = z$ is a linear bijective isomorphism.

[2] The converse is trivial: let $u_y(x) = (x, y)$, then $u_y \in X'$.

The dual of a L^p-space is a L^q-space under some conditions, as is shown by the following theorem.

Theorem 3.3.15 (Riesz for $L^p(\Omega)$)
$L^q(\Omega)$ is isometric isomorph to the dual space of $L^p(\Omega)$ if $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. If $p = 1$ then $q = \infty$, but L^∞ can not be associated with the dual of L^1.
Proof: See [1, page 40,41] ∎

With the notion of the dual in mind we can give a weaker definition for convergence. For contrast we also state the usual definition.

Definition 3.3.16 A sequence $\{u_n\}_n$ in X is called *(strongly) convergent* to $u \in X$ if

$$\lim_{n\to\infty} \|u_n - u\|_X = 0.$$

A sequence $\{u_n\}_n$ in X is called *weakly convergent* to $u \in X$ if

$$\lim_{n\to\infty} \varphi(u_n) = \varphi(u) \quad \forall \varphi \in X'.$$

As this defines the weakest topology which renders all elements of the dual continuous, we call it the *weak topology*.

Strong convergence implies weak convergence, but the converse is not true (cf. [22]).

Remark:
If X is a Hilbert space we have by the Riesz Representation Theorem **3.3.13** that weak convergence is equivalent to

$$\lim_{n\to\infty} (u_n, v)_X = (u, v)_X, \quad \forall v \in X.$$

3.3.2 The Generalised Lax-Milgram Theorem

Definition 3.3.17
(1) A symmetric bilinear (sesquilinear) functional $F : X \times X \to K$ is called *coercive* (or *strictly positive*) if there exists a $\gamma > 0$ such that for all $x \in X$

$$|F(x, x)| \geq \gamma \|x\|^2. \tag{3.20}$$

(2) A non-symmetric bilinear functional $F : X_1 \times X_2 \to K$ is called *sub-coercive* if there exists a $\gamma > 0$ such that:

$$\forall x \in X_1 \quad \exists z \in X_2,\ z \neq 0 \quad |F(x, z)| \geq \gamma \|x\|_{X_1} \|z\|_{X_2} \tag{3.21}$$

and

$$\forall z \in X_2,\ z \neq 0 \quad \exists x \in X_1 \quad |F(x, z)| > 0. \tag{3.22}$$

It is easy to see that a coercive functional also satisfies the properties (3.21) and (3.22) of a sub-coercive functional.

Condition (3.22) is equivalent to $\forall z \in X_2,\ z \neq 0\ F(\cdot, z) \not\equiv 0$. Condition (3.21) implies that $\exists \gamma > 0$ such that for all $x \in X_1$:

$$\sup_{z \in X_2} \frac{|F(x,z)|}{\|z\|_{X_2}} \geq \gamma \|x\|_{X_1}. \tag{3.23}$$

This means that we can find a largest possible γ as:

$$\gamma := \inf_{x \in X_1} \sup_{z \in X_2} \frac{|F(x,z)|}{\|x\|\ \|z\|}. \tag{3.24}$$

We now come to the central theorem of this chapter: the Lax-Milgram Theorem. It discusses, in an abstract setting, the existence as well as the uniqueness and boundedness of solutions to elliptic PDEs. We will state it in a more generalised version than usually, namely for the general, asymmetric instead of only the symmetric case. The treatment is such that we can easily generalise it for discrete equations later.

Theorem 3.3.18 (Generalised Lax-Milgram)
Let X_1 and X_2 be two Hilbert spaces and let $F : X_1 \times X_2 \to K$ be a bounded, sub-coercive, bilinear functional. Let $f \in X_2'$, then there exists one and only one $u_0 \in X_1$ such that $\forall v \in X_2$,

$$F(u_0, v) = f(v). \tag{3.25}$$

Further the problem is *stable*: small changes in the data f cause only small changes in the solution u_0:

$$\|u_0\|_{X_1} \leq \frac{1}{\gamma} \|f\|_{X_2'}.$$

Proof: We give the proof in six parts.
(1) As F is bounded we have a bounded linear functional $F(u, \cdot) : X_2 \to K$ for each $u \in X_1$. Now, according to the Riesz Theorem for every $u \in X_1$ a unique element $z \in X_2$ exists such that $F(u, \cdot) = (\cdot, z)_{X_2}$. This defines an operator $R : X_1 \to X_2$ such that

$$F(u, \cdot) = (\cdot, Ru)_{X_2}. \tag{3.26}$$

(2) The operator R is linear, due to the bi-linearity of F.

(3) To prove that R is a surjection we first prove that $R(X_1)$ is a closed subspace of X_2. We first note that:

$$\|Ru\|_{X_2} = \sup_{v \in X_2} \frac{|(Ru, v)|}{\|v\|_{X_2}} = \sup_{v \in X_2} \frac{|F(u,v)|}{\|v\|_{X_2}}.$$

(a) An easy consequence is that R is bounded, and so continuous:

$$\|R\|_{X_2 \leftarrow X_1} = \sup_{u \in X_1} \frac{\|Ru\|_{X_2}}{\|u\|_{X_1}} = \sup_{u \in X_1} \frac{1}{\|u\|} \sup_{v \in X_2} \frac{|F(u,v)|}{\|v\|_{X_2}} = \|F\|.$$

(b) F is sub-coercive, so we get using (3.21) or (3.23):

$$\|Ru\| = \sup_{v} \frac{|F(u,v)|}{\|v\|} \geq \gamma \|u\|. \tag{3.27}$$

(c) Now we show that $R(X_1)$ is closed or that every accumulation point of $R(X_1)$ is an element of it. So let v be an accumulation point of $R(X_1)$, then it is the limit point of some Cauchy-sequence $\{Ru_n\}$. Due to (3.27) we know $\|Ru_m - Ru_n\|_{X_2} \geq \gamma \|u_m - u_n\|$, so $\{u_n\}$ is a Cauchy sequence in X_1. Because X_1 is a Banach space $\exists! u \in X_1$, $u = \lim u_n$ and so $\exists! Ru \in X_2$ as limit of $\{Ru_n\}$. Clearly $v = Ru$, by continuity of R, so $v \in R(X_1)$. Hence it follows that $R(X_1) = \overline{R(X_1)} \subset X_2$.

(4) We prove that the range of R actually is all of X_2. To do so we first assume that $R(X_1) \neq X_2$. This means there exists a $v_0 \neq 0 \in X_2$ such that $v_0 \perp R(X_1)$, or $(v_0, Ru)_{X_2} = 0$ for all $u \in X_1$ (by Corollary **3.3.3**). By (3.26) $F(\cdot, v_0) \equiv 0$, which contradicts the sub-coercivity of F. This implies $R(X_1) = X_2$. Now we have proved that R is surjective. So we obtain $\forall v \in X_2 \quad \exists u \in X_1 \quad Ru = v$ which means $\exists R^{-1} : X_2 \to X_1$.

(5) In fact R^{-1} is the solution operator for the equation (3.25). This is seen as follows. If we take a right-hand-side of our equation $f \in X_2'$, then according to the Riesz theorem $\exists! v_0 \;\; (v, v_0) = f(v), \;\; \forall v \in X_2$. If we define $u_0 = R^{-1} v_0$ then

$$f(\cdot) = (\cdot, v_0) = (\cdot, Ru_0) = F(u_0, \cdot),$$

so u_0 is the solution of equation (3.25).

(6) Finally we prove the stability of the problem, using (3.27): $\gamma \|u_0\| \leq \|Ru_0\| = \|v_0\| = \|f\|$, or:

$$\|u_0\| \leq \frac{1}{\gamma} \|f\|.$$

Uniqueness is a direct consequence of this stability. ∎

3.3.3 Distributions and Sobolev Spaces

Solutions to (weak formulations of) elliptic PDEs are often found among special classes of functions: the so called Sobolev spaces. In this section we will discuss these spaces and some of the properties of their elements. Because it involves weak formulations of PDEs we first treat the subject of distributions.

Distributions

We define the function space $\mathcal{D}(\Omega)$ as the set of functions $\mathcal{C}_0^\infty(\Omega)$ supplied with the following topology. We say that a sequence $\varphi_n \in \mathcal{D}(\Omega)$ converges to $\varphi \in \mathcal{D}(\Omega)$, notation

$$\lim_{n\to\infty} \varphi_n = \varphi,$$

if and only if there exists a compact subset $K \subset \Omega$ such that for every function φ_n we have $\text{supp}(\varphi_n) \subset K$ and such that for all $|\alpha| \geq 0$

$$D^\alpha \varphi_n \to D^\alpha \varphi.$$

So $\mathcal{D}(\Omega)$ is the same set of functions as $\mathcal{C}_0^\infty(\Omega)$, but it has more structure.

Definition 3.3.19 A *distribution* (or *generalised function*) on Ω is a bounded linear functional on $\mathcal{D}(\Omega)$. The space of distributions is denoted by $\mathcal{D}'(\Omega)$, being the dual of $\mathcal{D}(\Omega)$.

Example 3.3.20 *(*Generalised functions*)*

A function $f \in L^p(\Omega)$ defines[3] a distribution $T_f : \mathcal{D}(\Omega) \to \mathbb{R}$ by:

$$T_f(\varphi) = \int_\Omega f(x)\varphi(x)dx, \quad \forall \varphi \in \mathcal{D}(\Omega).$$

In this way we can identify each L^p-function with a distribution from $\mathcal{D}'(\Omega)$. Sometimes we write $T_f(\varphi) =< f, \varphi >$, in which $< \cdot, \cdot >$ is the duality paring between $\mathcal{D}(\Omega)$ and $\mathcal{D}'(\Omega)$.

Example 3.3.21
Let $x_0 \in \Omega$ and define $T_{x_0} : \mathcal{D}(\Omega) \to \mathbb{R}$ as

$$T_{x_0}(\varphi) = \varphi(x_0).$$

This generalised function is called the *Dirac delta function* associated with point x_0 and it is also called δ_{x_0}. We write $< \delta_{x_0}, \varphi >$ or

$$\int_\Omega \delta_{x_0}(x)\varphi(x)dx.$$

However, the latter is rather formal because it has nothing to do with an actual integration.

[3] In fact we can extend the class of such functions to the *locally integrable* functons

$$L^1_{\text{loc}}(\Omega) = \left\{ f \mid f \in L^1(K) \ \forall \text{compact} K \subset \Omega \right\} .$$

The *topology* on $\mathcal{D}'(\Omega)$ is defined as follows. With $T_n, T \in \mathcal{D}'(\Omega)$ we say that

$$\lim_{n\to\infty} T_n = T$$

when:

$$\lim_{n\to\infty} T_n(\varphi) = T(\varphi), \quad \forall \varphi \in \mathcal{D}(\Omega).$$

Because of the analogy with weak convergence we will call this the *weak topology.*

Definition 3.3.22 The *distributional* or *generalised derivative* $D^\alpha T$ of a distribution T is defined by

$$D^\alpha T(\varphi) = (-1)^{|\alpha|} T(D^\alpha \varphi), \ \ \forall \varphi \in \mathcal{D}(\Omega).$$

If we have a differentiable f, say $f \in \mathcal{C}^1(\overline{\Omega})$, then we may consider the distribution T_f associated with f, and the distribution T_{Df} associated with the derivative Df of f. Now we see, for any $\varphi \in \mathcal{D}(\Omega)$

$$T_{Df}(\varphi) = \int_\Omega Df(x)\varphi(x)dx = -\int_\Omega f(x)D\varphi(x)dx = -T_f(D\varphi) = DT_f(\varphi).$$

This means that, for differentiable functions, the generalised derivative of the (generalised) function is the generalised function of the derivative. However, for all (generalised) functions all generalised derivatives (always) exist.

Remark:
The mapping $D^\alpha : \mathcal{D}'(\Omega) \to \mathcal{D}'(\Omega)$ is continuous because if

$$\lim_{n\to\infty} T_n = T \quad \text{in} \quad \mathcal{D}'(\Omega),$$

then $\forall \varphi \in \mathcal{D}(\Omega)$:

$$D^\alpha T_n(\varphi) = (-1)^{|\alpha|} T_n(D^\alpha \varphi) \to (-1)^{|\alpha|} T(D^\alpha \varphi) = D^\alpha T(\varphi).$$

Example 3.3.23 *(*Heaviside and delta-function*)*
We consider the *Heaviside function* $H(x)$, which yields 0 if $x < 0$ and 1 if $x \geq 0$. Of course, this is not an L^2-function, (if $\Omega = \mathbb{R}$) but it is a distribution:

$$T_H(\varphi) = \int_{\mathbb{R}} H(x)\varphi(x)\ dx = \int_0^\infty \varphi(x)\ dx,$$

which is clearly bounded as $\varphi \in \mathcal{C}_0^\infty(\Omega)$. We compute the derivative DT_H:

$$\begin{aligned} DT_H(\varphi) &= -T_H(D\varphi) = -\int_{\mathbb{R}} H(x)D\varphi(x)dx \\ &= -\int_0^\infty D\varphi(x)dx = \varphi(0) = T_{\delta_0}(\varphi). \end{aligned}$$

This means $DT_H = T_{\delta_0}$, i.e. the generalised derivative of the Heaviside-function is the Dirac delta-function.

Sobolev spaces

In the theory of elliptic PDEs, functions that, together with their (generalised) derivatives, are elements of $L^2(\Omega)$, play an important role. Therefore, we introduce the so called Sobolev spaces.

Definition 3.3.24 The *Sobolev space* $W^{n,p}(\Omega)$, $p \geq 1, \quad n = 0, 1, \ldots$, is the space of all $L^p(\Omega)$-functions u for which all distributional derivatives $D^\alpha u$, $|\alpha| \leq n$ are also elements of $L^p(\Omega)$:

$$W^{n,p}(\Omega) = \{u \in L^p(\Omega) | D^\alpha u \in L^p(\Omega), \ \forall \ |\alpha| \leq n\}.$$

This is a Banach space if we provide it with the norm:

$$\|u\|_{n,p,\Omega} := \left\{ \sum_{|\alpha| \leq n} \|D^\alpha u\|^p_{L^p(\Omega)} \right\}^{\frac{1}{p}}. \tag{3.28}$$

In case of $p = 2$ we denote the space $W^{n,2}(\Omega)$ as $H^n(\Omega)$. Supplied with the inner product

$$(u, v)_{W^{n,2}(\Omega)} = \sum_{|\alpha| \leq n} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}, \tag{3.29}$$

these Sobolev spaces are Hilbert spaces. We write $\|u\|_{n,\Omega} := \|u\|_{n,2,\Omega}$, or even $\|u\|_n := \|u\|_{n,2,\Omega}$.

Another way of saying that a functional u is an element of $W^{n,p}(\Omega)$ is that they are L^p-functions u^α, $0 \leq |\alpha| \leq n$ such that

$$\int_\Omega u^\alpha(x) \ \varphi(x) \ dx = (-1)^{|\alpha|} \int_\Omega u(x) \ D^\alpha \varphi(x) \ dx,$$

for all $\varphi \in \mathcal{D}(\Omega)$.

Example 3.3.25
The definition is not trivial. For instance the Heaviside function on an interval $\Omega = [a, b]$, $-\infty < a < 0 < b < \infty$ shows that $u \in L^p(\Omega)$ does not imply $Du \in L^p(\Omega)$.

Example 3.3.26
Consider the basic hat function on the interval $\Omega = [0, 1]$:

$$u(x) = \left\{ \begin{array}{ccc} 2x & \text{for} & x \in [0, \frac{1}{2}) \\ 2 - 2x & \text{for} & x \in [\frac{1}{2}, 1]. \end{array} \right\}.$$

This function is in $L^2(\Omega)$. To look if it is also in $H^1(\Omega)$ we must find an $L^2(\Omega)$-function du that is the derivative of u in distributional sense: $\int_\Omega du(x) \ \varphi(x) \ dx = -\int_\Omega u(x) \ D\varphi(x) \ dx$, for all $\varphi \in \mathcal{D}(\Omega)$. Now take the function du such that $du = \frac{\partial u}{\partial x}$, except for $x = 0, \ \frac{1}{2}, \ 1$ where we define it (arbitrarily) zero. This

function du is a simple function, so it is in L^2, so it remains to check if this is the generalised derivative of u.

$$\begin{aligned}
\int_\Omega du(x)\ \varphi(x)\ dx &= \int_0^{\frac{1}{2}} du(x)\ \varphi(x)\ dx + \int_{\frac{1}{2}}^1 du(x)\ \varphi(x)\ dx \\
&= \int_0^{\frac{1}{2}} \frac{\partial u}{\partial x}\ \varphi(x)\ dx + \int_{\frac{1}{2}}^1 \frac{\partial u}{\partial x}\ \varphi(x)\ dx \\
&= \left\{ \lim_{x\uparrow\frac{1}{2}} u(x)\ \varphi(x) - \lim_{x\downarrow 0} u(x)\varphi(x) \right\} - \int_0^{\frac{1}{2}} u(x)\ \frac{\partial \varphi}{\partial x}(x)\ dx \\
&+ \left\{ \lim_{x\uparrow 1} u(x)\varphi(x) - \lim_{x\downarrow\frac{1}{2}} u(x)\ \varphi(x) \right\} - \int_{\frac{1}{2}}^1 u(x)\ \frac{\partial \varphi}{\partial x}(x)\ dx \\
&= -\int_0^{\frac{1}{2}} u(x)\ \frac{\partial \varphi}{\partial x}(x)\ dx - \int_{\frac{1}{2}}^1 u(x)\ \frac{\partial \varphi}{\partial x}(x)\ dx \\
&= -\int_\Omega u(x)\ D\varphi(x)\ dx,
\end{aligned}$$

where we have used the fact that

$$\lim_{x\uparrow\frac{1}{2}} u(x) = \lim_{x\downarrow\frac{1}{2}} u(x)$$

and that $\varphi \in \mathcal{D}([0,1])$, so $\varphi(0) = \varphi(1) = 0$. ∎

The proof shows that it is essential that u is continuous in the point $\frac{1}{2}$, where its (generalised) derivative is discontinuous. This raises the question under which conditions elements of Sobolev spaces are continuous or even continuously differentiable.[4] This is the content of the next theorem.

Theorem 3.3.27 (Sobolev's Embedding Theorem)
Let $\Omega \subset \mathbb{R}^d$ be a bounded open region and let $n > k + d/p$, then every element from $W^{n,p}(\Omega)$ is k times continuously differentiable. Or:

$$n > k + d/p \Rightarrow W^{n,p}(\Omega) \subset \mathcal{C}^k(\overline{\Omega}). \tag{3.30}$$

Proof: [30, page 174] ∎

Example 3.3.28
Note that the dimension d plays a role. So for $d = 1$ the functions of $H^1(\Omega)$ are continuous, but for $d = 2$ they do not have to be and indeed it should be

[4] An element u of a Sobolev space is continuous, bounded or whatever, if there is an equivalent function u' such that this function has the property.

noted that the Sobolev's inequality (3.30) is *sharp*: there are functions in $H^1(\Omega)$ ($d = 2$) that are not continuous. An example of this is

$$f(x, y) = \log|\log(x^2 + y^2)| \text{ on } B_{\frac{1}{2}}(0, 0).$$

It is not continuous in $(x, y) = (0, 0)$, but it is not difficult to prove that $\|f\|_{L^2}$, $\|\frac{\partial f}{\partial x}\|_{L^2}$ and $\|\frac{\partial f}{\partial y}\|_{L^2}$ do exist. ∎

Example 3.3.29
Remember that, according to Lemma **3.3.6**, we have that $\mathcal{C}_0^\infty(\Omega)$ is dense in $L^2(\Omega)$. Unfortunately, for a true subset $\Omega \subset \mathbb{R}^d$, the space $\mathcal{C}_0^\infty(\Omega)$ is *not* dense in $W^{n,p}(\Omega)$. For instance, consider $H^1(\Omega)$. Assuming that $\mathcal{C}_0^\infty(\Omega)$ is not dense we compute the orthogonal complement:

$$0 = (u, \varphi)_{H^1} = (u, \varphi)_{L^2} + (Du, D\varphi)_{L^2} = (u, \varphi)_{L^2} - (\Delta u, \varphi)_{L^2} = (u - \Delta u, \varphi)_{L^2}.$$

So the orthogonal complement of $\mathcal{C}_0^\infty(\Omega)$ consists of those u that satisfy $u - \Delta u = 0$ distributionally. A non trivial example is the function $e^{(\xi,x)}$, with $|\xi| = 1$. This function is in $H^1(\Omega)$, but not in $H_0^1(\Omega)$ (in terms of the trace operator in the next section: the trace on Γ is not zero). Luckily however, $\mathcal{C}^\infty(\overline{\Omega})$ is dense in $W^{n,p}(\Omega)$ under the Sobolev norm. This implies that we might introduce $W^{n,p}(\Omega)$ as the closure of $\mathcal{C}^\infty(\overline{\Omega})$ in the norm $\|\cdot\|_{n,p,\Omega}$.

We now give the following definition.

Definition 3.3.30 $W_0^{n,p}(\Omega)$ is the closure of $\mathcal{C}_0^\infty(\Omega)$ in the norm $\|\cdot\|_{n,p,\Omega}$. In analogy to $W^{n,2}(\Omega) = H^n(\Omega)$ we write also $W_0^{n,2}(\Omega) = H_0^n(\Omega)$.

Corollary 3.3.31
$\mathcal{D}(\Omega)$ is dense in $H_0^n(\Omega)$.

From Lemma **3.3.6** we see that $H_0^0(\Omega) = H^0(\Omega)$. For $n \geq 1$ however $H_0^n(\Omega) \neq H^n(\Omega)$, as is shown by Example **3.3.29**.

We are interested in the dual spaces of the Sobolev spaces. First a definition.

Theorem 3.3.32 Let $p \neq \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$, then

$$(W_0^{p,m}(\Omega))' \cong W^{q,-m}(\Omega).$$

Proof: [1, page 48] ∎

For the Sobolev spaces with $p = q = 2$ we write $H^{-m} = (H_0^m)'$, $m > 0$. In particular we use $H^{-1}(\Omega) := W^{2,-1}(\Omega)$, the dual of $H_0^1(\Omega)$.

Trace operator

For a function $f \in \mathcal{C}^\infty(\overline{\Omega})$ we define the *trace* $\gamma_0 f$ as $\gamma_0 f = f|_\Gamma$ i.e. the restriction of f to Γ, the boundary of the domain Ω. [5]

If Ω is a bounded area and has a Lipschitz continuous boundary then one can show that the linear mapping

$$\gamma_0 : \mathcal{C}^\infty(\overline{\Omega}) \to L^2(\Gamma)$$

is bounded in the sense that

$$\|\gamma_0 u\|_{L^2(\Gamma)} \leq C \|u\|_{H^1(\Omega)}.$$

So one can extend the operator γ_0 to a bounded linear operator

$$\gamma_0 : H^1(\Omega) \to L^2(\Gamma).$$

$H_0^1(\Omega)$ is exactly the kernel of γ_0:

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : \ \ \gamma_0 u = 0\}.$$

We also define the trace operator γ_1:

$$\gamma_1 u := \frac{\partial u}{\partial n} = \sum_{i=1}^{d} \gamma_0 \left(\frac{\partial u}{\partial x_i} \right) n_i.$$

3.3.4 The Poincaré-Friedrichs Inequality.

To prove the coercivity of a bilinear functional one can sometimes use the following lemma.

Lemma 3.3.33 (Poincaré's Lemma)
If $\Omega \subset \mathbb{R}^d$ is a bounded and open set then for all $v \in H_0^1(\Omega)$

$$\int_\Omega |v|^2 dx \leq C(\Omega) \int_\Omega |\nabla v| dx. \tag{3.31}$$

Proof: It is sufficient to prove the statement for functions in $\mathcal{D}(\Omega)$ as $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$.
We enclose Ω in a rectangle S:

$$\Omega \subset S = [a_1, b_1] \times \cdots \times [a_d, b_d].$$

We extend u to S by setting $u(x) = 0$ whenever $x \in S \setminus \Omega$.
As $u \in \mathcal{D}(\Omega)$ we have $u \in \mathcal{D}(S)$ and for every $i = 1, \ldots, d$:

$$u(x) = \int_{a_i}^{x_i} \frac{\partial u}{\partial x_i}(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_d), dt$$

[5] For more on trace operators we refer to [1, page 113].

so we get by the Cauchy-Schwarz-Lemma (3.8):

$$\begin{aligned} |u(x)|^2 &\leq \int_{a_i}^{x_i} 1^2 dt \cdot \int_{a_i}^{x_i} |\frac{\partial u}{\partial x_i}(\ldots,t,\ldots)|^2 dt \\ &\leq (b_i - a_i) \cdot \int_{a_i}^{b_i} |\frac{\partial u}{\partial x_i}(\ldots,t,\ldots)|^2 dt \end{aligned}$$

Now integrate over S:

$$\begin{aligned} \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} |u(x)|^2 dx &\leq (b_i - a_i) \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} \int_{a_i}^{b_i} |\frac{\partial u}{\partial x_i}(\ldots,t,\ldots)|^2 \, dx \, dt \\ &= (b_i - a_i)^2 \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} |\frac{\partial u}{\partial x_i}(\ldots,t,\ldots)|^2 dx. \end{aligned}$$

So we get for every $i = 1, 2, \ldots, n$,

$$\int_\Omega |u(x)|^2 dx \leq (b_i - a_i)^2 \int_\Omega |\nabla_i u|^2 dx$$

and so

$$\int_\Omega |u(x)|^2 dx \leq C(\Omega) \int_\Omega |\nabla u|^2 dx.$$

■

Corollary 3.3.34
When the set Ω is bounded, the semi-norm $|\cdot|_{1,\Omega}$ is a norm over the space $H_0^1(\Omega)$ equivalent to the norm $\|\cdot\|_{1,\Omega}$.

The inequality (3.31) is also known as the Poincaré-Friedrichs inequality.

3.3.5 Variational formulations for differential equations

A Dirichlet problem

Let $\Omega \subset \mathbb{R}^d$ be an open bounded area with Lipschitz continuous boundary Γ. Let $u \in \mathcal{C}^2(\overline{\Omega})$ be a solution of the *Poisson problem with homogeneous Dirichlet boundary conditions*

$$-\Delta u = f \quad \text{in } \Omega, \tag{3.32}$$

$$u = 0 \quad \text{on } \Gamma, \tag{3.33}$$

then, by partial integration, for every $\varphi \in \mathcal{C}^\infty(\overline{\Omega})$

$$\begin{aligned} \int_\Omega \varphi f \, dx &= -\int_\Omega \varphi \Delta u \, dx = -\int_\Omega \varphi \sum_i \nabla_i \nabla_i u \, dx & (3.34) \\ &= \sum_i \int_\Omega \nabla_i \varphi \nabla_i u \, dx - \int_\Gamma \varphi (\nabla u \cdot \mathbf{n}) \, ds, & (3.35) \end{aligned}$$

where $\mathbf{n}$ is the outward unit normal. For all $\varphi \in \mathcal{D}(\Omega)$ we get

$$\int_\Omega \varphi f dx = \sum_i \int_\Omega \nabla_i \varphi \nabla_i u \; dx.$$

This leads to the following formulation of a problem similar to (3.32): find a u with $u = 0$ on Γ such that

$$\int_\Omega \nabla_i \varphi \nabla_i u \; dx = \int_\Omega \varphi f dx, \quad \forall \varphi \in \mathcal{C}_0^\infty(\Omega). \tag{3.36}$$

We see that a solution of (3.32) is also a solution of (3.36).

Equation (3.36) is of the following form: given the bilinear form $B : S \times V \to \mathbb{R}$ defined by

$$B(u, \varphi) = \sum_i \int_\Omega \nabla_i \varphi \nabla_i u \; dx$$

and the linear form $l : V \to \mathbb{R}$, defined by

$$l(\varphi) = \int_\Omega \varphi f \; dx,$$

find a $u \in S$ such that

$$B(u, \varphi) = l(\varphi) \quad \text{for all} \quad \varphi \in V. \tag{3.37}$$

In our case (3.36) we see that $B : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is a symmetrical bilinear form defined on $H_0^1(\Omega)$, and from Poincaré's Lemma **3.3.33** we get:

$$B(u, u) = \int_\Omega |\nabla u|^2 \; dx \geq \gamma \, \|u\|_{1,\Omega},$$

which tells us $B : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is coercive. Now it is an immediate consequence of the generalised theorem of Lax-Milgram **3.3.18** that for every $f \in H^{-1}(\Omega)$ there exists a unique solution $u \in H_0^1(\Omega)$ to (3.36) for which

$$\|u\|_{H_0^1(\Omega)} \leq \frac{1}{\gamma} \|f\|_{H^{-1}(\Omega)}.$$

If (3.32) has a solution, we can find this solution by solving (3.36).

It is clear that there will be cases where (3.36) has a solution but (3.32) hasn't: for instance when $u \in H_0^1(\Omega)$ but $u \notin \mathcal{C}_0^2(\Omega)$. Obviously we have generalised the meaning of (3.32). Hence we will call the solution $u \in H_0^1(\Omega)$ the *weak solution*.

More general problems

Let $u \in \mathcal{C}^2(\Omega)$ be a solution to

$$-\Delta u + cu = f \quad \text{in } \Omega, \tag{3.38}$$

$$u_n + \beta u = h \quad \text{on } \Gamma, \tag{3.39}$$

then for every $\varphi \in \mathcal{C}^\infty(\overline{\Omega})$:

$$\begin{aligned}\int_\Omega \varphi f dx &= -\int_\Omega \Delta u\ \varphi\ dx + \int_\Omega c\varphi\ u\ dx \\ &= \int_\Omega \nabla_i\varphi\nabla_i u\ dx + \int_\Omega c\varphi\ u\ dx - \int_\Gamma \varphi\nabla_i u\ n_i\ dx \\ &= \int_\Omega [\nabla_i\varphi\nabla_i u + c\ \varphi u\ ]\, dx - \int_\Gamma \varphi(h-\beta u)\ ds \\ &= \int_\Omega [\nabla_i\varphi\nabla_i u + c\ \varphi u\ ]\, dx + \int_\Gamma \beta\varphi u\ ds - \int_\Gamma \varphi h\ ds.\end{aligned}$$

This leads to the problem: find $u \in S$ such that

$$B(u,\varphi) = l(\varphi), \quad \forall \varphi \in V,$$

where

$$B(u,\varphi) = \int_\Omega \nabla_i\varphi\nabla_i u\ dx + \int_\Omega c\ \varphi\ u\ dx + \int_\Gamma \beta\ \varphi\ u\ ds$$

and

$$l(\varphi) = \int_\Omega \varphi\ f\ dx + \int_\Gamma \varphi\ h\ ds.$$

The bilinear operator $B : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is bounded and symmetric, and it is coercive if $c \geq 0$, $\beta \geq 0$, except when $c = 0$ and $\beta = 0$:

$$\begin{aligned}|B(\varphi,\varphi)| &= \left| \int_\Omega |\nabla\varphi|^2\ dx + c\int_\Omega |\varphi|^2\ dx + \beta\int_\Gamma |\varphi|^2\ ds \right| \\ &= \int_\Omega |\nabla\varphi|^2\ dx + c\int_\Omega |\varphi|^2\ dx + \beta\int_\Gamma |\varphi|^2\ ds \\ &\geq C\left\{|\varphi|^2 + |\nabla\varphi|^2\right\} = C\|\varphi\|_{1,\Omega}, \quad \forall \varphi \in H^1(\Omega),\end{aligned}$$

for some $C > 0$, and also

$$|B(\varphi,\varphi)| \leq C\left\{|\varphi|^2 + |\nabla\varphi|^2\right\} = C\|\varphi\|_{1,\Omega}, \quad \forall \varphi \in H^1(\Omega),$$

for some $C > 0$. From the Lax-Milgram Theorem we conclude that there exists one and only one solution in $H^1(\Omega)$. So, if the solution to (3.38) with boundary condition (3.39) exists, we can find it by solving the weak problem.

Symmetric problems and minimisation of quadratic functionals

In the equations (3.32) and (3.38), where $B(u,v)$ is a symmetric, coercive bilinear form, the functional $J(\varphi) = B(\varphi,\varphi)$ can be considered as a quadratic form. $B(\varphi,\varphi) \geq 0$ and $B(\varphi,\varphi) = 0$ if and only if $\varphi = 0$. For problem (3.32) this functional is defined on $X = H_0^1(\Omega)$, for problem (3.38) on $X = H^1(\Omega)$.

Theorem 3.3.35 Under the circumstances that $B(\varphi,\varphi)$ is a quadratic form, we can also characterise the solution of the problem: find $u \in X$ such that

$$B(u,\varphi) = l(\varphi) \text{ for all } \varphi \in X$$

as: find $u \in X$ which minimises the quadratic functional

$$J(u) = \frac{1}{2}B(u,u) - l(u).$$

Proof:

$$\begin{aligned} J(u+\varphi) &= \frac{1}{2}B(u+\varphi,u+\varphi) - l(u+\varphi) \\ &= \frac{1}{2}B(u,u) - l(u) + B(u,\varphi) - l(\varphi) + \frac{1}{2}B(\varphi,\varphi). \end{aligned}$$

For a u that satisfies $B(u,\varphi) = l(\varphi)$ we see that

$$J(u+\varphi) = J(u) + \frac{1}{2}B(\varphi,\varphi) \geq J(u).$$

Obviously, there is one and only one $u \in X$ that minimises $J(\cdot)$. ∎

Asymmetric problems

A quite different situation exists for problems with first derivatives and/or more general boundary values:

$$\begin{cases} -\Delta u + b_j \nabla_j u + cu = f & \text{on } \Omega, \\ u_n + \alpha u_s + \beta u = h & \text{on } \Gamma. \end{cases} \tag{3.40}$$

Then for $\varphi \in \mathcal{C}^\infty(\Omega)$

$$\int_\Omega \varphi f dx = \int_\Omega \left[\nabla_i \varphi \nabla_i u + \varphi\ b_j \nabla_j u + \varphi\ c\ u\right]\ dx - \int_\Gamma \varphi(h - \beta u - \alpha u_s) ds.$$

This again can be written in the form

$$B(u,\varphi) = l(\varphi), \quad \forall \varphi \in V,$$

but in this case the form $B(u,\varphi)$ is not symmetric, due to the first order derivative. To prove that the conditions of Lax-Milgram are satisfied is a bit more hard to do. As we have seen in Theorem **3.2.1** we can get rid of the first order terms by a transformation of the variables. This helps us to prove existence and uniqueness (as in Section 3.3.5), but for numerical purposes the transformation is of no use. We will see an example of this in the chapter concerning the convection-diffusion equation.

3.4 The technique of the finite element method

In this section we treat the practical aspects of the finite element method. We give a simple error estimate and we discuss the choice of basic functions (in one or two dimensions) and the construction of the discrete equations for an example problem. We will also discuss the use of isoparametric elements for the case of curved boundaries. The techniques explained here for the one and two-dimensional case can readily be generalised to three space dimensions.

3.4.1 The principles

The basic principle of the finite element method is the following. It is a Galerkin method applied to the variational problem: find $u \in S$ such that

$$B(u, \varphi) = l(\varphi), \ \forall \ \varphi \in V, \tag{3.41}$$

in which S and V are (infinitely dimensional) vector spaces. Usually we have a *conforming* finite element discretisation for which the discrete system of equations read: find $u_h \in S_h$ such that

$$B(u_h, \varphi_h) = l(\varphi_h), \ \forall \ \varphi_h \in V_h, \tag{3.42}$$

where S_h and V_h are finite dimensional subspaces of S and V, with $\dim(S_h) = \dim(V_h)$. We also may replace $B(\cdot, \cdot)$ with an approximation $B_h(\cdot, \cdot)$ and $l(\cdot)$ with an approximation $l_h(\cdot)$. The consequences of this are discussed in Section **3.5**.

If $B(\cdot, \cdot)$ is a symmetric and coercive bilinear form and if $S = V$ then we can show that this type of discretisation gives an approximate solution if $S_h = V_h \subset S = V$. The solution to (3.41) is the minimising function of a quadratic functional $J(u)$ over S. The discrete function u_h minimises the same functional over the subspace. This implies that an optimal solution is found in the norm induced by the problem (the *energy norm*). We describe this in the following theorem.

Theorem 3.4.1 Let $(\cdot, \cdot)_B = B(\cdot, \cdot)$ be the inner product induced by the bilinear form B and let $e_h = u_h - u$ be the error in the approximation, then, in the norm $\|.\|_B$ induced by B, the solution of (3.42) is in $S_h = V_h$ the optimal approximation of u, the solution to (3.41):

$$\|e_h\|_B \leq \inf_{v_h \in S_h} \|u - v_h\|_B.$$

Proof: First,

$$\|e_h\|_B^2 = B(e_h, e_h) = B(e_h, u_h - u).$$

Now because $B(u, v_h) = f(v_h)$ for all $v_h \in V$, and $B(u_h, v_h) = f(v_h)$ for all $v_h \in V_h$, we have $B(u_h - u, v_h) = B(e_h, v_h) = 0$ for all $v_h \in V_h$ and because $u_h \in S_h = V_h$ we get $B(e_h, u_h) = 0$. So

$$B(e_h, u_h - u) = B(e_h, v_h - u) \leq \|e_h\|_B \|u - v_h\|_B \ \ \forall v_h \in V_h.$$

This gives

$$\|e_h\|_B \leq \inf_{v_h \in V_h} \|u - v_h\|_B.$$

■

We want to investigate the approximation in other norms (L^p and Sobolev norms) and for more general problems. To this purpose we will first consider some specific approximation spaces V_h and in the next chapter discuss in more detail the error estimates that are valid when we use these approximations.

3.4.2 Piecewise Lagrange Interpolation in one dimension

We divide an interval $\Omega = [a, b] \subset \mathbb{R}$ in N (not necessarily equal) subintervals

$$e_j = [x_{j-1}, x_j] \quad j = 1, \ldots, N.$$

We denote this partition by

$$\Omega_N : \quad a = x_0 < x_1 < \ldots < x_N = b.$$

Further we define $h_j = x_j - x_{j-1}$ and $h = \max\{h_j\}$, the *meshwidth* of the *mesh* Ω_N.

Definition 3.4.2
(1) Let Ω' be a (connected) subinterval of Ω, then $P_k(\Omega')$ is the set of all polynomials of degree less or equal k on Ω'.
(2) We define $P_k(\Omega_N)$ as the space of continuous, piecewise polynomial functions on Ω:

$$P_k(\Omega_N) = \{f | f \in \mathcal{C}(\Omega); \;\; f|_{e_i} \in P_k(e_i), \;\; i = 1, \ldots, N\}.$$

A polynomial in $P_k(e_i)$ needs $k + 1$ (independent) values in order to be defined correctly and uniquely.

Lemma 3.4.3 $P_k(\Omega_N) \subset H^1(\Omega)$. Also $P_k(\Omega) \subset P_k(\Omega_N)$, for each partition Ω_N.
Proof: We prove the first, the second is obvious.
Take a polynomial $p \in P_k(\Omega_N)$, then $p \in L^2(\Omega)$. Define the distribution $\frac{dp}{dx}$ by $dp(x) = \frac{\partial p}{\partial x}(x)$, $x \neq x_i$ and $\frac{dp}{dx}(x_i) = 0$, then we can prove in a way similar to Example **3.3.26** (and using the continuity of p) that dp is the distributional derivative of p. Also $dp \in L^2(\Omega)$, so $p \in H^1(\Omega)$. ■

Now we want to determine the dimension and a basis for the space $P_k(\Omega_N)$ for $k = 1, 2, 3$. For this purpose we first define on the subinterval e_i a new variable $t = (x - x_{j-1})/h_j$, so that $t = 0$ corresponds with $x = x_{j-1}$ and $t = 1$ with $x = x_j$.

The case $k = 1$.

An element from $P_1(\Omega_N)$ is a piecewise linear function. Such a function is determined by its values at the points x_i. So there are $N+1$ basis functions. A natural choice is the set $\{\psi_j\}_{j=1}^N$ where

$$\psi_j(x_i) = \delta_{ij} \quad i,j = 0,\ldots,N. \tag{3.43}$$

These are the so called *standard hat functions*. In figure 3.1 the function ψ_i is given.

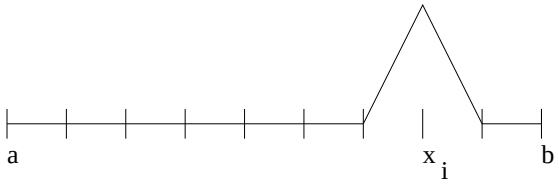


Figure 3.1: Hat function ψ_j.

It turns out that a basis function ψ_i vanishes everywhere on Ω except on the two subintervals to which x_i belongs. For each subinterval e_j there are two non-zero basis-functions:

$$\begin{aligned} \psi_{j-1}(x) &= 1-(x-x_{j-1})/h_j = (1-t), \\ \psi_j(x) &= (x-x_{j-1})/h_j = t. \end{aligned}$$

The case $k=2$.
An element from $P_2(\Omega_N)$ is a piecewise quadratic function. The functions are not completely determined by their values on the points in the partition Ω_N. We need one additional value in each interval. We take in each e_j an interior point, for instance $x_{j-\frac{1}{2}} = (x_{j-1}+x_j)/2$, the midpoint of the interval. We now choose the basis functions $\{\psi_i\} \subset P_2(\Omega_N)$ as

$$\psi_i(x_j) = \delta_{i,j} \quad i,j = 0,1/2,1,3/2,\ldots,(2N-1)/2,N.$$

The number of basis functions is $2N+1$. Again $\psi_i \neq 0$ on at most two segments e_i, namely those to which x_i belongs. On each segment e_j there are at most three basis functions unequal to zero

$$\begin{aligned} \psi_{j-1}(x) &= (1-t)(1-2t), \\ \psi_{j-\frac{1}{2}}(x) &= 4t(1-t), \\ \psi_j(x) &= -t(1-2t). \end{aligned}$$

In figure 3.2 the functions ψ_j and $\psi_{j-\frac{1}{2}}$ are shown. All basis-functions ψ_i, i integer, are similar to ψ_j (with the exception for the two functions at the

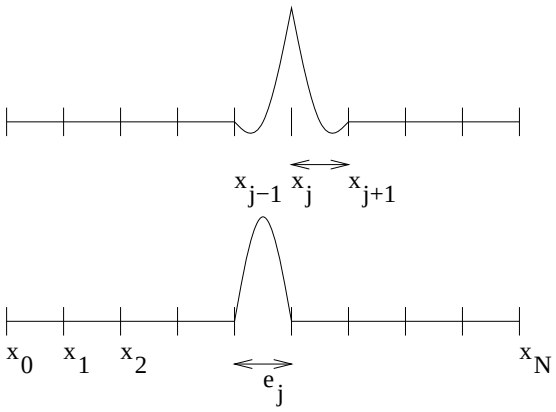


Figure 3.2: Second order Lagrange.

boundary). Note that the support of $\psi_{j-\frac{1}{2}}$ is completely contained in e_j. All functions $\psi_{i-\frac{1}{2}}$, i integer, are translated versions of $\psi_{j-\frac{1}{2}}$.

The case $k = 3$.
This can be treated completely analogous to the case $k = 2$. We get the space $P_3(\Omega_N)$ of piecewise cubic functions. Now we have two extra degrees of freedom for each interval e_j. We can choose $x_{j+k/3-1} = x_{j-1} + \frac{k}{3}h_j$ for $k = 1, 2$ and the basis $\{\psi_i\}$ is defined by

$$\psi_i(x_j) = \delta_{i,j} \quad i, j = 0, 1/3, 2/3, 1, 4/3, \ldots, N.$$

On each interval e_j we now have four non-zero functions:

$$\begin{aligned}
\psi_{j-1}(x) &= -\tfrac{1}{2}(3t-1)(3t-2)(t-1), \\
\psi_{j-\frac{2}{3}}(x) &= \tfrac{9}{2}(3t-2)t(t-1), \\
\psi_{j-\frac{1}{2}}(x) &= -\tfrac{9}{2}(3t-1)t(t-1), \\
\psi_j(x) &= \tfrac{1}{2}(3t-1)(3t-2)t.
\end{aligned}$$

These functions are shown in Figure 3.3. Again the basis functions corresponding to the interior points have their support in a single interval e_j. Obviously, the dimension of the approximation space $P_3(\Omega_N)$ is $3N + 1$.

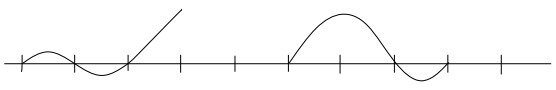


Figure 3.3: Third order Lagrange.

Remark:
For the same spaces $P_k(\Omega_N)$, $k = 1,\ 2,\ 3$, we can also choose different basis functions. One possibility is to use a *hierarchical basis*. This means that we start with the basis for the lower dimensional approximating space (smaller k or smaller N) and that we extend this basis to a basis of the current approximating space. For instance, we make such a basis for $k = 2$ by starting with the $N + 1$ basis functions already used as basis for the case $k = 1$. To this set we add one quadratic basis-function, for each interval e_j. We can take for instance $\psi_{j-1/2}$. Similarly the basis for $k = 3$ can be formed by adding one cubic basis function per interval to the (hierarchical) basis that was constructed for $k = 2$.

3.4.3 The construction of the discrete equations

As an example of a finite element discretisation we discuss the discretisation of a two-point boundary-value problem (TPBVP). We consider the following problem

$$\begin{aligned} -(a_2 u_x)_x + a_1 u_x + a_0 u &= s \quad \text{in} \quad \Omega, \\ u &= 0 \qquad\qquad\ \ \text{on} \quad \Gamma, \end{aligned}$$

where $\Omega = [a, b]$, $\Gamma = \partial\Omega$ and a_2, a_1, a_0 and s are functions on Ω. To apply the finite element method we first have to formulate the problem in its the weak form, so we seek an approximation for the function $u \in H_0^1(\Omega)$ that satisfies

$$(a_2 u_x, v_x) + (a_1 u_x, v) + (a_0 u, v) = (s, v),\ \forall\ v \in H_0^1(\Omega). \tag{3.44}$$

For our finite element discrete approximation we choose

$$S_h = \{u \in P_k(\Omega_N) \mid u(a) = u(b) = 0\} \subset H_0^1(\Omega),$$

and we take our weighting functions to be the same space, so $V_h = S_h$.

With $P_k(\Omega_N) = \text{Span}\,\{\psi_i\}$ we can write $u_h(x) = \sum_j c_j \psi_j(x)$. This yields the discrete system of $kN + 1$ equations

$$\sum_j c_j \int_\Omega a_2 \psi_j' \psi_i' + a_1 \psi_j' \psi_i + a_0 \psi_j \psi_i dx = \int_\Omega s \psi_i dx,$$

one equation for every weighting function ψ_i. Here the prime denotes differentiation with respect to x. The integration interval can be split into the subintervals e_j. We get

$$\sum_j c_j \{\sum_k \int_{e_k} a_2 \psi_j' \psi_i' + a_1 \psi_j' \psi_i + a_0 \psi_j \psi_i dx\} = \sum_k \int_{e_k} s \psi_i dx,$$

where we notice that the contributions from most of the subintervals vanish Thus the discrete equations are a linear system

$$\sum_j m_{ij} c_j = s_i, \tag{3.45}$$

in which

$$\begin{aligned} m_{ij} &= \sum_k \int_{e_k} a_2 \psi_j' \psi_i' + a_1 \psi_j' \psi_i + \psi_j \psi_i dx, \\ s_i &= \sum_k \int_{e_k} s \psi_i dx. \end{aligned}$$

From equation (3.45) the coefficients c_j are to be solved, to obtain the discrete solution

$$u_h(x) = \sum_j c_j \psi_j(x).$$

The matrix (m_{ij}) is called the *stiffness matrix* and the vector (s_i) the *load vector*. (m_{ij}) is composed of contributions from the separate intervals e_k. These contributions vanish if e_k is not a subset of the support of both ψ_i and ψ_j. The contribution of e_k to (m_{ij}) or (s_i) is called the *elementary stiffness matrix* or the *elementary load vector* corresponding to this interval.

In practice the linear system is built up by scanning every element in the partition and adding the elementary stiffness matrix and elementary load-vector to the full stiffness matrix or load-vector. This means that the matrix and vector elements are not computed one by one. This technique is practical not only for one-dimensional, but also for two and three-dimensional problems. It is called the *assembly* of the stiffness matrix and load vector.

Now we show explicitly how such a matrix/vector is drawn up for a two-point boundary-value problem. For simplicity we take the coefficients a_2, a_1 and a_0 constant. Define

$$\begin{aligned} m_{2,k,i,j} &= \int_{e_k} \psi_j' \psi_i' dx, \\ m_{1,k,i,j} &= \int_{e_k} \psi_j' \psi_i dx, \\ m_{0,k,i,j} &= \int_{e_k} \psi_j \psi_i dx, \end{aligned}$$

then the element matrix entries are

$$m_{ij} = \sum_k m_{k,i,j},$$

with

$$m_{k,i,j} = a_2 m_{2,k,i,j} + a_1 m_{1,k,i,j} + a_0 m_{0,k,i,j}.$$

Example 3.4.4

If we take piecewise linear basis-functions as in (3.43), for which $\psi_{k-1} = 1 - t$, $\psi_k = t$, $\psi'_{k-1} = -\frac{1}{h_k}$ and $\psi'_k = \frac{1}{h_k}$, we get

$$m_{k,i,j} = \frac{a_2}{h_k} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + a_1 \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} + a_0 h_k \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix}. \qquad (3.46)$$

Note that for convenience we have given only the essential part of the matrix: the elements $(k-1, k-1)$, $(k, k-1)$, $(k-1, k)$ and (k, k); all other elements vanish. [6]

The elementary stiffness matrices for higher order piecewise polynomials can be computed in a similar way.

Use of quadrature rules

The computation of the matrix entries is more laborious when the coefficients a_2, a_1 and a_0 are not (piecewise) constant. Only in special cases we can compute the integrals $m_{k,i,j}$ exactly. In most cases it is necessary to compute the integral using a *quadrature rule*:

$$\int_a^b f(x)\, dx \approx \sum_{m=0}^{l} w_m f(x_m),$$

where w_m is the weight for quadrature node x_m. Replacing the integral by a summation causes a new error, so that the accuracy of the quadrature may influence the accuracy of the discretisation. We usually want to use quadrature rules that preserve the order of accuracy of the discretisation method. The requirements for such a quadrature rule are discussed in section **3.5**.

An interesting quadrature in combination with Lagrangian interpolation is the method that uses the same nodal points for the integration. Assume that the approximation uses polynomials of degree l (so $V_h = P_l(\Omega_N)$) and take $y_{k,m}$, $m = 0, \ldots, l$ as the nodes in e_k:

$$x_{k-1} = y_{k,0} < y_{k,1} < \cdots < y_{k,l} = x_k.$$

Or rather,

$$y_{k,m} := x_{k+\frac{m}{l}-1},$$

so that

$$\psi_i(y_{k,m}) = \delta_{i,k+\frac{m}{l}-1}.$$

The elementary matrix entry becomes ($k = 0, \cdots, N$ and $i = n_i + \frac{m_i}{l} - 1$, $j = n_j + \frac{m_j}{l} - 1$ where $n_i, n_j = 0, \cdots, N$, $m_i, m_j = 0, \cdots, l$)

$$m_{k,i,j} \approx \sum_{m=0}^{l} w_m \left\{ a_2(y_{k,m}) \psi'_j(y_{k,m}) \psi'_i(y_{k,m}) \right.$$

[6] Be aware of the boundaries however, see a next section.

$$
\begin{aligned}
&+ a_1(y_{k,m})\psi_j'(y_{k,m})\psi_i(y_{k,m}) + a_0(y_m)\psi_j(y_{k,m})\psi_i(y_{k,m})\big\} \\
= &\sum_{m=0}^{l} w_m \big\{a_2(y_{k,m})\psi_j'(y_{k,m})\psi_i'(y_{k,m}) \qquad (3.47)\\
&+ a_1(y_{k,m})\psi_j'(y_{k,m})\delta_{i,k+m/l-1} + a_0(y_{k,m})\delta_{j,k+m/l-1}\delta_{i,k+m/l-1}\big\} \\
= &\sum_{m=0}^{l} w_m \big\{a_2(y_{k,m})\psi_j'(y_{k,m})\psi_i'(y_{k,m})\big\} \\
&+a_1(y_{k,m_i})\psi_j'(y_{n_i,m_i})\, w_{m_i} + a_0(x_i)\delta_{i,j} w_{m_i}.
\end{aligned}
$$

Example 3.4.5

In case of piecewise linear approximation ($l = 1$) and the trapezoidal rule ($w_0 = w_1 = \frac{1}{2}$), we get

$$
\begin{aligned}
m_{k,i,j} \;\approx\; & \frac{a_2(x_{k-1}) + a_2(x_k)}{2h_k} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \qquad (3.48)\\
&+\frac{1}{2}\begin{pmatrix} -a_1(x_{k-1}) & a_1(x_{k-1}) \\ -a_1(x_k) & a_1(x_k) \end{pmatrix} + \frac{h_k}{2}\begin{pmatrix} a_0(x_{k-1}) & 0 \\ 0 & a_0(x_k) \end{pmatrix}.
\end{aligned}
$$

We observe that the zero order term contribution is a diagonal matrix, which is different from formula (3.46) for constant coefficients. Using the quadrature also for the elementary load vector, we obtain

$$
s_{k,i} \approx \frac{h_k}{2}\begin{pmatrix} s(x_{k-1}) \\ s(x_k) \end{pmatrix}.
$$

Remark:
It is known that for an accurate quadrature we better take the nodes within each element $[x_{j-1}, x_j]$ not equidistant. To obtain the most accurate quadrature (and assuming that the endpoints of the interval are be nodal points) the nodes should be placed as for Lobatto quadrature. These points, then, may determine the nodal points for the Lagrange interpolation.

Example 3.4.6
As a last example, we give the construction of the element matrix based on piecewise quadratic functions and the Simpson quadrature ($l = 2$, $w_0 = w_2 = \frac{1}{6}$, $w_1 = \frac{2}{3}$). For simplicity we choose $a_2(x)$ piecewise constant on the partitioning Ω_N. We obtain

$$
\begin{aligned}
m_{k,i,j} \;\approx\; & \frac{a_{2,k}}{6h_k}\begin{pmatrix} 14 & -16 & 2 \\ -16 & 32 & -16 \\ 2 & -16 & 14 \end{pmatrix} \\
+ \;& \frac{1}{6}\begin{pmatrix} -3a_1(x_{k-1}) & 4a_1(x_{k-1}) & -a_1(x_{k-1}) \\ -4a_1(x_{k-\frac{1}{2}}) & 0 & 4a_1(x_{k-\frac{1}{2}}) \\ a_1(x_k) & -4a_1(x_k) & 3a_1(x_k) \end{pmatrix} \qquad (3.49)
\end{aligned}
$$

$$+ \quad \frac{h_k}{6} \begin{pmatrix} a_0(x_{k-1}) & 0 & 0 \\ 0 & 4a_0(x_{k-\frac{1}{2}}) & 0 \\ 0 & 0 & a_0(x_k) \end{pmatrix}.$$

For a_1 and a_0 also piecewise constant this formula simplifies to

$$m_{k,i,j} \approx \tfrac{a_{2,k}}{6h_k} \begin{pmatrix} 14 & -16 & 2 \\ -16 & 32 & -16 \\ 2 & -16 & 14 \end{pmatrix} + \tfrac{a_{1,k}}{6} \begin{pmatrix} -3 & 4 & -1 \\ -4 & 0 & 4 \\ 1 & -4 & 3 \end{pmatrix} + \\ + \tfrac{a_{0,k}h_k}{6} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.50}$$

Again the contribution of the zero-order term is a diagonal matrix. The elementary load-vector follows after direct computation, using the same Simpson quadrature

$$s_{k,i} \approx \frac{h}{6} \begin{pmatrix} s(x_{k-1}) \\ 4s(x_{k-\frac{1}{2}}) \\ s(x_k) \end{pmatrix}. \tag{3.51}$$

For our two-point boundary-value problem and piecewise second and higher-order functions the shape of the complete matrix and right-hand vector is given in Figure 3.4.

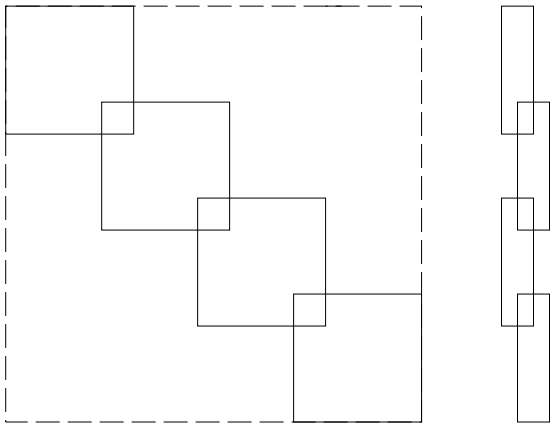


Figure 3.4: Matrix and right-hand-side constructed by means of C^0 piecewise cubic polynomials.

Static condensation, lumping

The solution of the general matrix system (3.45) is not a topic we will discuss here. See for instance [6]. We however make some remarks.

First notice, that due to the relatively small support of the basis functions, we have a linear system that is *sparse*: only a fairly limited amount of the matrix elements is non-zero.

The use of Lagrange interpolation has a nice advantage. We can reduce the set of equations in figure 3.4 to a tridiagonal form by so-called *static condensation*. With this we mean the elimination of the unknowns belonging to the internal nodes x_i, for which the basis functions ψ_i have a support that is completely contained in a single element e_k. The elimination of these unknowns is a strictly local process that can be done elementwise: it causes no changes in the contributions of other elements to the stiffness matrix and the load vector. The same technique can be applied for 2- or 3-dimensional problems, for each basis function $\psi_i(x)$ that has its support completely in a single element $\mathbf{e}_k$.

Let a linear system be partitioned as

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

and let the second set of variables and equations be available for elimination, then after elimination we get the system

$$(A_{11} - A_{12}A_{22}^{-1}A_{21})x_1 = b_1 - A_{12}A_{22}^{-1}b_2,$$

so there remains just a number of additional contributions in the matrix and right-hand-side.

When all interior variables are eliminated in this way we get the tridiagonal system of equations shown in figure 3.5.

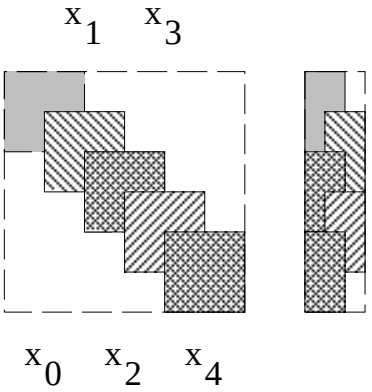


Figure 3.5: Matrix and right-hand side after lumping.

We have seen that the zero order (not differentiated) terms consist only of diagonal elements if we use a quadrature. The technique in which these terms are placed on the main diagonal is called *lumping* [7]. This technique is interesting

[7] Originally the off-diagonal elements where added to the main diagonal without any explanation.

for two reasons:

1. The part $a_0(x)u(x) = s(x)$ of the equation is discretised pointwise. This can be preferable when $a_0(x)$ or $s(x)$ is strongly varying or dominating.

2. For an initial-boundary-value problem (IBVP), when there is time dependence in the problem, we can use lumping together with the method of *semi-discretisation* to reduce the IBVP to an explicit set of ordinary differential equations. As an example we may consider a parabolic equation of the form $\varphi_t = L\varphi$, where L is a differential operator in the space variables (for instance $L\varphi = p\varphi_{xx} + q\varphi_x + r\varphi + s$), then a semi-discretisation with $S_h = V_h = \text{Span}\{\varphi_i\}$ yields the following set of equations: find $\varphi_h(x,t) = \sum_j c_j(t)\varphi_j(x)$ such that

$$(\frac{\partial}{\partial t}\varphi_h, \varphi_i) = (L\varphi_h, \varphi_i), \quad \forall \varphi_i.$$

So we have to compute the set $\{c_j(t)\}$ for which

$$\sum_j \frac{dc_j(t)}{dt}(\varphi_j, \varphi_i) = (L\varphi_h(t,\cdot), \varphi_i), \quad \forall \varphi_i.$$

The matrix (φ_j, φ_i) is called the *mass matrix*. By using the technique "lumping" we can recast it to a diagonal matrix $(\varphi_j, \varphi_i) = w_i\delta_{ij}$ so that we obtain an *explicit* set of ordinary differential equations for the coefficients $\{c_i(t)\}$:

$$\frac{dc_i(t)}{dt} = \frac{1}{w_i}(L\varphi_h(t,\cdot), \varphi_i), \ \forall\ i.$$

Treatment of boundary conditions

Until now we have limited ourselves to problems with homogeneous Dirichlet boundary conditions. In this section we treat also inhomogeneous and more general boundary conditions.

First we handle inhomogeneous Dirichlet boundary conditions. Consider $\Omega = [a,b]$ and the equation

$$Lu := -(a_2u_x)_x + a_1u_x + a_0u = s \quad \text{on} \quad \Omega, \tag{3.52}$$

$$u = g \quad \text{on} \quad \Gamma. \tag{3.53}$$

The weak form of this problem can be written as: find a $u \in H^1(\Omega)$ such that

$$u = g \ \text{ on } \ \Gamma$$

and

$$B(u,v) := (a_2u_x, v_x) + (a_1u_x, v) + (a_0u, v) = (s, v), \quad \forall v \in H_0^1(\Omega).$$

For theoretical purposes, this problem is easily reduced to the problem with homogeneous boundary conditions by choosing a function $g \in H^1(\Omega)$ (the extension of the g defined on Γ) that satisfies the boundary conditions [8]. Defining the function $w := u - g$ we have $w \in H_0^1(\Omega)$ and w satisfies the equation

$$Lw = s - Lg,$$

or,

$$B(w, v) = (s, v) - B(g, v).$$

Thus the problem is reduced to the problem with homogeneous boundary conditions. This gives existence and uniqueness.

In practice it is also easy to implement non-homogeneous Dirichlet boundary conditions. Because the value of the function u at the boundary points is given we have to compute less unknowns. This means that there are no discrete equations needed for the boundary points.

We can handle Dirichlet boundary conditions in two ways. To show this, we first assume the equations are constructed elementwise as shown in the previous section (figure 3.4). The two approaches are as follows.

I) Replace the equations for the boundary points by equations that force the boundary conditions. For example, in the stiffness matrix put 1 on the main diagonal and 0 at the other entries of this row, and put the boundary value in the right-hand-side. This forces the boundary value to be part of the solution. In this way we get a linear system of equations of the original form, of which the size corresponds with the total number of nodes.

II) In the other approach (see Figure 3.6) we eliminate the known "unknowns" on the boundary. The set of equations is reduced, but more important is that the right-hand-side has to be adjusted at a number of places. Now the size of the system corresponds with the interior nodes. In practice, this is less convenient, in particular if we have a combination of Dirichlet and Neumann conditions. In this case, however, a symmetric operator $B(u, v)$ guarantees a symmetric stiffness matrix.

If we consider more general boundary conditions than Dirichlet boundary conditions, we see that the whole procedure can be carried out in the same manner as before. In the discrete operator and right-hand-side additional terms will appear due to the boundary condition. See for instance the $B(u, \varphi)$ and the $l(\varphi)$ of section **3.3.5**. Every boundary integral gives a contribution to the elementary matrix/right-hand-side at each boundary point. In practice we first add all contributions from the interior elements to the stiffness-matrix and later the contributions from the boundary segments are added. Also these boundary integrals can be computed by quadrature.

[8] That we can actually find such a function puts an extra condition on the boundary condition and on the smoothness of the boundary Γ. With γ_0 the trace operator we have the requirement that

$$\gamma_0 u = \gamma_0 g.$$

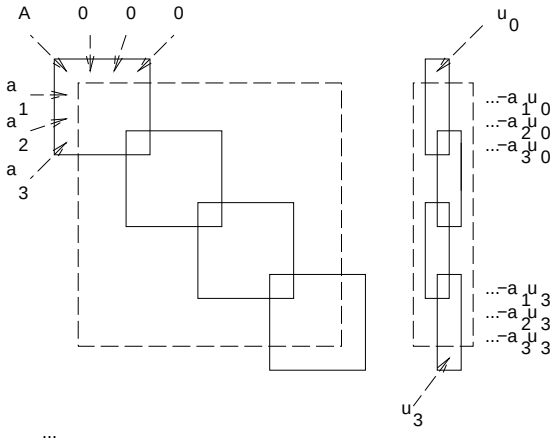


Figure 3.6: Treatment of boundary conditions.

3.4.4 Other piecewise polynomial approximations in one dimension

In every finite element method we use an approximation of the solution of the form

$$u_h(x) = \sum_j c_j \varphi_i(x),$$

where the use of continuous, piecewise polynomial basis functions $\varphi_i(x)$ with a small support is preferred. Not only Lagrange interpolation but also Hermite interpolation can be used. Here the functions $\varphi_i(x)$ are not only determined by their *function values* at the nodes, but also by the *values of the derivatives*. The approximating basis functions thus may have larger smoothness, which gives us the opportunity to solve problems which contain higher order derivatives. We may find solutions in $H_0^k(\Omega)$ for $k > 1$. For instance it allows the solution of the homogeneous Dirichlet problem for the biharmonic operator Δ^2:

$$\begin{array}{lll} \Delta^2 u = f & \text{in} & \Omega, \\ u = \partial_n u = 0 & \text{on} & \Gamma. \end{array}$$

A polynomial of degree k needs $k+1$ (independent) values on an interval to be defined correctly.

Definition 3.4.7 Let f be a function in $\mathcal{C}^{m-1}([a,b])$, then the *Hermite interpolate* of f on a partition Ω_N is the function $p \in P_{2m-1}(\Omega_N) \cap \mathcal{C}^{m-1}([a,b])$ for which, for all $x_i \in \Omega_N$ and $k = 0, 1, \ldots, m-1$,

$$D^k p(x_i) = D^k f(x_i).$$

Note that the Hermitian interpolation of a polynomial q of degree $2m-1$ is the polynomial itself. Moreover it is not just an element of $\mathcal{C}^{m-1}([a,b])$, but even from $\mathcal{C}^{2m-1}([a,b])$.

A usual Hermite-basis for a finite element method in one space dimension is that of the cubic Hermite splines (m=2): on every interval $[x_i, x_{i+1}]$ the $P_3(\Delta)$-function is determined by four parameters $f(x_i)$, $f'(x_i)$, $f(x_{i+1})$ and $f'(x_{i+1})$. For every element $e_j = [x_j, x_{j+1}]$ there are four non-zero basis-functions: $\varphi_{0,j}$, $\varphi_{1,j}$, $\varphi_{0,j+1}$ and $\varphi_{1,j+1}$ such that $D^k\varphi_{l,j}(x_i) = \delta_{kl}\delta_{ij}$. Let $t = (x - x_j)/(x_{j+1} - x_j)$ then the form of these functions is on element e_j

$$\begin{aligned}
\varphi_{0,j}(t) &= (t-1)^2(2t+1),\\
\varphi_{0,j+1}(t) &= t^2(3-2t),\\
\varphi_{1,j}(t) &= t(t-1)^2(x_{j+1}-x_j),\\
\varphi_{1,j+1}(t) &= t^2(t-1)(x_{j+1}-x_j).
\end{aligned}$$

Extending these functions to the entire interval $[a, b]$ such that $D^k\varphi_{l,j}(x_i) = \delta_{kl}\delta_{ij}$ we get functions $\varphi_{j,k}(x) \in H^2([a, b])$, because the derivatives are continuous. Except for the functions that are associated with the boundary points, all functions $\varphi_{l,j}$ have a support that extends over two subintervals e_k.

The smoothness of the approximating function $u_h(x)$ implies that it can be used for solving fourth order elliptic problems like

$$\begin{aligned}
au_{xxxx} - bu_{xx} + cu &= f, \qquad (3.54)\\
u(a) = u'(a) = u(b) &= u'(b) = 0,
\end{aligned}$$

of which the corresponding variational formulation is as follows: find $u \in H_0^2(\Omega)$ such that

$$\int_\Omega a\, u_{xx}\varphi_{xx} + b\, u_x\, \varphi_x + cu\, \varphi\, dx = \int f\, \varphi\, dx,$$

for all $\varphi \in H_0^2(\Omega)$.

This problem can not be solved by Lagrangian interpolation because the space spanned by the Lagrange interpolation polynomials $\{\varphi_i\}$ is not in $H^2(\Omega)$.

The discrete set of equations for the two point boundary value problem discretised with Hermite interpolation polynomials has a block-tridiagonal form (see Figure 3.7).

Exercise 3.4.8
Consider the fourth order problem ($\Omega = [0, 1]$)

$$\begin{aligned}
u_{xxxx} - 2u_{xx} + u &= 1,\\
u(0) = u(1) = u_x(0) &= u_x(1) = 0.
\end{aligned}$$

Prove that the analytic solution is

$$\begin{aligned}
u(x) &= 1 + e^x(c_1 + c_2x) + e^{-x}(c_3 + c_4x),\\
c_1 &= -\frac{(2e-1)(e^2-2e-1)}{e^4-6e^2+1},\\
c_2 &= \frac{(e-1)(e^2-2e-1)}{e^4-6e^2+1},\\
c_3 &= -1 - c_1,\\
c_4 &= c_3 - c_1 - c_2.
\end{aligned}$$

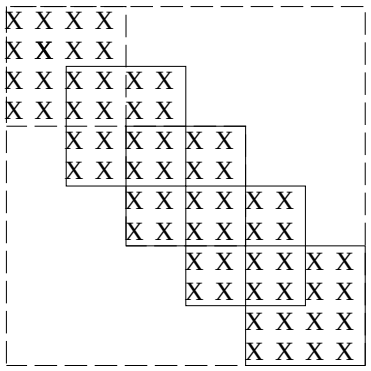


Figure 3.7: Block-tridiagonal matrix.

Make a uniform discretisation with $h = 1/N$, compute and solve the linear system. This gives discrete approximations for $u(x_i)$ and $u'(x_i)$. Compute

$$\max_i |u(x_i) - u_h(x_i)|$$

and

$$\max_i |u'(x_i) - u'_h(x_i)|,$$

for $N = 10, 20, 40$. Compare the order of convergence you observe with the order you expected.

Other piecewise polynomials that might be used are for instance the *splines*. These are $P_m(\Delta) \cap \mathcal{C}^{m-1}([a,b])$-functions. Such functions have a single degree of freedom for each interval and they have a maximum smoothness for a polynomial of degree m. However, in general, the support of these basic spline stretches over many adjacent elements. Application for a finite element method is therefore unusual, because it means that many non-zero elements will appear in each row of the stiffness matrix. In other words: the use of splines causes a non-sparse system of linear equations (although in the one-dimensional case it will be a band system).

3.4.5 Approximation in two dimensions

In the introduction to discretisation methods we have seen some possible representations of a two-dimensional function. Not all of them are useful for the solution of second order elliptic problems by a (conforming) finite element method. The most important which remain are shown in Figure 3.8. A dot indicates that at this point we take the function value, a circle that we also take the value of the derivative, two circles the value of two derivatives, etc.

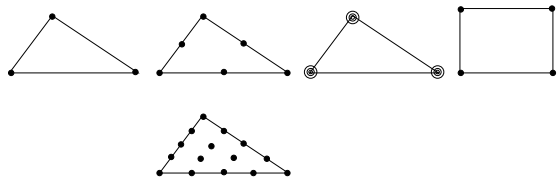


Figure 3.8: Some representations of a two-dimensional function.

Among these are the piecewise Lagrange interpolation polynomials on a triangulation. Notice that the basis-functions defined by $\varphi_i \in P_k(\Omega_N)$ (functions with $|\alpha| \leq k$) and $\varphi_i(x_j) = \delta_{ij}$ are indeed continuous.

Example 3.4.9
On the standard triangle shown in figure 3.9 we find basis functions for a $P_2(\Delta)$ Lagrange approximation by

$$\begin{aligned}
\Phi_1(x,y) &= 2\ (x+y-1)\ (x+y-\tfrac{1}{2}),\\
\Phi_2(x,y) &= 2x\ (x-\tfrac{1}{2}),\\
\Phi_3(x,y) &= 2y\ (y-\tfrac{1}{2}),\\
\Phi_4(x,y) &= 4x\ y,\\
\Phi_5(x,y) &= -4y\ (x+y-1),\\
\Phi_6(x,y) &= -4x\ (x+y-1).
\end{aligned}$$

That the dimension of $P_2(\Delta)$ is indeed 6 can be easily seen because any element is of the form

$$p(x,y) = \sum_{|\alpha|\leq 2} a_\alpha\ x^{\alpha_1}\ y^{\alpha_2}.$$

This is only possible for the set of values

$$\alpha = (\alpha_1, \alpha_2) \in \{(0,0),(1,0),(0,1),(2,0),(0,2),(1,1)\}.$$

To express the symmetry in this formulas more clearly, we may use *barycentric coordinates* (r,s,t) defined by $r = x$, $s = y$ and $t = 1 - x - y$. [9] Then we get

$$\begin{aligned}
\Phi_1(r,s,t) &= t\ (2t-1),\\
\Phi_2(r,s,t) &= r\ (2r-1),\\
\Phi_3(r,s,t) &= s\ (2s-1),\\
\Phi_4(r,s,t) &= 4\ r\ s,\\
\Phi_5(r,s,t) &= 4\ s\ t,\\
\Phi_6(r,s,t) &= 4\ r\ t.
\end{aligned} \tag{3.55}$$

[9] So: $t = 0 \Rightarrow x + y = 1$, $r = 0 \Rightarrow x = 0$ and $s = 0 \Rightarrow y = 0$. Which give the three boundaries of the triangle.

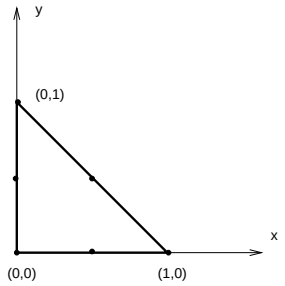


Figure 3.9: Lagrange for a triangle.

These functions are also used for isoparametrical elements, which we shall discuss in the following section.

The number of degrees of freedom for a general k-th order polynomial in 2 or 3 dimensions is exactly the number of points that can be distributed nicely over a triangle or tetrahedron. If, on a rectangle or a parallelepiped, a number of nodes is distributed in a regular way we will not be able to construct exactly all k-th order polynomials. For a parallelepiped one usually takes polynomials of the form $P_k(x) \cdot P_k(y)$ (so the tensor product of the two spaces) and for the nodes one takes the Cartesian product of the one-dimensional partitions.

Example 3.4.10
We consider the bilinear interpolation on a rectangle. A bilinear basis function is of the form

$$a + bx + cy + dxy,$$

or,

$$(a_1 + a_2 x)(b_1 + b_2 y).$$

The first form shows clearly that such a function has four degrees of freedom on each rectangle and the second form shows that the space of bilinear functions is the product of the two spaces of linear functions.

Having four degrees of freedom, such a bilinear function is completely determined by the function values at the vertices of the rectangle.

Note that in two dimensions, on triangles, the piecewise Hermite interpolation polynomials are continuous, but are not in $\mathcal{C}^1(\Omega)$. If we want to find a piecewise polynomial in $\mathcal{C}^1(\Omega)$, then we have to use a relatively intricate construction. One of the more simple ones is a $P_5(\Omega)$-function which has 21 degrees of freedom for each triangle (the Argyris triangle, cf [4, p. 71], see Figure 3.10).
Note that the function u along a boundary is a 5th order function determined by its function value and first and second order derivatives at the vertices and the value of $\frac{\partial u}{\partial \mathbf{n}}$ ($\mathbf{n}$ is the outward unit normal) at the edges midpoints.

The construction of the discrete equations and the treatment of the boundary

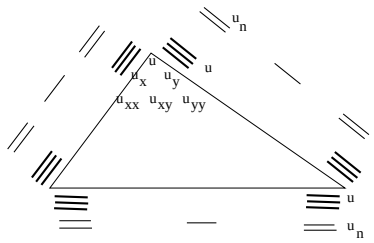


Figure 3.10: The Argyris triangle.

conditions is in principle similar to the one-dimensional case: the matrix and the right-hand-side will be build up elementwise. To handle the boundary conditions we have to compute boundary integrals. When basis functions $\phi_j(x)$ have a support that is completely contained in a single element, then we can eliminate the corresponding variables by *static condensation*. When computing the elementary matrices we again use a quadrature and also the combination of the elementary basis-functions for Lagrange interpolation and an associated quadrature can lead to *lumping*.

3.4.6 Isoparametric elements

By triangulation of the domain, the solution of a problem on polygonal areas can be performed in a relatively simple way. However when there is a *curved boundary* we can approximate it by means of *isoparametric elements*. To define these elements we first introduce the following

Definition 3.4.11 Let $v = \{z_1, \ldots, z_{N_k}\}$ be a set of N_k points in $\mathbb{R}^2$. Then v is called a *k-unisolvent set* if for every sequence of real numbers $(\alpha_1, \ldots, \alpha_{N_k})$ there is precisely one polynomial of degree $\leq k$ such that

$$P_k(z_i) = \alpha_i, \quad i = 1, \ldots, N_k.$$

Examples of this are: for $k = 1$: the vertices of a triangle; $k = 2$: the vertices and the midpoints of the edges. Examples for $k > 2$ are easily found.

An isoparametric element of degree $k = 2$ is defined as follows. Let $v = \{\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_6\}$ be a 2-unisolvent set in $\mathbb{R}^2$. The matching isoparametric element is

$$I_\Delta = \{(x,y) | (x,y)^T = \sum_{i=1}^{6} \vec{z}_i \Phi_i(r,s,t);\ \ 0 \leq r,s,t \leq 1;\ \ r+s+t = 1\},$$

where r, s and t are the barycentric coordinates as introduced in Section 3.4.5 and Φ_i the functions defined in (3.55). On the isoparametric element we now have a parametrisation in r, s, t, with $r + s + t = 1$, and basis functions can be

defined on this element like it was done in case of triangular elements

$$u_h(x)|_{I_\Delta} = \sum_{i=1}^{6} a_i \Phi_i(r,s,t).$$

We see that with k-th order isoparametric elements the boundary is approximated by piecewise k-th order polynomials.

The dependent and independent variables (u and (x,y) respectively) are both parametrised by the same type of functions. This is the reason to call these elements isoparametric elements.

3.5 Error estimates for the finite element method

In this section we derive error estimates for the finite element method. First we will give the discrete version of the Generalised Lax-Milgram theorem. It gives the uniqueness of the solution to the discrete equation and it gives a first estimate of the error.

This theorem is in fact only applicable when we use finite dimensional subspaces of our original Hilbert spaces. We have a more general case when we don't have such subspaces or when the operators in the variational equation are replaced by approximations (for instance by quadrature). We give also error estimates for this case.

The error estimates depend on how good we can interpolate elements of Banach spaces in subspaces of these Banach spaces, so we have to discuss the interpolation theory in Banach spaces, preceded by a necessary discussion of the formalism of the finite element method. This will give us estimates in the Sobolev norms $\|.\|_{m,q,\Omega}$. We will also give an estimate in the L^2-norm, but for this we need additional requirements on the problem we consider.

We also discuss pointwise convergence and superconvergence, i.e the phenomenon that at some particular points the accuracy of the approximation is better than the global error estimates indicate. Finally we study the influence of the use of particular quadrature rules on Banach space interpolation and on superconvergence.

3.5.1 The discrete version of the Generalised Lax-Milgram

In Section **3.3.2** we saw that -under certain conditions- a variational problem has a solution, which is bounded by the right-hand-side of the equation. This Generalised Lax-Milgram Theorem **3.3.18** can be applied not only to continuous problems, but also to the associated discrete problems, with trial functions in $S_h \subset S$ and test functions in $V_h \subset V$. Taking these discrete spaces as subspaces of the continuous spaces we obtain a so called *conforming finite element methods*.

In the following theorem we show that the discrete solution obtained in this way is "quasi-optimal". This means that, apart from a constant factor, it is (in the S-norm) not worse than the best approximation of the solution $u \in S$ in the discrete space S_h.

Theorem 3.5.1 (Discrete Generalised Lax-Milgram)
Let the requirements of the Generalised Lax-Milgram Theorem **3.3.18** be satisfied. So, let X^1 and X^2 be two Hilbert spaces and let $F : X^1 \times X^2 \to \mathbb{R}$ be a bilinear functional that is bounded and sub-coercive. Further, let $X_h^1 \subset X^1$ and $X_h^2 \subset X^2$ be linear subspaces, such that

$$\forall x_h \in X_h^1 \quad \exists z_h \in X_h^2,\ z_h \neq 0 \quad |F(x_h, z_h)| \geq \gamma_h \|x_h\|\ \|z_h\|,$$

and

$$\forall z_h \in X_h^2,\ z_h \neq 0 \quad \exists x_h \in X_h^1 \quad |F(x_h, z_h)| > 0.$$

Finally, let $f \in (X^2)'$, and $u_0 \in X^1$ be the solution of

$$F(u_0, v) = f(v) \quad \text{for all } v \in X^2.$$

Then there exists a unique $u_h \in X_h^1$, the solution of the discrete problem

$$F(u_h, v_h) = f(v_h) \ \text{ for all } v_h \in X_h^2,$$

and the following error estimate holds

$$\|u_h - u_0\|_{X^1} \leq \left[1 + \frac{\|F\|}{\gamma_h}\right] \inf_{x_h \in X_h^1} \|u_0 - x_h\|_{X^1}. \tag{3.56}$$

Proof: Let $R : X^1 \to X^2$ be the Riesz mapping as used in the proof of the Generalised Lax-Milgram Theorem **3.3.18**

$$(Ru, v)_{X^2} = F(u, v), \ \text{ for all } \ u \in X^1,\ v \in X^2,$$

and let $S : X_h^1 \to X_h^2$ be the analogue mapping for X_h^1 and X_h^2, i.e.,

$$(Su_h, v_h)_{X^2} = F(u_h, v_h), \ \text{ for all } \ u_h \in X_h^1, v_h \in X_h^2.$$

Then $\|R\| = \|F\|$ and $\|S^{-1}\| \leq \gamma_h^{-1}$.

Let $P_h^1 : X^1 \to X_h^1$ and $P_h^2 : X^2 \to X_h^2$ be orthogonal projections

$$\begin{array}{ccc} X^1 & \xrightarrow{R} & X^2 \\ {\scriptstyle P_h^1}\downarrow & & \downarrow{\scriptstyle P_h^2}. \\ X_h^1 & \xrightarrow{S} & X_h^2 \end{array}$$

i) Consider point (5) in the GLM theorem. With $v_0 \in X^2$ such that $f(v) = (v_0, v)_{X^2}$ for all $v \in X^2$ (Riesz) and $v_h \in X_h^2$ such that $f(v) = (v_h, v)_{X^2}$ for all $v \in X_h^2$ (Riesz) we have $(v_0 - v_h, P_h^2 v) = 0$ for all $v \in X^2$; so Corollary **3.3.4** gives that $P_h^2(v_0 - v_h) = 0$. It follows that $P_h^2 v_0 = P_h^2 v_h = v_h$, so $v_h = P_h^2 v_0$.

From Lax Milgram we know that $u_0 = R^{-1} v_0$ and $u_h = S^{-1} v_h$, so that $u_h = S^{-1} P_h^2 R u_0$.

ii) For every $x_h \in X_h^1$ we have $S x_h^1 = P_h^2 R x_h^1$ because, for all $v_h \in X_h^2$,

$$\begin{aligned} (P_h^2 R x_h^1, v_h) &= (R x_h^1, P_h^2 v_h) = (R x_h^1, v_h) \\ &= F(x_h^1, v_h) = (S x_h^1, v_h). \end{aligned}$$

We want to know how well the discrete solution approximates the solution of the continuous problem, so we are interested in $\|u_h - u_0\|$. First we look at

$$\begin{aligned}
\|u_h - P_h^1 u_0\| &= \|S^{-1} P_h^2 R u_0 - P_h^1 u_0\| \\
&\leq \|S^{-1}\| \, \|P_h^2 R u_0 - S P_h^1 u_0\| \\
&= \|S^{-1}\| \, \|P_h^2 R u_0 - P_h^2 R P_h^1 u_0\| \\
&\leq \|S^{-1}\| \, \|P_h^2\| \, \|R\| \, \|u_0 - P_h^1 u_0\| \\
&\overset{(*)}{=} \|S^{-1}\| \, \|R\| \, \|u_0 - P_h^1 u_0\|.
\end{aligned}$$

In $(*)$ we used that $\|P_h^2\| = 1$ by Corollary **3.3.4**. Hence

$$\begin{aligned}
\|u_h - u_0\| &\leq \|u_h - P_h^1 u_0\| + \|P_h^1 u_0 - u_0\| \\
&\leq \{1 + \|S^{-1}\|\|R\|\} \, \|u_0 - P_h^1 u_0\|.
\end{aligned}$$

From this easily follows

$$\|u_h - u_0\|_{X^1} \leq \{1 + \|F\|/\gamma_h\} \inf_{x_h \in X_h^1} \|u_0 - x_h\|_{X^1}.$$

■

Corollary 3.5.2
Given a problem that satisfies the conditions of the Generalised Lax Milgram Theorem, then for a conforming finite element method a sufficient condition for convergence is the existence of a family of subspaces (X_h^1) of X^1 such that, for each $u \in X^1$,

$$\lim_{h \to 0} \inf_{x_h \in X_h^1} \|u - x_h\| = 0.$$

So the problem we were facing is reduced to an approximation problem: evaluate the distance $d(u, X_h^1) = \inf_{x_h \in X_h^1} \|u - x_h\|$ between a function $u \in X^1$ and a subspace $X_h^1 \subset X^1$. Thus, if we can prove that $d(u, X_h^1) = \mathcal{O}(h^k)$, for some $k > 0$, we have as an immediate consequence that

$$\|u - u_h\| = \mathcal{O}(h^k)$$

We then will call k the *order of convergence*.

The norm in which the error is measured is more or less induced by the problem: the conditions on coercivity and continuity are stated in the X^1-norm. For an elliptic problem of order $2m$ this will usually be the $H^m(\Omega)$-norm. This will be discussed later.

Measuring the error in other norms (and especially the L^2-norm) requires a further analysis of the problem.

3.5.2 More general error estimate

In this section we study an error estimate for more general problems. We can approximate both the bilinear form B and the linear form f by discrete versions B_h and f_h. These discrete operators will work on some spaces V_h and S_h that are now not necessarily contained in V and S. This means we have to expect a supplementary *consistency error* in our estimates. The estimate thus obtained will also be valid for non-conforming finite element methods.

We consider a solution $u^* \in S \subset S^*$ (S^* a normed linear space) of the problem: find $u \in S$ such that

$$B(u,v) = f(v), \quad \forall v \in V. \tag{3.57}$$

Let $B_h : S^* \times V^* \to \mathbb{R}$ and $f_h : V^* \to \mathbb{R}$. We also consider the discrete problem associated with (3.57): find $u_h \in S_h$ such that

$$B_h(u_h, v_h) = f_h(v_h), \quad \forall v_h \in V_h, \tag{3.58}$$

in which $S_h \subset S^*$ and $V_h \subset V^*$, but not necessarily $S_h \subset S$, $V_h \subset V$.

Note that we have assumed no relation between B and B_h or f and f_h. What can we say about $\|u_h - u^*\|_{S^*}$?

Theorem 3.5.3 Assume there exists a $u_h \in S_h$ that satisfies equation (3.58). Let $B_h : S^* \times V^* \to \mathbb{R}$ be bounded with $\|B\| = M_h$, so that

$$\exists M_h > 0 \ \ \forall s \in S^*,\ v \in V^* \quad |B_h(s,v)| \le M_h \|s\|_{S^*} \|v\|_{V^*}.$$

Let $B_h : S_h \times V_h \to \mathbb{R}$ be coercive:

$$\exists \alpha_h > 0 \ \ \forall s_h \in S_h \ \exists v_h \in V_h \quad \alpha_h \|s_h\|_{S^*} \|v_h\|_{V^*} \le |B_h(s_h, v_h)|.$$

Then the following estimate is valid

$$\begin{aligned} \|u_h - u^*\|_{S^*} \quad &\le \quad \left[1 + \frac{M_h}{\alpha_h}\right] \inf_{s_h \in S_h} \|u^* - s_h\|_{S^*} \\ &+ \quad \frac{1}{\alpha_h} \sup_{v_h \in V_h} \frac{|B_h(u^*, v_h) - f_h(v_h)|}{\|v_h\|_{V^*}}. \end{aligned} \tag{3.59}$$

Proof: In the proof norms are either $\|.\|_{S^*}$ or $\|.\|_{V^*}$.
Because B_h is coercive we know that for each $s_h \in S_h$ there exists a $w_h \in V_h$ such that

$$\begin{aligned} \alpha_h \|u_h - s_h\| \ \|w_h\| \quad &\le \quad |B_h(u_h - s_h, w_h)| \\ &= \quad |B_h(u^* - s_h, w_h) - B_h(u^*, w_h) + B_h(u_h, w_h)| \\ &\le \quad |B_h(u^* - s_h, w_h)| + |B_h(u_h, w_h) - B_h(u^*, w_h)| \\ &\le \quad M_h \|u^* - s_h\| \ \|w_h\| + |B_h(u^*, w_h) - f_h(w_h)|, \end{aligned}$$

because u_h is a solution to (3.58). It follows that for arbitrary $s_h \in S_h$

$$\|u_h - s_h\| \leq \frac{M_h}{\alpha_h}\|u^* - s_h\| + \frac{1}{\alpha_h}\frac{|B_h(u^*, w_h) - f_h(w_h)|}{\|w_h\|}.$$

Now choose an arbitrary $s_h \in S_h$ then

$$\begin{aligned}\|u_h - u^*\| &\leq \|u_h - s_h\| + \|s_h - u^*\| \\ &\leq \left[1 + \frac{M_h}{\alpha_h}\right]\|u^* - s_h\| + \frac{1}{\alpha_h}\frac{|B_h(u^*, w_h) - f_h(w_h)|}{\|w_h\|}.\end{aligned}$$

From this the theorem follows. ∎

We see that the error is determined by

1. *the conditioning of the discrete problem:* M_h/α_h;
2. *the interpolation error:* $\inf_{s_h} \|u^* - s_h\|_{S^*}$;
3. *the consistency error:* $|B_h(u^*, v_h) - f_h(v_h)|$.

Remarks:

- In the proof of the theorem we have not used the fact that u^* is a solution to (3.57). However we can expect the consistency error to be rather big if we choose a u^* that is not a solution.
- The conditioning of the problem can also be formulated as: there exist $\alpha_h > 0$ and $M_h > 0$ so that for all $s_h \in S_h$

$$\alpha_h\|s_h\|_{S_h} \leq \sup_{v_h \in V_h}\frac{|B_h(s_h, v_h)|}{\|v_h\|_{V_h}} \leq M_h\|s_h\|_{S_h}.$$

- If the discrete spaces *are* embedded in continuous spaces , i.e. if $S_h \subset S$ and $V_h \subset V$, and if B_h is the restriction of a $B : S \times V \to \mathbb{R}$ and f_h the restriction of a $f : V \to \mathbb{R}$ to these subspaces, i.e. if

$$B_h(u_h, v_h) = B(u_h, v_h) \text{ and } f_h(v_h) = f(v_h),\ \forall u_h \in S_h,\ \forall v_h \in V_h,$$

then

$$B_h(u^*, w_h) - f_h(w_h) = B(u^*, w_h) - f(w_h) = 0,\ \forall w_h \in V_h.$$

It follows that in this case the estimate is reduced to the result from Theorem **3.5.1**. A difference is the fact that for the present estimate we have assumed the existence of the solution u^*.

It turns out from theorem **3.5.3** that, apart from studying interpolation theory for Banach spaces, we also have to study the approximation of bilinear and linear operators. In this we will restrict ourselves to errors caused by quadrature. See Section **3.5.6**.

3.5.3 The formalisation of the finite element method

We do not want to describe the theory of interpolation by piecewise polynomials in all detail (for instance: we will prove few theorems) but we will show some of the main results in the next section. To gain some insight in the derivation of these results, we show in this section a formal way to introduce finite elements.

Let $\Omega \subset \mathbb{R}^d$ be an open bounded domain with a Lipschitz continuous boundary $\partial\Omega$. A *partition* Δ_h of Ω is defined as

$$\Delta_h = \{\ \Omega_e \quad | \quad \bigcup_e \overline{\Omega_e} = \overline{\Omega},\ \Omega_{e_i} \cap \Omega_{e_j} = \emptyset \text{ if } i \neq j,$$
$$\Omega_e \neq \emptyset,\ \partial\Omega_e \text{ Lipschitz continuous}\}.$$

Let $u \in \mathcal{C}^m(\overline{\Omega})$ with $m \geq 0$. For a *local finite element approximation of u on* $\Omega_e \in \Delta_h$ we choose a set of functions $P_e = \{\phi_j^e\}_{j=1,\cdots,N_e}$ and we approximate u on $\overline{\Omega_e}$ by

$$u_h(x) = \sum_{j=1}^{N_e} a_j^e \phi_j^e(x) \ \text{ for } x \in \overline{\Omega_e}.$$

As a rule we choose $\{\phi_j^e\}$ such that

1. there exists a set of *nodal points* of Ω_e: $\{b_j^e \mid b_j^e \in \overline{\Omega_e},\ i = 1, \cdots, N_e\}$ and
2. there exists a set of derivatives D^{α_i}, where α_i are multi-integers, such that

$$D^{\alpha_i}\phi_j^e(b_i^e) = \delta_{ij}. \tag{3.60}$$

Usually we choose $\{\phi_j^e\}$ such that it contains all polynomials of degree k

$$P_k(\overline{\Omega_e}) \subset \ \text{Span}(\phi_j^e).$$

We formalise the above construction.

Definition 3.5.4 Let P be a set of real-valued functions defined over a domain Ω. Let L be a finite set of linearly independent linear functionals l_i, $i = 1, \cdots, N$ over P. Then L is *P-unisolvent* if for every set of scalars α_i, $i = 1, \cdots, N$, there is a unique $p \in P$ such that

$$l_i(p) = \alpha_i.$$

Corollary 3.5.5
Let L be a P-unisolvent set then there exists a unique set of piecewise polynomials $\{p_j,\ j = 1, \cdots, N,\ p_j \in P\}$ such that,

$$l_i(p_j) = \delta_{ij}. \tag{3.61}$$

Example 3.5.6
Take $\Omega = [0, 1]$ and make a partition $0 = x_1 < x_2 < \cdots < x_{N+1} = 1$. Let $\Omega_i = [x_i, x_{i+1}]$, $P_{\Omega_i} = P_1(\Omega_i)$ and $L_{\Omega_i} = \{l_i, l_{i+1}\}$, with $l_i(f) = f(x_i)$ and $l_{i+1}(f) = f(x_{i+1})$. Then $\{L_{\Omega_i}\}$ is a $\{P_{\Omega_i}\}$-unisolvent set: take any two numbers α and β then there is one and only one polynomial $p \in P_{\Omega_i}$ such that

$$p(x_i) = \alpha \text{ and } p(x_{i+1}) = \beta.$$

Definition 3.5.7 A *finite element* is a triple $\{\Omega_e, P_e, L_e\}$ in which

$$P_e = \{\phi_j^e \in \mathcal{C}^\infty(\overline{\Omega_e}),\ j = 1, \cdots, N_e\}$$

are the *basis functions of the finite element* and the P_e-unisolvent set

$$L_e = \{l_i^e \in \left[\mathcal{C}^m(\overline{\Omega_e})\right]',\ i = 1, \cdots, N_e\}$$

gives the *degrees of freedoms* of the finite element.

Definition 3.5.8 A *global finite element approximation* of a function u on $\overline{\Omega}$ is defined by

$$u_h(x) = \sum_{j=1}^{N} a_j \phi_j(x)$$

for $x \in \overline{\Omega}$, where we choose the set $P = \{\phi_j\}$ such that

1. there exists a partition $\Delta_h = \{\Omega_e\}_e$ of Ω;

2. for every $\Omega_e \in \Delta_h$ there is a finite element $\{\Omega_e, P_e, L_e\}$;

3. by $L = \cup_e L_e$ we denote the set of *degrees of freedom of the global approximation*. The number of (different) elements in L is $N = \dim(L)$, so $N \leq \sum_e N_e$, and N is the total number of degrees of freedom;

4. by P we denote the set of basis functions of the global approximation; P is the set of N functions given by

$$P = \{\phi_j \ \mid \exists e \ \ni, \phi_j|_{\Omega_e} \in P_e\ ;\ l_k(\phi_m) = \delta_{km}\}\,.$$

The P-unisolvent set L of degrees of freedom implies a natural P-interpolation of a $\mathcal{C}^m(\overline{\Omega})$- function

$$\Pi : \mathcal{C}^m(\overline{\Omega}) \to \mathrm{Span}(P). \tag{3.62}$$

This Π is defined as

$$v \to \Pi v = \sum_j l_j(v)\phi_j. \tag{3.63}$$

Π is a projection because $l_i(\Pi v) = l_i(v)$ for all $l_i \in L$ and any $v \in \mathcal{C}^m(\overline{\Omega})$.

Usually we take the same type of L_e and P_e on every Ω_e in Δ. So we take a set of finite elements $\{\Omega_e, P_e, L_e\}$, that are all is equivalent to a *reference* or *master element* $(\hat{\Omega}, \hat{P}, \hat{L})$. Then all finite elements are constructed from $(\hat{\Omega}, \hat{P}, \hat{L})$ by an affine or isoparametric transformation. Error estimates for the master element are then easily transferred to every element.

3.5.4 Interpolation theory and applications

Interpolation theory in Sobolev spaces

In Section **3.5.3** we saw the construction of an interpolation operator $\Pi : \mathcal{C}^m(\overline{\Omega}) \rightarrow \text{Span}(P)$. It is clear that a function on an area Ω is approximated better if more degrees of freedom are available. This can be done either by making the elements Ω_e smaller or by taking a higher order of approximation.

In this section we choose for the first option and study the quality of the approximation if we take smaller and smaller elements Ω_e. In fact, we consider a sequence $\{\Delta_h\}_{h\rightarrow 0}$ of partitions, in which h denotes the largest diameter of the elements $\Omega_e \in \Delta_h$. Further we assume that the refinement of the whole area will take place in a regular way so that also the smallest of the inscribed circles of the elements in the partition will be $\mathcal{O}(h)$.

On each partition Δ_h we assume a global finite element approximation and we denote by $S_h = \text{Span}(P)$ the space of approximating functions. In this way we also introduce a sequence of projections $\{\Pi_h\}_{h\rightarrow 0}$ with $\Pi_h : \mathcal{C}^m(\overline{\Omega}) \rightarrow S_h$. With an appropriate choice of the functionals in L (imposed by Sobolev's Lemma), the operator $\Pi_h : W^{l,p}(\Omega) \rightarrow S_h$ can be defined

$$\Pi_h u = \sum_j l_j(u)\phi_j.$$

We are interested in the behaviour of the *error of approximation* $\|u - \Pi_h u\|$ for $h \rightarrow 0$. A result is given in theorem **3.5.18** below. The proof of this theorem and a number of preceding lemmas will not be given explicitly. They can be found in [4], or its more recent version [18].

The plan of our treatment is as follows. First we define on each element in the partition a projection operator which shall be used for the approximation. Assuming that all elements are similar to one another (and thus there will be a unique reference element) we can find estimates for these projection operators on each element. From this we can find error estimates for the global approximation. We apply this to analyse a second-order problem in detail.

Definition 3.5.9
Let Ω_e be a bounded polygon then define

$$\begin{aligned} h_e &= \text{diam}(\Omega_e), \\ \rho_e &= \sup\{\text{diam(S)};\ \text{S a ball contained in } \Omega_e\}, \\ \text{meas}(\Omega_e) &= \textstyle\int_{\Omega_e} dx. \end{aligned}$$

Definition 3.5.10 Two open subsets Ω and $\hat{\Omega}$ of $\mathbb{R}^n$ are *affine-equivalent* if there is an invertible affine mapping

$$F : \hat{x} \in \mathbb{R}^n \rightarrow F(\hat{x}) = B\hat{x} + b \in \mathbb{R}^n$$

such that $\Omega = F(\hat{\Omega})$.

If we denote the outer and the inner diameter by respectively h and ρ, then we find

$$\|B\| \leq h_{\Omega_e}/\rho_{\hat{\Omega}} \quad \text{and} \quad \|B^{-1}\| \leq h_{\hat{\Omega}}/\rho_{\Omega_e}\,.$$

Usually we have the following correspondences between points and functions

$$\hat{x} \in \hat{\Omega} \rightarrow x = F(\hat{x}) \in \Omega, \tag{3.64}$$

$$(\hat{v} : \hat{\Omega} \rightarrow \mathbb{R}) \rightarrow (v = \hat{v} \circ F^{-1} : \Omega \rightarrow \mathbb{R}). \tag{3.65}$$

So we have

$$\hat{v}(\hat{x}) = v(x).$$

Analogously we define the affine-equivalence of two finite elements and all finite elements in a partitioning are affine equivalent to a master element $(\hat{\Omega}, \hat{P}, \hat{L})$ if for each $\Omega_e \in \Delta_h$ there exists an affine mapping F_e such that for this element

- $x = F_e(\hat{x})$,
- $\psi_j^e(x) = \hat{\psi}_j(\hat{x})$,
- $l_i^e(\psi_j^e(x)) = \hat{l}_i(\hat{\psi}_j(\hat{x}))$.

Before we give the basic error theorems we recall the definitions of the norms and semi-norms with which we work.

$$|v|_{m,p,\Omega_e} := \left(\sum_{|\alpha|=m} \int_{\Omega_e} |D^\alpha v|^p \; dx \right)^{\frac{1}{p}},$$

$$\|v\|_{m,p,\Omega_e} := \left(\sum_{l \leq m} |v|^p_{l,p,\Omega_e} \right)^{\frac{1}{p}} = \left(\sum_{|\alpha| \leq m} \int_{\Omega_e} |D^\alpha v|^p \; dx \right)^{\frac{1}{p}},$$

$$|v|_{m,\infty,\Omega_e} := \max_{|\alpha|=m} \{\text{ess.sup}_x |D^\alpha v(x)|\},$$

$$\|v\|_{m,\infty,\Omega_e} := \max_{|\alpha|\leq m} \{\text{ess.sup}_x |D^\alpha v(x)|\}.$$

Lemma 3.5.11 Assume the integers $k \geq 0$ and $m \geq 0$ and the numbers $p, q \in [1, \infty]$ are such that the Sobolev spaces $W^{k+1,p}(\hat{\Omega})$ and $W^{m,q}(\hat{\Omega})$ satisfy the inclusion

$$W^{k+1,p}(\hat{\Omega}) \hookrightarrow W^{m,q}(\hat{\Omega}).$$

Further let $\hat{\Pi} \in \mathcal{L}(W^{k+1,p}(\hat{\Omega}); W^{m,q}(\hat{\Omega}))$ be a polynomial preserving mapping:

$$\forall \hat{p} \in P_k(\hat{\Omega}),\ \hat{\Pi}\hat{p} = \hat{p}.$$

For every open set Ω_e that is affine-equivalent to $\hat{\Omega}$ define the mapping Π_e by

$$(\Pi_e v)\hat{} = \hat{\Pi}\hat{v},$$

with $\hat{v} \in W^{k+1,p}(\hat{\Omega})$ and $v \in W^{k+1,p}(\Omega_e)$ corresponding as in (3.65). Then there exists a constant $C = C(\hat{\Pi}, \hat{\Omega})$ such that for all affine-equivalent sets Ω_e and all $v \in W^{k+1,p}(\Omega_e)$

$$|v - \Pi_e v|_{m,q,\Omega_e} \leq C(\text{meas}(\Omega_e))^{1/q-1/p} \ \frac{h_e^{k+1}}{\rho_e^m} |v|_{k+1,p,\Omega_e},$$

with h_e, ρ_e and $\text{meas}(\Omega_e)$ as in definition **3.5.9**.

We specialise this immediately to affine-equivalent finite element families. We formulate a number of requirements on the spaces we use:

1. we have a set of degrees of freedom $\hat{L}$ which uses derivatives up to certain order s. By Sobolev's Lemma, this sets requirements on the Sobolev spaces. This gives condition (3.66),
2. the projections will be measured in a weaker norm $\|\cdot\|_{m,q,\Omega_e}$: condition (3.67),
3. the approximating functions will at least have to include all polynomials up to certain order: condition (3.68).

Lemma 3.5.12 Let $(\hat{\Omega}, \hat{P}, \hat{L})$ be a master element for which s is the largest derivative in $\hat{L}$. If for some $k \geq 0$, $m \geq 0$ and $p, q \in [1, \infty]$

$$W^{k+1,p}(\hat{\Omega}) \hookrightarrow \mathcal{C}^s(\hat{\Omega}), \tag{3.66}$$
$$W^{k+1,p}(\hat{\Omega}) \hookrightarrow W^{m,q}(\hat{\Omega}), \tag{3.67}$$
$$P_k(\hat{\Omega}) \subset \hat{P} \subset W^{m,q}(\hat{\Omega}), \tag{3.68}$$

then there is a $C = C(\hat{\Omega}, \hat{P}, \hat{L})$ such that for affine-equivalent finite elements (Ω_e, P_e, L_e) and all $v \in W^{k+1,p}(\Omega_e)$

$$|v - \Pi_e v|_{m,q,\Omega_e} \leq C(\text{meas}(\Omega_e))^{1/q-1/p} \ \frac{h_e^{k+1}}{\rho_e^m} |v|_{k+1,p,\Omega_e}.$$

Obviously, the term $\frac{h_e^{k+1}}{\rho_e^m}$ appearing in this lemma is undesirable, so we would like to dispose of the ρ_e. This can be done provided the elements do not become "flat" in the limit: for $h_e \to 0$ we want that $\rho_e = \mathcal{O}(h_e)$.

Definition 3.5.13 A sequence of partitionings $\{\Delta_h\}$ of finite elements $\{(\Omega_e, P_e, L_e)\}_h$ [10] is *regular* if the following two conditions are satisfied:

1. There exists a constant σ such that

$$\forall e, \ \frac{h_e}{\rho_e} \leq \sigma.$$

[10] Note that we use e as a kind of parameter of the family.

2. The diameter $h = \max_{\Omega_e \in \Delta_h} h_e$ tends to zero.

Remark:
The regularity of the elements can also be stated in a more geometrical sense: let θ_Ω denote the smallest angle in an element Ω, then there must be a $\theta_0 > 0$ such that

$$\forall \Omega \in \mathcal{F}_h,\ \theta_\Omega \geq \theta_0 > 0.$$

For regular families we can convert the error estimate **3.5.12** in estimates for the associated *norms*.

Lemma 3.5.14 Let there be given a regular affine family of finite elements (Ω_e, P_e, L_e) whose reference element $(\hat{\Omega}, \hat{P}, \hat{L})$ satisfies conditions (3.66), (3.67) and (3.68). Then there exists a constant $C = C(\hat{\Omega}, \hat{P}, \hat{L})$ such that for all members Ω of the family and all $v \in W^{k+1,p}(\Omega)$,

$$\|v - \Pi_e v\|_{m,q,\Omega_e} \leq C(\text{meas}(\Omega_e))^{1/q-1/p}\ h_e^{k+1-m} |v|_{k+1,p,\Omega_e}.$$

Thus, we can give an error estimate on each element in a partition Δ_h of our domain Ω_e. Assuming the family is affine-equivalent we would like to give an error estimate for the *global* finite element approximation. The assumptions above have to be applied in an even more strict sense: on each level of refinement.

Assumptions:
(H1) We consider a *regular family* of finite elements $\mathcal{F}_h$. That is, there exists a constant $\sigma > 0$ such that

$$\forall\, \Omega_e \in \cup_h \mathcal{F}_h,\ \frac{h_e}{\rho_e} \leq \sigma.$$

and

$$h = \max_{\Omega_e \in \mathcal{F}_h} h_e \to 0\,.$$

(H2) All the finite elements $(\Omega_e, P_e, L_e),\ \Omega_e \in \cup_h \mathcal{F}_h$ are affine-equivalent to a single reference element $(\hat{\Omega}, \hat{P}, \hat{L})$.

(H3) All the elements are of class $\mathcal{C}^0$.

To summarise the result we define a norm and a semi-norm for *piecewise smooth functions* on a given partitioning Δ_h.

Definition 3.5.15 For functions that are (together with their derivatives up to order m) integrable on the elements $\Omega_e \in \Delta_h$, we introduce the following norm and seminorm if $p < \infty$

$$\|v\|_{m,p,\Delta_h} := \left\{ \sum_{\Omega_e \in \Delta_h} \|v\|^p_{m,p,\Omega_e} \right\}^{\frac{1}{p}},$$

$$|v|_{m,p,\Delta_h} := \left\{ \sum_{\Omega_e \in \Delta_h} |v|^p_{m,p,\Omega_e} \right\}^{\frac{1}{p}}.$$

In case $p = 2$ we drop the p in the subscripts. We denote the space of functions with finite norm $\|.\|_{m,\Delta_h}$ by $H^m_{\Delta_h}$.

If $p = \infty$ we define

$$\|v\|_{m,\infty,\Delta_h} := \max_{\Omega_e \in \Delta_h} \|v\|_{m,\infty,\Omega_e},$$

$$|v|_{m,\infty,\Delta_h} := \max_{\Omega_e \in \Delta_h} |v|_{m,\infty,\Omega_e}.$$

Example 3.5.16
The norm $\|\cdot\|_{m,p,\Delta_h}$ on the partitioning Δ_h of Ω is not the same as the norm $\|\cdot\|_{m,p,\Omega}$ on Ω itself. This is seen as follows. Take $\Omega = [a, b]$ and consider the piecewise linear function in figure 3.11. This function is an element of $H^1(\Omega)$,

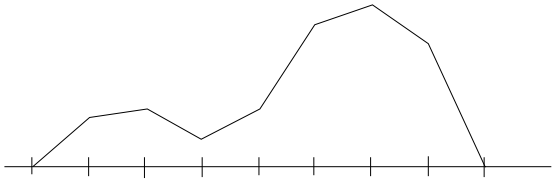


Figure 3.11: Piecewise H^2-function.

but it is not in $H^2(\Omega)$. If we take however a partition Δ_h such that the elements Ω_e correspond with the smooth parts of the function, it is piecewise H^2, so in $H^2_{\Delta_h}(\Omega)$.

A number of approximation results is summarised in the following lemma.

Lemma 3.5.17 Let $\mathcal{F}_h$ be a regular family of finite elements satisfying (H1), (H2) and (H3) that are all affine equivalent with the master element $(\hat{\Omega}, \hat{L}, \hat{P})$. Let the requirements (3.66), (3.67) and (3.68) be satisfied for every element. Consider the global interpolation operator

$$\Pi_h : W^{k+1,p}(\Omega) \to S_h \subset W^{m,p}(\Omega)$$

defined by

$$\Pi_h|_{\Omega_e} = \Pi_e$$

then there is a constant $C = C(\hat{\Omega}, \hat{L}, \hat{P}) > 0$, such that,

$$|v - \Pi_e v|_{m,p,\Omega_e} \leq C\ h_e^{k+1-m}\ |v|_{k+1,p,\Omega_e},$$

and

$$|v - \Pi_h v|_{m,p,\Delta_h} \leq C\ h^{k+1-m}\ |v|_{k+1,p,\Delta_h}.$$

From this easily follows

$$\|v - \Pi_e v\|_{m,p,\Omega_e} \leq C\ h_e^{k+1-m}\ |v|_{k+1,p,\Omega_e}, \tag{3.69}$$

and

$$\|v - \Pi_h v\|_{m,p,\Omega} \leq C\ h^{k+1-m}\ |v|_{k+1,p,\Omega}. \tag{3.70}$$

Remark:
If the functions v and $\Pi_h v$ are not sufficiently smooth estimate (3.70) is not valid, but we still have

$$\|v - \Pi_h v\|_{m,p,\Delta_h} \leq C\ h^{k+1-m}\ |v|_{k+1,p,\Delta_h}.$$

As a direct consequence of this lemma we can state the basic interpolation theorem that can be applied to elliptic problems. We thus take $p = q = 2$.

Theorem 3.5.18 Under conditions as in lemma **3.5.17**, for a regular family of finite element approximations, the following error estimate is valid

$$\|u - \Pi_h u\|_{m,\Omega} \leq C\ h^{k+1-m}\ |u|_{k+1,\Delta_h}.$$

Here C is independent of h, $0 \leq m \leq k+1$, $u \in H^{k+1}(\Omega)$ and Π_h is the interpolation operator that leaves invariant all polynomials of degree k or less.■

Definition 3.5.19 The spaces S_h, based on a regular family $\mathcal{F}_h$ of finite elements, such that the requirements of lemma **3.5.12** are satisfied for $p = 2$, i.e.they are k-th degree piecewise polynomials in $H^m(\Omega)$, are denoted by $S_h^{k,m}(\Omega)$, $0 \leq m \leq k+1$.

The spaces $S_h^{k,m}(\Omega)$ contain all k-th order piecewise polynomials and are subsets of $H^m(\Omega)$. By the preceding theorem we know that for a sequence $\{S_h^{k,m}(\Omega)\}_{h\to 0}$, $h \leq h_0$, $u \in H^r(\Omega)$, $r \geq 0$ and $0 \leq s \leq \min(r,m)$ there exists a $u_h \in S_h^{k,m}(\Omega)$ and a constant $C \geq 0$, independent of h, such that

$$\|u - u_h\|_s \leq C\ h^{\sigma}\ \|u\|_r,$$

where $\sigma = \min(k+1-s, r-s)$.

Corollary 3.5.20
Under the same conditions as in theorem **3.5.18**, let u be the solution to a second order elliptic problem and let u_h be the associated discrete solution. Then

$$\|u - u_h\|_{1,\Omega} = \mathcal{O}(h^k).$$

Sufficient for convergence for a second order problem is thus $k = 1$, or approximation by piecewise linear functions.
Proof: Because we consider a second order problem we have $m = 1$ and the theorem follows by combining theorem **3.5.1** and theorem **3.5.18**. ■

Inverse inequalities

We now treat a property of piecewise polynomial spaces that has no immediate use at this point. However it is an easy consequence of the foregoing and we will use it in section **3.5.6**.

Definition 3.5.21 (H4) A family $\mathcal{T}_h$ of triangulations satisfies an *inverse assumption* if there exists a ν such that

$$\forall\ \mathrm{T} \in \cup_h \mathcal{T}_h,\ \frac{h}{h_{\mathrm{T}}} \leq \nu.$$

In fact an inverse assumption implies that the elements can not differ too much in size, which in turn implies that functions in the approximation space have derivatives that will not "blow up", unless the function itself "blows up".

It is clear that this assumption is by no means restrictive in practice: the 'usual' finite element spaces satisfy an inverse assumption.

For such families of spaces we establish the following equivalence between semi-norms. Note that σ appears in the regularity assumption (H1).

Theorem 3.5.22 Let there be given a regular family of triangulations $\mathcal{T}_h$ in $\mathbb{R}^d$, satisfying an inverse assumption. Let $0 \leq m \leq k$, $p, q \in [1, \infty]$, and

$$\hat{P} \subset W^{k,p}(\hat{\Omega}) \cap W^{m,q}(\hat{\Omega}),$$

then $\exists C = C(\nu, \sigma, k, m, p, q)$ such that for all $v_h \in V_h$

$$|v_h|_{k,p,\Delta_h} \leq \frac{C}{\mathrm{meas}(\Omega_h)^{\max\{0,1/q-1/p\}} h^{k-m}} |v_h|_{m,q,\Delta_h}, \tag{3.71}$$

with the usual adaptation if p or q is ∞.
Proof: See [4, p. 141]. ∎

Example 3.5.23
Assume $V_h \in W^{l,r}$ for some (l, r), then the above inequalities can be given for the seminorms $|.|_{l,r,\Omega}$. For instance suppose assumption (H3) is valid and that $\hat{P} \subset H^1(\Omega)$, then

$$|v_h|_{0,\infty,\Omega} \leq \frac{C}{h^{n/2}} |v_h|_{0,\Omega},$$

and

$$|v_h|_{1,\Omega} \leq \frac{C}{h} |v_h|_{0,\Omega}. \tag{3.72}$$

If $\hat{P} \subset W^{1,\infty}(\Omega)$, then

$$|v_h|_{1,\infty,\Omega} \leq \frac{C}{h} |v_h|_{0,\infty,\Omega}.$$

Example 3.5.24
We can also easily give inequalities for the norms: for instance, from (3.72) we find

$$\|v_h\|_{1,\Omega} \leq \frac{C}{h}|v_h|_{0,\Omega}. \tag{3.73}$$

This can be applied for the following case. Take $\Omega = [a, b]$. Consider $S_h^{1,1}(\Omega)$, the space of piecewise linear functions on a regular partition Δ_h, then (3.73) is valid for each $v_h \in S_h^{1,1}(\Omega)$.

Motivated by this we come to the following definition, suited for the function spaces $S_h^{k,m}(\Omega)$.

Definition 3.5.25 A family of piecewise polynomial spaces $\{S_h^{k,m}(\Omega)\}_{h\to 0}$ has the *inverse property* if

$$\exists C: \ \forall s \leq m, \ h^m \ \|u_h\|_m \leq C \ h^s \ \|u_h\|_s, \ \forall u_h \in S_h^{k,m}(\Omega).$$

Error estimate in the L^2-norm

In Section **3.5.4** we saw that under conditions we can assure that $\|u-u_h\|_{1,\Omega} = \mathcal{O}(h^k)$. This means that at least $\|u - u_h\|_{0,\Omega} = \mathcal{O}(h^k)$. In this section we show that, under mild additional assumptions, $\|u - u_h\|_{0,\Omega} = \mathcal{O}(h^{k+1})$.

Theorem 3.5.26 (Aubin-Nitsche Lemma)
Let $Lu = f$ be a problem of order $2m$ and let $V_h \subset V = H^m(\Omega)$. The continuous problem is to find u such that $B(u, v) = f(v)$ for all $v \in V$. The discrete problem is: find u_h such that $B(u_h, v_h) = f(v_h)$ for all $v_h \in V_h$. Assume that for all $g \in H^{-s}$ there is a solution to $L^T z = g$ (the *adjoint problem*), such that

$$\|z\|_{2m-s} \leq C \ \|g\|_{-s} \ , \quad s \in [0, m] \, ,$$

(the *regularity* of the adjoint problem). Then

$$\|u - u_h\|_{s,\Omega} \leq C \ h^{m-s} \ \|u - u_h\|_{m,\Omega}. \tag{3.74}$$

Proof: Define the error $e = u_h - u$. For any $0 \leq s \leq m$:

$$\|e\|_s = \sup_{g\in H^{-s}} \frac{|g(e)|}{\|g\|_{-s}} \, ,$$

We first evaluate $|g(e)|$. Let z be the solution to the adjoint problem $B(w, z) = g(w)$, for all $w \in H^s(\Omega)$. In particular $B(e, z) = g(e)$, or, as $B(e, v_h) = 0$ for all $v_h \in V_h$,

$$B(e, z - v_h) = g(e).$$

So

$$|g(e)| \quad \leq \quad \|B\| \ \|e\|_m \ \inf_{v_h\in V_h} \|z - v_h\|_m.$$

Now, provided $r \leq k+1-m$,

$$|g(e)| \leq \|B\| \ \|e\|_m \ C \ h^r \|z\|_{m+r},$$

due to theorem **3.5.18**. If we put $r = m - s$ and provided $2m - s \leq k+1$,

$$\begin{aligned} |g(e)| &\leq \|B\| \ \|e\|_m \ C \ h^{m-s} \ \|z\|_{2m-s} \\ &\leq \|B\| \ \|e\|_m \ C \ h^{m-s} \ \|g\|_{-s}, \end{aligned}$$

because of the regularity of the adjoint problem. So, if we compute the error in the s-norm, we get

$$\|e\|_s = \sup_{g \in H^{-s}} \frac{|g(e)|}{\|g\|_{-s}} \leq C \ h^{m-s} \ \|e\|_m.$$

■

Corollary 3.5.27
For the error in L^2-norm for a second order problem we get

$$\|u - u_h\|_{0,\Omega} \leq C \ h^{k+1} |u|_{k+1,\Delta_h},$$

provided $k \geq 1$.

Remarks:

- The regularity of the adjoint problem is not very restrictive. For example in case of a symmetric problem it is a trivial matter.

3.5.5 Pointwise error estimate and superconvergence

In this section we restrict ourselves to a second order linear ordinary differential equation on $\Omega = [a, b] \subset \mathbb{R}$

$$Ly \equiv -\frac{d}{dx}(a_2(x)\frac{d}{dx}y) + a_1(x)\frac{d}{dx}y + a_0(x)y = s(x), \tag{3.75}$$

with $a_2(x) \neq 0$ and homogeneous Dirichlet boundary conditions

$$y(a) = y(b) = 0.$$

We will use the *Greens function* for equation (3.75): i.e. the function $G(x;\xi)$ such that [11]

$$y(x) = -\int_a^b G(x;\xi) \ s(\xi) \ d\xi, \tag{3.76}$$

[11] The function G is the resolvent kernel of the differential equation.

for any function $s(x)$. A priori, it is not clear that such a function $G(x;\xi)$ can exist, but we will construct it. Therefore, let φ_1 and φ_2 be two independent solutions to the adjoint (transposed) equation

$$L^T y \equiv -\frac{d}{dx}(a_2(x)\frac{d}{dx}y) - \frac{d}{dx}(a_1(x)y) + a_0(x)y = 0,$$

with boundary conditions $\varphi_1(a) = 0$, $\varphi_1'(a) = 1$, $\varphi_2(b) = 0$ and $\varphi_2'(b) = 1$.

Now $G(x;\xi)$ is constructed as

$$G(x;\xi) = \begin{cases} \varphi_1(x)\varphi_2(\xi)/z(x) & \text{if } x < \xi \\ \varphi_1(\xi)\varphi_2(x)/z(x) & \text{if } \xi < x, \end{cases} \tag{3.77}$$

where[12]

$$z(x) := a_2(x)\left(\varphi_1(x)\varphi_2'(x) - \varphi_1'(x)\varphi_2(x)\right).$$

Remark:
It is easily seen that this function $z(x)$ satisfies the equation $a_2 z' + a_1 z = 0$. So z has a unique sign. If $z \equiv 0$ then φ_1 and φ_2 are linearly dependent. (In that particular case the homogeneous problem has a non-trivial solution and $G(x,\xi)$ does not exist.)

By substitution of (3.76) in (3.75), it is not hard to verify that $G(x;\xi)$ is the Green's function indeed. Other properties of this Greens function $G(x;\xi)$ are:

1. $G(x;\cdot) \in H_0^1(\Omega) \cap C^2(\ (a,x)\ \cup\ (x,b)\ )$,
2. $L^T G(x;\cdot) \equiv 0$ on $(a,x) \cup (x,b)$,
3. $$\text{jump}_{\xi=x}\frac{\partial}{\partial\xi}G(x;\xi) := \lim_{h\to 0}\frac{\partial G(x;\xi+h)}{\partial\xi} - \lim_{h\to 0}\frac{\partial G(x;\xi-h)}{\partial\xi} = \frac{1}{a_2(x)}.$$

If x_i is a nodal point in a partition Δ_h of Ω, then $G(x_i,\cdot)$ can be approximated well in $S_h^{k,1}(\Omega)$. Under this condition we see that

$$\inf_{v_h\in S_h^{k,1}} \|G(x_i;\cdot) - v_h\|_{m,\Delta_h} \le C\ h^{k+1-m}\ \|G(x_i;\cdot)\|_{k+1,\Delta_h}.$$

So, for $m = 1$

$$\inf_{v_h\in S_h^{k,1}(\Omega)} \|G(x;\cdot) - v_h\|_1 \le \begin{cases} C\ h^k\ |G|_{k+1,\Delta_h} & \text{if} \quad x = x_i, \\ C\ h^0\ |G|_1 & \text{if} \quad x \ne x_i. \end{cases} \tag{3.78}$$

The last estimate is true because in this case $G(x;\cdot)$ is not piecewise H^{k+1} on Δ_h, but it still is continuous, so in H^1.

[12]The determinant $\varphi_1\varphi_2' - \varphi_1'\varphi_2 = \begin{vmatrix} \varphi_1 & \varphi_1' \\ \varphi_2 & \varphi_2' \end{vmatrix}$ is called the *Wronskian*.

Theorem 3.5.28 If equation (3.75) is discretised by a finite element method with $S_h = V_h = S_h^{k,1}(\Omega)$, then we have the pointwise error estimate

$$|y(x) - y_h(x)| \leq \begin{cases} C\ h^{2k} & \text{if} \quad x = x_i \in \Delta_h, \\ C\ h^k & \text{if} \quad x \notin \Delta_h. \end{cases} \tag{3.79}$$

Proof: $y \in H_0^1(\Omega)$ is the solution to $B(y,v) = f(v)$, $\forall v \in H_0^1(\Omega)$. Similarly $y_h \in S_h^{k,1}(\Omega)$ is the solution to $B(y_h, v_h) = f(v_h)$, $\forall v_h \in S_h^{k,0}(\Omega)$, so for $e := y - y_h$ we have $B(e, v_h) = 0\ \forall v_h \in S_h^{k,1}(\Omega)$. Now

$$y(x) = -\int_a^b G(x;\xi)\ f(\xi)\ d\xi = -\int G(x;\xi) L y(\xi) d\xi = -B(y, G(x;\cdot)),$$

and analogously for y_h, so

$$\begin{aligned} e(x) &= -B(e, G(x;\cdot)) \\ &= -B(e, G(x;\cdot) - v_h),\ \forall v_h \in S_h^{k,1}(\Omega). \end{aligned}$$

This gives

$$|e(x)| \leq \|B\|\ \|e\|_1\ \inf_{v_h} \|G(x;\cdot) - v_h\|_1.$$

Because B is bounded, $\|e\|_1 = \mathcal{O}(h^k)$ and (3.78) the theorem follows. ∎

The phenomenon that the order of accuracy of an approximation at particular points is better than the estimate in a global norm, is called *superconvergence*.

Remark:
If this same problem is tackled with functions in $S^{3,2}(\Omega)$, there will be no superconvergence. Then estimate (3.78) fails to be better at the points of Δ_h.

3.5.6 The influence of the quadrature rule

In the previous sections we have given estimates for the error $\|y - y_h\|$, that we obtain when we solve the discrete equation instead of the continuous equation. As we have seen in section **3.4.3** it may be so that we have to compute the integrals in the discrete equations using a quadrature rule. So we not only have the error $\|y - y_h\|$ but also an additional error $\|y_h - \tilde{y_h}\|$ (which can be associated with the consistency error of theorem **3.5.3**). This could disturb all the error estimates we have given so far. In order to take this into account we require that the additional error should be at most of the same order as the discretisation error. This will lead to some requirement for accuracy of the quadrature rule.

In this section we will formulate these requirements for two cases: first we show under which conditions the *global error estimate* is preserved (as in theorem **3.5.18**) and further we discuss the how we can preserve the *pointwise superconvergence* of theorem **3.5.28**.

Definition 3.5.29 A quadrature rule of the form

$$\int_\Omega f\ dx \approx \sum_i w_i\ f(x_i)$$

is called to be of *degree* t if it integrates any polynomial f of degree t exactly.

Theorem 3.5.30 Given a repeated quadrature rule of degree t, applied to a partition Δ_h, then there exists, for $f \in \mathcal{C}(\Omega) \cap W^{t+1,p}(\Omega_e)$, $\forall \Omega_e$, a piecewise polynomial $\Pi_h f$ of degree t such that it interpolates f on the nodes, i.e.

$$f(x_i) = \Pi_h f(x_i)$$

and such that

$$\|f - \Pi_h f\|_{m,p,\Delta_h} \leq C\ h^{t+1-m}\ |f|_{t+1,p,\Delta_h}.$$

Proof: Take a $(t+1)$-points interpolating polynomial of degree t on every Ω_e: each of this gives the estimate of lemma **3.5.17**. Provided the set of nodes contains the quadrature points the estimate is valid. ■

Corollary 3.5.31
When a quadrature rule of degree t is used the quadrature error is $\mathcal{O}(h^{t+1})$:

$$\int_{\Omega_e} |f - \Pi_h f|\ dx \leq C\ h^{t+1} \sum_{|\alpha|=t+1} \int_{\Omega_e} |D^\alpha f(x)|\ dx. \tag{3.80}$$

Example 3.5.32
A $t+1$-point Newton-Coates formula has degree t. An n-points Gauss quadrature has degree $2n-1$; an n-point Lobatto formula has degree $2n-3$.

In the following we will restrict ourselves to second order elliptic problems, and we use a quadrature of degree t, we will assume that all the coefficients in the partial differential equation are (piecewise) $C^{t+1}(\Omega)$-functions.

In order to study the influence of the quadrature on the accuracy of the FEM we have to distinguish three different problems:

1. **the continuous problem**
   find $y \in S$ such that $B(y, v) = f(v)$, for all $v \in V$;

2. **the discrete problem**
   find $y_h \in S_h$ such that $B(y_h, v_h) = f(v_h)$ for all $v_h \in V_h$;

3. **the discrete problem with quadrature**
   find $\tilde{y}_h \in S_h$ such that $B_h(\tilde{y}_h, v_h) = f_h(v_h)$, for all $v_h \in V_h$.

Here f_h is a linear and B_h is a bilinear functional in which the integrals (see Section **3.4.3**) are replaced by a repeated quadrature rule of degree t on the partition Δ_h.

To compute the additional error we first have to make an estimate of the amount by which f_h deviates from f, and B_h deviates from B.

Lemma 3.5.33 When using a quadrature rule of degree t we have

$$|f(v_h) - f_h(v_h)| \leq C\ h^{t+1} \|v_h\|_{k,\Delta_h} \tag{3.81}$$

and

$$|B(\tilde{y}_h, v_h) - B_h(\tilde{y}_h, v_h)| \leq C\ h^{t+1}\ \|\tilde{y}_h\|_{k,\Delta_h}\ \|v_h\|_{k,\Delta_h}. \tag{3.82}$$

Proof: We only give the proof for f_h and f, the other goes analogously. Using the operator Π_h, as introduced in Theorem 3.5.30, we have

$$E := |f(v_h) - f_h(v_h)| = \sum_e |\int_{\Omega_e} ((fv) - \Pi_h(fv))\ dx|.$$

Using (3.80) we have

$$E \leq C\ h^{t+1}\ \sum_e \int_{\Omega_e} |D^{t+1}(fv)|\ dx,$$

so, by Leibniz' rule and the fact that $v_h \in V_h$,

$$E \leq C\ h^{t+1} \sum_{l=0}^{k} \sum_e \int_{\Omega_e} |(D^{t+1-l}f)(D^l v_h)|\ dx \leq C\ h^{t+1}\ \|v_h\|_{k,\Delta_h}.$$

■

The following theorem states the conditions under which the accuracy as given in theorem **3.5.18** is preserved.

Theorem 3.5.34 Let $S = V = H^1(\Omega)$, Δ_h a regular partition of Ω and assume S_h and V_h satisfy the inverse property. Further, let $B : S_h \times V_h \to \mathbb{R}$ be bounded and coercive and let f_h and B_h be the approximations of f and B obtained by quadrature $t \geq 2k - 2$. Then

$$\|y - \tilde{y}_h\|_{1,\Omega} = \mathcal{O}(h^k).$$

Proof: Let y_h be the solution of the discrete problem and $\tilde{y}_h$ the solution to the discrete equation with quadrature. We seek a upper bound of $\|y_h - \tilde{y}_h\|$ by considering

$$\|y_h - \tilde{y}_h\|_1 \leq \frac{1}{\gamma_h} \sup_{v_h \in V_h} \frac{|B(y_h - \tilde{y}_h, v_h)|}{\|v_h\|_1},$$

in which γ_h is the (discrete) coercivity constant of B.

So compute

$$\begin{aligned}
|B(y_h - \tilde{y}_h, v_h)| &\leq |B(y_h, v_h) - B_h(\tilde{y}_h, v_h)| + |B_h(\tilde{y}_h, v_h) - B(\tilde{y}_h, v_h)| \\
&\overset{i}{=} |f(v_h) - f_h(v_h)| + |B_h(\tilde{y}_h, v_h) - B(\tilde{y}_h, v_h)| \\
&\overset{ii}{\leq} C\ h^{t+1}\ \|v_h\|_{k,\Delta_h} + C\ h^{t+1}\ \|\tilde{y}_h\|_{k,\Delta_h}\ \|v_h\|_{k,\Delta_h} \\
&\overset{iii}{\leq} C\ h^{t+2-k}\ \|v_h\|_1 + \\
&\quad C\ h^{t+2-k}\ \|v_h\|_1\ \{\|y_h\|_{k,\Delta_h} + \|y_h - \tilde{y}_h\|_{k,\Delta_h}\} \\
&\overset{iiii}{\leq} C\ h^{t+2-k}\ \|v_h\|_1\ \{1 + \|y_h\|_{k,\Delta_h}\} + \\
&\quad C\ h^{t+3-2k}\ \|v_h\|_1\ \|\tilde{y}_h - y_h\|_{1,\Delta_h}.
\end{aligned}$$

We used (i) that y_h is a solution of the discrete problem and $\tilde{y}_h$ is the solution to the discrete problem with quadrature, (ii) the estimates of Lemma **3.5.33** for the quadrature, (iii) the fact that V_h has the inverse property and $(iiii)$ that S_h has the inverse property.

Thus,

$$\|y_h - \tilde{y}_h\|_1\{1 - C\ h^{t+3-2k}\} \leq C\ h^{t+2-k}\ \{1 + \|y_h\|_{k,\Delta_h}\}.$$

Provided h is small enough ($C\ h^{t+3-2k} < 1$), we have

$$\|y_h - \tilde{y}_h\|_1 \leq C\ h^{t+2-k}\ \{1 + \|y_h\|_{k,\Delta_h}\}. \tag{3.83}$$

To preserve the accuracy t should satisfy both $t + 2 - k \geq k$ and $t + 3 - 2k > 0$, both of which are satisfied if $t \geq 2k - 2$. ∎

For the superconvergence as treated in Section **3.5.5** we now answer the same question: what should be the degree of the quadrature to preserve the pointwise superconvergence accuracy?

Theorem 3.5.35 Let the two point boundary value problem (3.75) be discretised by a finite element method with $S_h = V_h = S_h^{k,1}(\Omega)$, and let the quadrature rule be of degree t, with $t \geq 2k - 1$, then

$$|y(x) - \tilde{y}_h(x)| = \begin{cases} \mathcal{O}(h^{2k}), & x \in \Delta_h, \\ \mathcal{O}(h^k), & x \notin \Delta_h. \end{cases}$$

Proof: We introduce $G_i = G(x_i, \cdot)$, x_i a nodal point, and consider

$$|y_h(x_i) - \tilde{y}_h(x_i)| = |B(y_h - \tilde{y}_h, G_i)|.$$

First

$$B(y_h - \tilde{y}_h, G_i) \quad = \quad B(y_h - \tilde{y}_h, G_i - v_h) + B(y_h - \tilde{y}_h, v_h)$$

$$
\begin{aligned}
&= B(y_h - \tilde{y}_h, G_i - v_h) + B(y_h, v_h) - B_h(\tilde{y}_h, v_h) \\
&\quad + B_h(\tilde{y}_h, v_h) - B(\tilde{y}_h, v_h) \\
&= B(y_h - \tilde{y}_h, G_i - v_h) + f(v_h) - f_h(v_h) \\
&\quad + B_h(\tilde{y}_h, v_h) - B(\tilde{y}_h, v_h),
\end{aligned}
$$

so, using the continuity of B

$$
\begin{aligned}
|B(y_h - \tilde{y}_h, G_i)| &\leq M\ \|y_h - \tilde{y}_h\|_1\ \|G_i - v_h\|_1 + |f(v_h) - f_h(v_h)| \\
&\quad + |B_h(\tilde{y}_h, v_h) - B(\tilde{y}_h, v_h)| \\
&\leq M\ \|y_h - \tilde{y}_h\|_1\ \|G_i - v_h\|_1 + C\ h^{t+1}\ \|v_h\|_{k,\Delta_h} \\
&\quad + C\ h^{t+1}\ \|\tilde{y}_h\|_{k,\Delta_h}\ \|v_h\|_{k,\Delta_h} \\
&\overset{(3.83)}{\leq} M\ C\ h^{t+2-k}\ \{1 + \|y_h\|_{k,\Delta_h}\}\ \|G_i - v_h\|_1 \\
&\quad + C\ h^{t+1}\ \|v_h\|_{k,\Delta_h}\ \{1 + \|y_h\|_{k,\Delta_h}\} \\
&\leq M\ C\ h^{t+2-k}\ \{1 + \|y_h\|_{k,\Delta_h}\}\ \|G_i - v_h\|_1 \\
&\quad + C\ h^{t+1}\ \|G_i - v_h\|_{k,\Delta_h}\ \{1 + \|y_h\|_{k,\Delta_h}\} \\
&\quad + C\ h^{t+1}\ \|G_i\|_{k,\Delta_h}\ \{1 + \|y_h\|_{k,\Delta_h}\} \\
&\overset{(3.78)}{\leq} C\ h^{t+1}\ \|G_i\|_{k,\Delta_h} + C\ h^{t+1}\ \|G_i\|_{k,\Delta_h} \\
&\leq C\ h^{t+1}\ \|G_i\|_{k,\Delta_h}.
\end{aligned}
$$

According to theorem **3.5.28** $y(x) - y_h(x) = \mathcal{O}(h^{2k})$ in nodal points $x = x_i \in \Delta_h$, so we conclude that by the condition $t+1 \geq 2k$ also $y(x) - \tilde{y_h}(x) = \mathcal{O}(h^{2k})$ in nodal points. ■

Example 3.5.36
A Gaussian quadrature on k points or a Lobatto quadrature on $k+1$ points are sufficiently accurate to preserve superconvergence.

Acknowledgement

In 1992 H. Eleveld prepared an earlier version of these notes on finite element methods, based on lectures given at the University of Amsterdam. Final corrections 2004.

Bibliography

[1] R.A. Adams. *Sobolev Spaces*, volume vol 65 of *Pure and applied mathematics series*. Academic Press, New York, 1975.

[2] D. Braess. *Finite Elements*. Cambridge University Press, 1997.

[3] S. C. Brenner and L. Ridgeway Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, 1994.

[4] P. G. Ciarlet. *The finite element method for elliptic problems*, volume vol 4 of *Studies in mathematics and its application*. North-Holland, Amsterdam, 1978.

[5] G.E. Forsythe and W.R. Wasow. *Finite-Difference Methods for Partial Differential Equations*. Wiley, New York, 1960.

[6] G. H. Golub and Ch. F. van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, London, 1989. 2nd edition.

[7] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. B. G. Teubner, Stuttgart, 1986.

[8] A. Harten, P. D. Lax, and B. Van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Review*, 25:35–61, 1983.

[9] P.W. Hemker and B. Koren. Defect correction and nonlinear multigrid for the steady Euler equations. Technical Report NM-N8801, CWI, 1988.

[10] P.W. Hemker and B. Koren. Efficient multi-dimensional upwinding for the steady Euler equations. Technical Report NM-R9107, CWI, 1991.

[11] C. Hirsch. *Numerical Computation of Internal and External Flows, Volume 1: Fundamentals of Numerical Discretisation*. J. Wiley, 1988.

[12] C. Hirsch. *Numerical Computation of Internal and External Flows, Volume 2: Computational Methods fir Inviscid and Viscous Flow*. J. Wiley, 1990.

[13] R. Jeltsch. Properties of discrete hyperbolic conservation laws. Technical Report 88-54, Inst. f. Geometrie u. Praktische Mathematik, RWTH, Aken, 1988.

[14] C. Johnson, A. Szepessy, and P. Hansbo. Defect correction and nonlinear multigrid for the steady Euler equations. Technical Report 21, Chalmers Univ. Techn., 1987.

[15] B. Koren. *Multigrid and Defect Correction for the Steady Navier-Stokes equations, Applications to aerodynamics.* Number 74 in CWI Tract series. CWI, Amsterdam, 1991.

[16] P. D. Lax. *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, volume 11 of *Regional Conference Series in Applied Mathematics.* SIAM Publication, Philadelphia, 1973.

[17] R.J. LeVeque. *Numerical Methods for Conservation Laws.* Lectures in Mathematics. Birkhäuser, Basel, 1990.

[18] J. L. Lions (eds.) P. G. Ciarlet. *Handbook of Numerical Analysis, Vol II: Finite Element Methods, part I.*

[19] S. V. Patankar. *Numerical Heat Transfer and Fluid Flow.* Hemisphere McGraw-Hill, 1980.

[20] D. W. Peaceman. *Fundamentals of Numerical Reservoir Simulation.* Elsevier Science Publ., 1977.

[21] R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial Value Problems.* Interscience Publ., 1967.

[22] W. Rudin. *Functional Analysis.* McGraw-Hill, New Delhi, etc., 1973.

[23] J. Smoller. *Shock Waves and Reaction Diffusion Equations*, volume 258 of *Grundlehren der mathematische Wissenschaften.* Springer Verlag, 1983.

[24] S.P. Spekreijse. *Multigrid Solution of the Steady Euler Equations*, volume 46 of *CWI Tracts.* CWI, 1988.

[25] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations.* Wadswordth and Brooks, Pacific Grove, Calif., 1989.

[26] R. Struijs, H. Deconinck, and G. Coussement. Multidimensional upwind schemes for the compressible flow equations. Technical report, Von Karman Institute, 1989.

[27] P. A. Raviart. V. Girault. *Finite Elements Methods for Navier Stokes Equations*, volume vol. 5 of *Springer series in Computational Mathematics.* Springer-Verlag, Berlin, et.c, 1986.

[28] B. Van Leer. On numerical dispersion by upwind differencing. Technical Report 85-51, Delft Univ. Techn., 1985.

[29] B. Van Leer. Flux vector splitting for the 1990's. Technical report, Univ. Michigan, 1990.

[30] K. Yosida. *Functional Analysis.* Springer Verlag, Berlin, etc., 1974.

Index

adjoint problem, 101
affine-equivalent, 94
amplification factor, 34
asymmetric problem, 68
Aubin-Nitsche Lemma, 101

backward Euler method, 33
Banach space, 50
barycentric coordinates, 84
bilinear functional, 54
bilinear operator, 53
block centered, 21
boundary conditions, 46
bounded operator, 53
box method, 28
Bubnov-Galerkin, 28
Burgers' equation, 12, 13

Cauchy Riemann, 22
Cauchy-Schwarz inequality, 50
cell centered, 21
cell vertex, 21
central approximation, 32
characteristic polynomial, 5
characteristics, 7
coercive, 56
collocation method, 27
compact support, 51
compatibility condition, 24
complete, 50
conforming method, 69
conjugate exponents, 54
conjugate-linear operator, 53
conservation law, 6
consistency error, 91
convection equation, 31
convective flux, 2
convergence
    strong, 56
    weak, 56
Cranck-Nicolson method, 33
curved boundary, 86

degree, 105
degrees of freedoms, 93
delta function, 59
dense, 50
diffusion equation, 3, 29, 46
diffusive flux, 2
Dirac delta function, 59
Dirichlet boundary conditions, 46
Discrete Lax-Milgram Theorem, 88
discrete solution, 87
distance, 50
distribution, 59
distributional derivative, 60
divergence, 2
downwind approximation, 32
dual space, 54

elementary load vector, 74
elementary stiffness matrix, 74
elliptic, 5, 6
elliptic equation, 46
energy norm, 69
entropy condition, 15, 16
equivalent differential equation, 37
Euler equations, 6

finite difference approximation, 19
finite element, 93
finite element approximation, 21
finite volume approximation, 21
finite volume method, 28

forward Euler method, 33
function spaces, 51
functionals, 53

Galerkin method, 28
Gauss' theorem, 2
generalised derivative, 60
generalised function, 59
global discretisation error, 36
global finite element approximation, 93
gradient, 2
Greens function, 102

Heaviside function, 60
Hermite interpolate, 81
hierarchical basis, 73
Hilbert space, 50
hyperbolic, 5, 7

IBVP, 79
initial-boundary-value problem, 79
inner product, 49
inner product space, 49
interpolation error, 91
inverse property, 101
inviscid Burgers' equation, 8
isoparametric elements, 86

Lagrange interpolation, 70
Laplace equation, 46
Laplace's equation, 3
Lax-Milgram Theorem, generalised, 57
linear functional, 53
linear operator, 53
linear space, 49
linearisation, 30
Lipschitz continuous, 51
load vector, 74
local discretisation error, 36
lumping, 78

mass matrix, 79
master element, 93
mesh, 19, 70
meshwidth, 70
metric, 50
mixed boundary conditions, 46
modified equation, 37
multi-index, 51
multi-integer, 51

Neumann boundary conditions, 46
nodal points, 92
non-unique solution, 14
nonconforming method, 90
norm, 49
normed linear space, 49
numerical diffusion, 37

Oleinik, 15, 16
order of the differential equation, 1
ordinary differential equation, 1

parabolic, 5
partial differential equation, 2
partitioning, 19
PDE, 1
Petrov-Galerkin, 28
piecewise Lagrange interpolation, 70
piecewise polynomials, 21
Poincaré's Lemma, 64
Poisson equation, 3
positive
    strictly, 56
positive definite, 46
potential equation, 46
principle part, 3, 5, 46
projection Π_h, 94, 98, 105

quadrature rule, 75
quasi-linear, 3

Rankine-Hugoniot, 12
reference element, 93
regular family, 97
regularity, 101

scalar field, 49
scalar product, 49
semi-discretisation, 31, 79
semi-norm, 49
separable, 50

sesquilinear functional, 54
sesquilinear operator, 53
Sobolev space, 61
Sobolev's Embeddings Theorem, 62
spectral method, 21
splines, 83
staggered grid, 22
standard hat functions, 71
static condensation, 78
steady equations, 4
stiffness matrix, 74
Stokes equations, 24
strictly positive, 56
sub-coercive, 56
superconvergence, 104
symmetric operator, 53
symmetric problem, 67

topology, 60
trace, 64
triangle inequality, 49

unisolvent set, 86
upwind approximation, 32

variational method, 26
vector space, 49
vertex centered, 21

wave equation, 3
weak solution, 11, 66
weak topology, 60
weakly convergent, 56
weighted residual method, 27
weighting function, 54
Wronskian, 103