An Introduction to the Finite Element Method (FEM) for Differential Equations

Mohammad Asadzadeh

January 13, 2012

Contents

0 Introduction **5**
- 0.1 Preliminaries . . . 5
- 0.2 Trinities . . . 6

1 Partial Differential Equations **15**
- 1.1 Differential operators, superposition . . . 17
  - 1.1.1 Exercises . . . 20
- 1.2 Some equations of mathematical physics . . . 21
  - 1.2.1 Exercises . . . 30

2 Polynomial Approximation in 1d **33**
- 2.1 Overture . . . 33
- 2.2 Exercises . . . 56

3 Polynomial Interpolation and Numerical Integration in 1d **59**
- 3.1 Preliminaries . . . 59
- 3.2 Lagrange interpolation . . . 67
- 3.3 Numerical integration, Quadrature rule . . . 71
  - 3.3.1 Composite rules for uniform partitions . . . 73
  - 3.3.2 Gauss quadrature rule . . . 77
- 3.4 Exercises . . . 81

4 Linear Systems of Equations **85**
- 4.1 Direct methods . . . 86
- 4.2 Iterative methods . . . 93
- 4.3 Exercises . . . 103

5 Two-point boundary value problems **107**
5.1 A Dirichlet problem . . . 107
5.2 A mixed Boundary Value Problem . . . 112
5.3 The finite element method (FEM) . . . 115
5.4 Error estimates in the energy norm . . . 116
5.5 FEM for convection–diffusion–absorption boundary value problems . . . 123
5.6 Exercises . . . 131

6 Scalar Initial Value Problems **139**
6.1 Solution formula and stability . . . 140
6.2 Galerkin finite element methods for IVP . . . 141
6.2.1 The continuous Galerkin method . . . 142
6.2.2 The discontinuous Galerkin method . . . 144
6.3 A posteriori error estimates . . . 146
6.3.1 A posteriori error estimate for cG(1) . . . 146
6.3.2 A posteriori error estimate for dG(0) . . . 153
6.3.3 Adaptivity for dG(0) . . . 155
6.4 A priori error analysis . . . 156
6.4.1 A priori error estimates for the dG(0) method . . . 156
6.5 The parabolic case ($a(t) \geq 0$) . . . 160
6.5.1 Some examples of error estimates . . . 164
6.6 Exercises . . . 167

7 Initial Boundary Value Problems in 1d **171**
7.1 Heat equation in 1d . . . 171
7.1.1 Stability estimates . . . 173
7.1.2 FEM for the heat equation . . . 179
7.1.3 Error analysis . . . 183
7.1.4 Exercises . . . 190
7.2 The wave equation in 1d . . . 192
7.2.1 Wave equation as a system of hyperbolic PDEs . . . 193
7.2.2 The finite element discretization procedure . . . 194
7.2.3 Exercises . . . 197
7.3 Convection - diffusion problems . . . 199
7.3.1 Finite Element Method . . . 201
7.3.2 The Streamline-diffusion method (SDM) . . . 204
7.3.3 Exercises . . . 206

8 Piecewise polynomials in several dimensions 209
8.1 Introduction 209
8.2 Piecewise linear approximation in 2 D 212
8.2.1 Basis functions for the piecewise linears in 2 D 212
8.2.2 Error estimates for piecewise linear interpolation 216
8.2.3 The L_2 projection 217
8.3 Exercises 218

9 Riesz and Lax-Milgram Theorems 221
9.1 Preliminaries 221
9.2 Riesz and Lax-Milgram Theorems 226
9.3 Exercises 230

10 The Poisson Equation 233
10.1 Stability 233
10.2 Error Estimates for FEM 234
10.3 Exercises 243

11 The heat equation in $\mathbb{R}^N$ 247
11.1 Stability 248
11.2 Exercises 252

12 The wave equation in $\mathbb{R}^N$ 255
12.1 Exercises 256

13 Convection - diffusion problems 259
13.1 A convection-diffusion model problem 259
13.1.1 Finite Element Method 262
13.1.2 The Streamline - diffusion method (SDM) 264
13.1.3 Exercises 266

Acknowledgment. I wish to thank Niklas Eriksson and Bengt Svensson who have read the entire material and made many valuable suggestions. Niklas has contributed to a better presentation of the text as well as to simplifications and corrections of many key estimates that has substantially improved the quality of this lecture notes. Bengt has made all *xfig* figures. I have used some material from hand written lecture notes by Kenneth Eriksson and Bernt Wennberg. Otherwise I have followed Eriksson et al [12].

Chapter 0

Introduction

This note presents an introduction to the Galerkin finite element method (FEM), as a general tool for numerical solution of differential equations (both ODEs and PDEs). Iteration procedures are included in order to efficiently compute the numerical solutions to matrix equations. Interpolation techniques are presented to derive basic *a priori* and *a posteriori* error estimates necessary in determining qualitative/quantitative properties of the approximate solutions. Some theoretical aspects as existence, uniqueness, stability and convergence are discussed as well. Galerkin's method for solving a general differential equation (both PDEs and ODEs) is based on seeking an approximate solution, which is

1. Easy to differentiate and integrate

2. Spanned by a set of "nearly orthogonal" basis functions in a finite-dimensional vector space.

3. *Galerkin orthogonality*. Roughly speaking, this means that: the error between the exact and approximate solutions is orthogonal to the finite dimensional vector space containing the approximate solution.

0.1 Preliminaries

• A *differential equation* is a relation between an unknown function u and its derivatives $u^{(k)}$, $1 \leq k \leq N$, where k and N are integers.

• If the function $u(x)$ depends on only one variable ($x \in \mathbb{R}$), then the equation is called an *ordinary differential equation* (ODE).

• The *order* of the differential equation is determined by the order of the highest derivative (N) of the function u that appears in the equation.

• If the function $u(x,t)$ depends on more than one variable, then the differential equation is called a *partial differential equation* (PDE), e.g.:

$$u_t(x,t) - u_{xx}(x,t) = 0,$$

is a *homogeneous* PDE of second order whereas

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

is a *non-homogeneous* PDE of second order.

• A solution to a differential equation is a function; e.g. $u(x)$, $u(x,t)$ or $u(x,y)$.

• In general the solution u cannot be expressed in terms of elementary functions and numerical methods are the only way to solve the differential equation by constructing *approximate solutions*. Then the main questions are: *how close is the approximate solution to the exact solution?* and how and in which environment does one measure this *closeness*? In which extent does the approximate solution preserve the physical quality of the exact solution? These are some of the questions that we want to deal with in this text.

• A linear ordinary differential equation of order n has the general form:

$$L(t,D)u = u^{(n)}(t) + a_{n-1}(t)u^{(n-1)}(t) + \ldots + a_1(t)u'(t) + a_0(t)u(t) = b(t),$$

where $D = d/dt$ denotes the derivative, and $u^{(k)} := D^k u$, with $D^k := \frac{d^k}{dt^k}$, $1 \leq k \leq n$ (the k-th order derivative). The corresponding *linear differential operator* is denoted by

$$L(t,D) = \frac{d^n}{dt^n} + a_{n-1}(t)\frac{d^{n-1}}{dt^{n-1}} + \ldots + a_1(t)\frac{d}{dt} + a_0(t).$$

0.2 Trinities

Below we introduce the so called *trinities* classifying the main ingredients involved in the process of identifying and modeling problems in differential equations, see Karl E. Gustafson [14] for details.

The usual three operators involved in differential equations are

$$\text{Laplace operator} \qquad \Delta_n = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \ldots + \frac{\partial^2}{\partial x_n^2}, \tag{0.2.1}$$

$$\text{Diffusion operator} \qquad \frac{\partial}{\partial t} - \Delta_n, \tag{0.2.2}$$

$$\text{D'Alembert operator} \qquad \Box = \frac{\partial^2}{\partial t^2} - \Delta_n, \tag{0.2.3}$$

where we have the space variable $\mathbf{x} := (x_1, x_2, x_3, \ldots, x_n) \in \mathbb{R}^n$, the time variable $t \in \mathbb{R}^+$, and $\partial^2/\partial x_i^2$ denotes the second partial derivative with respect to x_i, $1 \leq i \leq n$. We also define a first order operator, namely the gradient operator ∇_n which is a vector valued operator and is defined as follows

$$\nabla_n = \Big(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n}\Big).$$

Classifying general second order PDEs in two dimensions

The usual three classes of second order partial differential equations are *elliptic*, *parabolic* and *hyperbolic* ones.

A general second order PDE with constant coefficients in 2-D

$$Au_{xx}(x,y)+2Bu_{xy}(x,y)+Cu_{yy}(x,y)+Du_x(x,y)+Eu_y(x,y)+Fu(x,y)+G=0.$$

Here we introduce the *discriminant* $d = AC - B^2$: a quantity that specifies the role of the coefficients in determining the equation type.

Discriminant	**Type of equation**	**Solution behavior**
$d = AC - B^2 > 0$	Elliptic	stationary energy-minimizing
$d = AC - B^2 = 0$	Parabolic	smoothing and spreading flow
$d = AC - B^2 < 0$	Hyperbolic	a disturbance-preserving wave

Example 0.1. *Below are the classes of the most common differential equations:*

Elliptic	**Parabolic**	**Hyperbolic**
Potential equation	Heat equation	Wave Equation
$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$	$\frac{\partial u}{\partial t} - \Delta u = 0$	$\frac{\partial^2 u}{\partial t^2} - \Delta u = 0$
$u_{xx}(x,y) + u_{yy}(x,y) = 0$	$u_t(t,x) - u_{xx}(x,t) = 0$	$u_{tt}(x,t) - u_{xx}(x,t) = 0$
$A = C = 1, B = 0$	$B = C = 0, A = -1$	$A = -1, B = 0, C = 1$
$d = AC - B^2 = 1 > 0$	$d = AC - B^2 = 0$	$d = AC - B^2 = -1 < 0.$

Second order differential equations with variable coefficients in 2-D
In the variable coefficients case, one can only have a local classification.

Example 0.2. *Consider the Tricomi operator of gas dynamics:*

$$Lu(x,y) = yu_{xx} + u_{yy}.$$

Here the coefficient y is not a constant and we have $A = y$, $B = 0$, and $C = 1$. Hence $d = AC - B^2 = y$ and consequently, the domain of ellipticity is $y > 0$, and so on, see Figure 1.

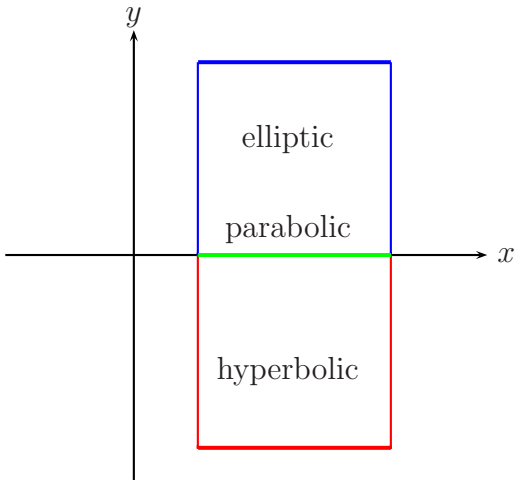


Figure 1: Tricomi: an example of a variable coefficient classification.

• **Summing up and generalizing to n space variables we have**

Mathematical Quantity	**Surname**	**Physical Name**	**Classification Type**
Δ_n	Laplacian	Potential operator	Elliptic
$\frac{\partial}{\partial t} - \Delta_n$	Heat	Diffusion operator	Parabolic
$\Box = \frac{\partial^2}{\partial t^2} - \Delta_n$	d'Alembertian	Wave operator	Hyperbolic

The usual three types of problems in differential equations

1. Initial value problems (IVP)

The simplest differential equation is $u'(x) = f(x)$ for $a < x \leq b$, but also $(u(x) + c)' = f(x)$ for any constant c. To determine a unique solution a specification of the initial value $u(a) = u_0$ is generally required. For example for $f(x) = 2x,\ 0 < x \leq 1$, we have $u'(x) = 2x$ and the general solution is $u(x) = x^2 + c$. With an *initial value* of $u(0) = 0$ we get $u(0) = 0^2 + c = c = 0$. Hence the unique solution to this initial value problem is $u(x) = x^2$. Likewise for a time dependent differential equation of *second order* (two time derivatives) the initial values for $t = 0$, i.e. $u(x, 0)$ and $u_t(x, 0)$, are generally required. For a PDE such as the heat equation the initial value can be a *function* of the space variable.

Example 0.3. *The wave equation, on the real line, augmented with the given initial data:*

$$\begin{cases} u_{tt} - u_{xx} = 0, & -\infty < x < \infty, \quad t > 0, \\ u(x,0) = f(x), \quad u_t(x,0) = g(x), & -\infty < x < \infty, \quad t = 0. \end{cases}$$

2. Boundary value problems (BVP)

a. Boundary value problems in $\mathbb{R}$

Example 0.4. *Consider the stationary heat equation:*

$$-[a(x)u'(x)]' = f(x), \qquad \textit{for } \ 0 < x < 1.$$

In order to determine a solution $u(x)$ uniquely (see Remark 0.1 below), the differential equation is complemented by boundary conditions imposed at the boundary points $x = 0$ and $x = 1$; for example $u(0) = u_0$ and $u(1) = u_1$, where u_0 and u_1 are given real numbers.

b. Boundary value problems in $\mathbb{R}^n$

Example 0.5. *The Laplace equation in $\mathbb{R}^n$, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$:*

$$\begin{cases} \Delta_n u = \dfrac{\partial^2 u}{\partial x_1^2} + \dfrac{\partial^2 u}{\partial x_2^2} + \ldots + \dfrac{\partial^2 u}{\partial x_n^2} = 0, & \mathbf{x} \in \Omega \subset \mathbb{R}^n, \\ u(\mathbf{x}) = f(\mathbf{x}), & \mathbf{x} \in \partial\Omega \text{ (boundary of } \Omega). \end{cases}$$

Remark 0.1. *In general, in order to obtain a unique solution for a (partial) differential equation, one needs to supply physically adequate boundary data. For instance in Example 0.1 for the potential problem $u_{xx} + u_{yy}$, stated in a rectangular domain $(x, y) \in \Omega := (0, 1) \times (0, 1)$, to determine a unique solution we need to give 2 boundary conditions in the x-direction, i.e. the numerical values of $u(0, y)$ and $u(1, y)$, and another 2 in the y-direction: the numerical values of $u(x, 0)$ and $u(x, 1)$; whereas to determine a unique solution for the wave equation $u_{tt} - u_{xx} = 0$, it is necessary to supply 2 initial conditions in the time variable t, and 2 boundary conditions in the space variable x.*

3. Eigenvalue problems (EVP)

Let $\mathbf{A}$ be a given matrix. The relation $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, $\mathbf{v} \neq 0$, is a linear equation system where λ is an *eigenvalue* and $\mathbf{v}$ is an *eigenvector*. In the example below we introduce the corresponding terminology for differential equations.

Example 0.6. *A differential equation describing a steady state vibrating string is given by*

$$-u''(x) = \lambda u(x), \qquad u(0) = u(\pi) = 0,$$

where λ is an eigenvalue and $u(x)$ is an eigenfunction. $u(0) = 0$ and $u(\pi) = 0$ are boundary values.

The differential equation for a time dependent vibrating string with small displacement, which is fixed at the end points, is given by

$$\begin{cases} u_{tt}(x,t) - u_{xx}(x,t) = 0, & 0 < x < \pi, \quad t > 0, \\ u(0,t) = u(\pi,t) = 0, \quad t > 0, & u(x,0) = f(x), \quad u_t(x,0) = g(x). \end{cases}$$

Using separation of variables, see also Folland [11], this equation can be split into two eigenvalue problems: Insert $u(x,t) = X(x)T(t) \neq 0$ (a nontrivial solution) into the differential equation to get

$$u_{tt}(x,t) - u_{xx}(x,t) = X(x)T''(t) - X''(x)T(t) = 0. \tag{0.2.4}$$

Dividing (0.2.4) *by $X(x)T(t) \neq 0$ separates the variables, viz*

$$\frac{T''(t)}{T(t)} = \frac{X''(x)}{X(x)} = \lambda = C \quad \textit{(independent of } x \textit{ and } t\textit{)}. \tag{0.2.5}$$

Consequently we get 2 ordinary differential equations (2 eigenvalue problems):

$$X''(x) = \lambda X(x), \qquad \textit{and} \qquad T''(t) = \lambda T(t). \tag{0.2.6}$$

The usual three types of boundary conditions

1. Dirichlet boundary condition

(the solution is known at the boundary of the domain),

$$u(\mathbf{x},t) = f(\mathbf{x}), \quad \text{for} \quad \mathbf{x} = (x_1, x_2, \ldots, x_n) \in \partial\Omega, \quad t > 0.$$

2. Neumann boundary condition

(the derivative of the solution in the direction of the outward normal is given)

$$\frac{\partial u}{\partial \mathbf{n}} = \mathbf{n} \cdot \operatorname{grad}(u) = \mathbf{n} \cdot \nabla u = f(\mathbf{x}), \qquad \mathbf{x} \in \partial\Omega$$

$\mathbf{n} = \mathbf{n}(\mathbf{x})$ is the *outward unit normal* to $\partial\Omega$ at $\mathbf{x} \in \partial\Omega$, and

$$\mathbf{grad}(u) = \nabla u = \Big(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_n}\Big).$$

3. Robin boundary condition
(a combination of 1 and 2),

$$\frac{\partial u}{\partial \mathbf{n}} + k \cdot u(\mathbf{x}, t) = f(\mathbf{x}), \qquad k > 0, \quad \mathbf{x} = (x_1, x_2, \ldots, x_n) \in \partial\Omega.$$

Example 0.7. *For $u = u(x, y)$ we have $\mathbf{n} = (n_x, n_y)$, with $|\mathbf{n}| = \sqrt{n_x^2 + n_y^2} = 1$ and $\mathbf{n} \cdot \nabla u = n_x u_x + n_y u_y$.*

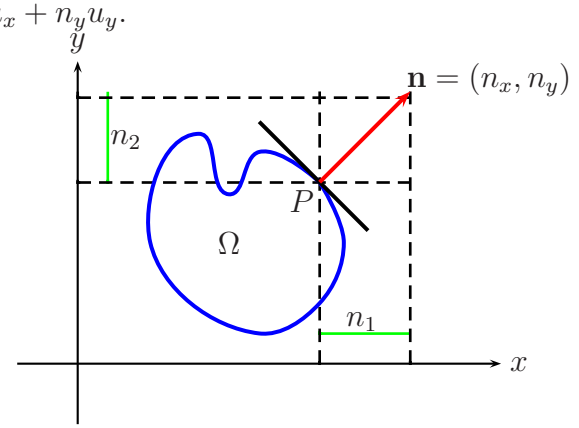


Figure 2: The domain Ω and its outward normal $\mathbf{n}$ at a point $P \in \partial\Omega$.

Example 0.8. *Let $u(x, y) = x^2 + y^2$. We determine the normal derivative of u in direction $\mathbf{v} = (1, 1)$. The gradient of u is the vector valued function $\nabla u = 2x \cdot \mathbf{e}_x + 2y \cdot \mathbf{e}_y$, where $\mathbf{e}_x = (1, 0)$ and $\mathbf{e}_y = (0, 1)$ are the unit orthonormal basis in $\mathbb{R}^2$: $\mathbf{e}_x \cdot \mathbf{e}_x = \mathbf{e}_y \cdot \mathbf{e}_y = 1$ and $\mathbf{e}_x \cdot \mathbf{e}_y = \mathbf{e}_y \cdot \mathbf{e}_x = 0$. Note that $\mathbf{v} = \mathbf{e}_x + \mathbf{e}_y = (1, 1)$ is not a unit vector. The normalized $\mathbf{v}$ is obtained viz $\hat{\mathbf{v}} = \mathbf{v}/|\mathbf{v}|$, i.e.*

$$\hat{\mathbf{v}} = \frac{\mathbf{e}_x + \mathbf{e}_y}{|\mathbf{e}_x + \mathbf{e}_y|} = \frac{(1, 1)}{\sqrt{1^2 + 1^2}} = \frac{(1, 1)}{\sqrt{2}}.$$

Thus with $\nabla u(x, y) = 2x \cdot \mathbf{e}_x + 2y \cdot \mathbf{e}_y$, we get

$$\hat{\mathbf{v}} \cdot \nabla u(x, y) = \frac{\mathbf{e}_x + \mathbf{e}_y}{|\mathbf{e}_x + \mathbf{e}_y|} \cdot (2x \cdot \mathbf{e}_x + 2y \cdot \mathbf{e}_y).$$

which gives

$$\hat{\mathbf{v}} \cdot \nabla u(x, y) = \frac{(1, 1)}{\sqrt{2}} \cdot [2x(1, 0) + 2y(0, 1)] = \frac{(1, 1)}{\sqrt{2}} \cdot (2x, 2y) = \frac{2x + 2y}{\sqrt{2}}.$$

The usual three questions

I. In theory

I1. Existence, at least one solution u

I2. Uniqueness, either one solution or no solutions at all

I3. Stability, continuous dependence of solutions on the data

Note. A property that concerns behavior, such as growth or decay, of perturbations of a solution as time increases is generally called a stability property.

II. In applications

II1. Construction, of the solution.

II2. Regularity, how smooth is the found solution.

II3. Approximation, when an exact construction of the solution is impossible.

Three general approaches to analyzing differential equations

1. Separation of Variables Method

The separation of variables technique reduces the (PDEs) to simpler eigenvalue problems (ODEs). Also called *Fourier method, or solution by eigenfunction expansion* (Fourier analysis).

2. Variational Formulation Method

Variational formulation or the multiplier method is a strategy for extracting information by multiplying a differential equation by suitable test functions and then integrating. This is also referred to as *The Energy Method* (subject of our study).

3. Green's Function Method

Fundamental solutions, or solution of integral equations (is the subject of an advanced PDE course).

Chapter 1

Partial Differential Equations

We recall the common notation $\mathbb{R}^n$ for the real Euclidean spaces of dimension n with the elements $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. In the most applications n will be $1, 2, 3$ or 4 and the variables x_1, x_2 and x_3 denote coordinates in one, two, or three space dimensions, whereas x_4 represents the time variable. In this case we usually replace (x_1, x_2, x_3, x_4) by a most common notation: (x, y, z, t). Further we shall use the subscript notation for the partial derivatives, viz:

$$u_{x_i} = \frac{\partial u}{\partial x_i}, \quad \dot{u} = u_t = \frac{\partial u}{\partial t}, \quad u_{xy} = \frac{\partial^2 u}{\partial x \partial y}, \quad u_{xx} = \frac{\partial^2 u}{\partial x^2}, \quad \ddot{u} = \frac{\partial^2 u}{\partial t^2}, \quad \text{etc.}$$

A more general notation for partial derivatives of a sufficiently smooth function u (see definition below) is given by

$$\frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdot \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdot \ldots \cdot \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} u,$$

where $\frac{\partial^{\alpha_i}}{\partial x_i^{\alpha_i}}$, $1 \leq i \leq n$, denotes the partial derivative of order α_i with respect to the variable x_i, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a multi-index of integers $\alpha_i \geq 0$ and $|\alpha| = \alpha_1 + \ldots + \alpha_n$.

Definition 1.1. *A function f of one real variable is said to be of class $\mathcal{C}^{(k)}$ on an interval I if its derivatives $f', \ldots, f^{(k)}$ exist and are continuous on I. A function f of n real variables is said to be of class $\mathcal{C}^{(k)}$ on a set $S \subset \mathbb{R}^n$ if all of its partial derivatives of order $\leq k$, i.e. $\partial^{|\alpha|} f / (\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n})$ with the multi-index $|\alpha| = \alpha_1 + \ldots + \alpha_n \leq k$, exist and are continuous on S.*

A key defining property of a partial differential equation (PDE) is that there is more than one independent variable and a PDE is a relation between an unknown function $u(x_1, \ldots, x_n)$ and its partial derivatives:

$$F(x_1, \ldots, x_n, u, u_{x_1}, u_{x_2}, \ldots, u_{x_1x_1}, \ldots, \partial^{|\alpha|}u/\partial x_1^{\alpha_1} \ldots \partial x_l^{\alpha_l}, \ldots) = 0. \quad (1.0.1)$$

The order of an equation is defined to be the order of the highest derivative in the equation. The most general PDE of first order in two independent variables can be written as

$$F(x, y, u(x, y), u_x(x, y), u_y(x, y)) = F(x, y, u, u_x, u_y) = 0. \quad (1.0.2)$$

Likewise the most general PDE of second order in two independent variables is of the form

$$F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0. \quad (1.0.3)$$

It turns out that, when the equations (1.0.1)-(1.0.3) are considered in bounded domains $\Omega \subset \mathbb{R}^n$, in order to obtain a *unique solution* (see below) one should provide conditions at the boundary of the domain Ω called *boundary conditions*, denoted, e.g. by $B(u) = f$ or $B(u) = 0$ (as well as conditions for $t = 0$, *initial conditions*; denoted, e.g. by $I(u) = g$ or $I(u) = 0$), as in the theory of ordinary differential equations. B and I are expressions of u and its partial derivatives, stated on the whole or a part of the boundary of Ω (or, in case of I, for $t = 0$), and are associated to the underlying PDE. Below we shall discuss the choice of relevant initial and boundary conditions for a PDE.

A *solution* of a PDE of type (1.0.1)-(1.0.3) is a function u that identically satisfies the corresponding PDE, and the associated initial and boundary conditions, in some region of the variables x_1, $x_2, \ldots, x_n$ (or x, y). Note that a solution of an equation of order k has to be k times differentiable. A function in $\mathcal{C}^{(k)}$ that satisfies a PDE of order k is called a classical (or strong) solution of the PDE. We sometimes also have to deal with solutions that are not classical. Such solutions are called weak solutions. In this note, in the variational formulation for finite element methods, we actually deal with weak solutions. For a more thorough discussion on weak solutions see, e.g. any textbook in distribution theory.

Definition 1.2 (Hadamard's criteria)**.** *A problem consisting of a PDE associated with boundary and/or initial conditions is called well-posed if it fulfills the following three criteria:*
1. **Existence** *The problem has a solution.*

2. **Uniqueness** *There is no more than one solution.*
3. **Stability** *A small change in the equation or in the side (initial and/or boundary) conditions gives rise to a small change in the solution.*

If one or more of the conditions above does not hold, then we say that the problem is *ill-posed*. The fundamental theoretical question of PDEs is whether the problem consisting of the equation and its associated side conditions is well-posed. However, in certain engineering applications we might encounter problems that are ill-posed. In practice, such problems are unsolvable. Therefore, when we face an ill-posed problem, the first step should be to modify it appropriately in order to render it well-posed.

Definition 1.3. *An equation is called linear if in* (1.0.1)*, F is a linear function of the unknown function u and its derivatives.*

Thus, for example, the equation $e^{x^2y}u_x + x^7u_y + \cos(x^2+y^2)u = y^3$ is a linear equation, while $u_x^2 + u_y^2 = 1$ is a nonlinear equation. The nonlinear equations are often further classified into subclasses according to the type of their nonlinearity. Generally, the nonlinearity is more pronounced when it appears in higher order derivatives. For example, the following equations are both nonlinear

$$u_{xx} + u_{yy} = u^3 + u. \tag{1.0.4}$$

$$u_{xx} + u_{yy} = |\nabla u|^2 u. \tag{1.0.5}$$

Here $|\nabla u|$ denotes the norm of the gradient of u. While (1.0.5) is nonlinear, it is still linear as a function of the highest-order derivative. Such a nonlinearity is called *quasilinear*. On the other hand in (1.0.4) the nonlinearity is only in the unknown solution u. Such equations are called *semilinear*.

1.1 Differential operators, superposition

We recall that we denote the set of continuous functions in a domain D (a subset of $\mathbb{R}^n$) by $\mathcal{C}^0(D)$ or $\mathcal{C}(D)$. Further, by $\mathcal{C}^{(k)}(D)$ we mean the set of all functions that are k times continuously differentiable in D. *Differential and integral operators* are examples of mappings between function classes as $\mathcal{C}^{(k)}$. We denote by $\mathcal{L}[u]$ the operation of a mapping (operator) $\mathcal{L}$ on a function u.

Definition 1.4. *An operator $\mathcal{L}$ that satisfies*

$$\mathcal{L}[\beta_1 u_1 + \beta_2 u_2] = \beta_1 \mathcal{L}[u_1] + \beta_2 \mathcal{L}[u_2], \qquad \forall \beta_1, \beta_2 \in \mathbb{R}, \tag{1.1.1}$$

where u_1 and u_2 are functions, is called a linear operator. We may generalize (1.1.1) *as*

$$\mathcal{L}[\beta_1 u_1 + \ldots + \beta_k u_k] = \beta_1 \mathcal{L}[u_1] + \ldots + \beta_k \mathcal{L}[u_k], \quad \forall \beta_1, \ldots, \beta_k \in \mathbb{R}, \quad (1.1.2)$$

i.e. $\mathcal{L}$ maps any linear combination of u_j's to corresponding linear combination of $\mathcal{L}[u_j]$'s.

For instance the *integral operator* $\mathcal{L}[f] = \int_a^b f(x)\,dx$ defined on the space of continuous functions on $[a, b]$ defines a linear operator from $\mathcal{C}[a, b]$ into $\mathbb{R}$, which satisfies both (1.1.1) and (1.1.2).

A *linear partial differential operator* $\mathcal{L}$ that transforms a function u of the variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ into another function $\mathcal{L}u$ is given by

$$\mathcal{L}[\bullet] = a(\mathbf{x})\,\bullet + \sum_{i=1}^{n} b_i(\mathbf{x}) \frac{\partial\,\bullet}{\partial x_i} + \sum_{i,j=1}^{n} c_{ij}(\mathbf{x}) \frac{\partial^2\,\bullet}{\partial x_i \partial x_j} + \ldots \quad (1.1.3)$$

where $\bullet$ represents any function u in, say $\mathcal{C}^{(\ell)}$, and the dots at the end indicate higher-order derivatives, but the sum contains only *finitely* many terms.

The term *linear* in the phrase *linear partial differential operator* refers to the following fundamental property: if $\mathcal{L}$ is given by (1.1.3) and u_j, $1 \leq j \leq k$, are any set of functions possessing the requisite derivatives, and β_j, $1 \leq j \leq k$, are any constants then the relation (1.1.2) is fulfilled. This is an immediate consequence of the fact that (1.1.1) and (1.1.2) are valid for $\mathcal{L}$ replaced with the derivative. A linear differential equation defines a linear differential operator: the equation can be expressed as $\mathcal{L}[u] = F$, where $\mathcal{L}$ is a linear operator and F is a given function. The differential equation $\mathcal{L}[u] = 0$ is called a *homogeneous equation*. For example, define the operator $\mathcal{L} = \partial^2/\partial x^2 - \partial^2/\partial y^2$. Then

$$\mathcal{L}[u] = u_{xx} - u_{yy} = 0,$$

is a homogeneous equation, while the equation

$$\mathcal{L}[u] = u_{xx} - u_{yy} = x,$$

is an example of a *nonhomogeneous equation*. In a similar way we may define another type of constraint for the PDEs that appears in many applications:

the boundary conditions. In this regard the linear boundary conditions are defined as operators B satisfying

$$B(\beta_1 u_1 + \beta_2 u_2) = \beta_1 B(u_1) + \beta_2 B(u_2), \qquad \forall \beta_1, \beta_2 \in \mathbb{R}, \tag{1.1.4}$$

at the boundary of a given domain Ω.

The Superposition principle. An important property of the linear operators is that if the functions u_j, $1 \leq j \leq k$, satisfy the linear differential equations $\mathcal{L}[u_j] = F_j$, and the boundary conditions (linear) $B(u_j) = f_j$ for $j = 1, 2, \ldots, k$, then any linear combination $v := \sum_{i=1}^{\ell} \beta_i u_i$, $\ell \leq k$, satisfies the equation $\mathcal{L}[v] = \sum_{i=1}^{\ell} \beta_i F_i$ as well as the boundary condition $B(v) = \sum_{i=1}^{\ell} \beta_i f_i$. In particular, if each of the functions u_i, $1 \leq i \leq \ell$, satisfies the homogeneous equation $\mathcal{L}[u] = 0$ and the homogeneous boundary condition $B(u) = 0$, then every linear combination of them satisfies that equation and boundary condition too. This property is called the *superposition principle*. It allows to construct complex solutions through combining simple solutions: suppose we want to determine all solutions of a differential equation associated with a boundary condition viz,

$$\mathcal{L}[u] = F, \qquad B(u) = f. \tag{1.1.5}$$

We consider the corresponding, *simpler homogeneous problem*:

$$\mathcal{L}[u] = 0, \qquad B(u) = 0. \tag{1.1.6}$$

Now it suffices to find just one solution, say v of the original problem (1.1.5). Then, for any solution u of (1.1.5), $w = u - v$ satisfies (1.1.6), since $\mathcal{L}[w] = \mathcal{L}[u] - \mathcal{L}[v] = F - F = 0$ and $B(w) = B(u) - B(v) = f - f = 0$. Hence we obtain a general solution of (1.1.5) by adding the general solution w of (1.1.6) to any particular solution of (1.1.5).

Following the same idea one may apply superposition to split a problem involving several inhomogeneous terms into simpler ones each with a single inhomogeneous term. For instance we may split (1.1.5) as

$$\begin{aligned} \mathcal{L}[u_1] &= F, \qquad B(u_1) = 0, \\ \mathcal{L}[u_2] &= 0, \qquad B(u_2) = f, \end{aligned}$$

and then take $u = u_1 + u_2$.

The most important application of the superposition principle is in the homogeneous case: linear homogeneous differential equations satisfying homogeneous boundary conditions (which we repeat from above).

The Superposition principle for the homogeneous case. *If the functions u_j, $1 \le j \le k$, satisfy* (1.1.6)*: the linear differential equation $\mathcal{L}[u_j] = 0$ and the boundary conditions (linear) $B(u_j) = 0$ for $j = 1, 2, \ldots, k$, then any linear combination $v := \sum_{i=1}^{\ell} \beta_i u_i$, $\ell \le k$, satisfies the same equation and boundary condition:* (1.1.6).

Finally, the superposition principle is used to prove uniqueness of solutions to linear PDEs.

1.1.1 Exercises

Problem 1.1. *Consider the problem*

$$u_{xx} + u = 0, \quad x \in (0, \ell); \qquad u(0) = u(\ell) = 0.$$

Clearly the function $u(x) \equiv 0$ is a solution. Is this solution unique? Does the answer depend on ℓ?

Problem 1.2. *Consider the problem*

$$u_{xx} + u_x = f(x), \quad x \in (0, \ell); \qquad u(0) = u'(0) = \frac{1}{2}[u'(\ell) + u(\ell)].$$

a) Is the solution unique? (f is a given function).
b) Under what condition on f a solution exists?

Problem 1.3. *Suppose u_i, $i = 1, 2, \ldots, N$ are N solutions of the linear differential equation $\mathcal{L}[u] = F$, where $F \neq 0$. Under what condition on the constant coefficients c_i, $i = 1, 2, \ldots, N$ is the linear combination $\sum_{i=1}^{N} c_i u_i$ also a solution of this equation?*

Problem 1.4. *Consider the nonlinear ordinary differential equation $u_x = u(1 - u)$.*
a) Show that $u_1(x) \equiv 1$ and $u_2(x) = 1 - 1/(1 + e^x)$ both are solutions, but $u_1 + u_2$ is not a solution.
b) For which value of c_1 is $c_1 u_1$ a solution? What about $c_2 u_2$?

Problem 1.5. *Show that each of the following equations has a solution of the form $u(x,y) = f(ax+by)$ for a proper choice of constants a, b. Find the constants for each example.*

a) $u_x + 3u_y = 0$. *b)* $3u_x - \pi u_y = 0$. *c)* $2u_x + eu_y = 0$.

Problem 1.6. *a) Consider the equation $u_{xx} + 2u_{xy} + u_{yy} = 0$. Write the equation in the coordinates $s = x$, $t = x - y$.*
b) Find the general solution of the equation.
c) Consider the equation $u_{xx} - 2u_{xy} + 5u_{yy} = 0$. Write it in the coordinates $s = x + y$ and $t = 2x$.

Problem 1.7. *a) Show that for $n = 1, 2, 3, \ldots$, $u_n(x,y) = \sin(n\pi x)\sinh(n\pi y)$ satisfies*

$$u_{xx} + u_{yy} = 0, \quad u(0,y) = u(1,y) = u(x,0) = 0.$$

b) Find a linear combination of u_n's that satisfies $u(x,1) = \sin 2\pi x - \sin 3\pi x$.
c) Solve the Dirichlet problem

$$\begin{aligned} u_{xx} + u_{yy} = 0, &\quad u(0,y) = u(1,y) = 0, \\ u(x,0) = 2\sin \pi x, &\quad u(x,1) = \sin 2\pi x - \sin 3\pi x. \end{aligned}$$

1.2 Some equations of mathematical physics

Below we introduce some of the basic partial differential equations of mathematical physics that will be the subject of our studies throughout the book. These equations all involve a fundamental differential operator of order two, called the *Laplacian*, acting on $\mathcal{C}^{(2)}(\mathbb{R}^n)$ and defined as follows:

$$\nabla^2 u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2}, \qquad u \in \mathcal{C}^{(2)}(\mathbb{R}^n). \tag{1.2.1}$$

Basically, there are three types of fundamental physical phenomena described by differential equations involving the Laplacian:

$$\begin{aligned} \nabla^2 u &= F(\mathbf{x}), && \textit{The Poisson equation} \\ u_t - k\nabla^2 u &= F(\mathbf{x},t), && \textit{The heat equation} \\ u_{tt} - c^2\nabla^2 u &= F(\mathbf{x},t), && \textit{The wave equation.} \end{aligned} \tag{1.2.2}$$

Here F is a given function. If $F \neq 0$, then the equations (1.2.2) are *inhomogeneous*. In the special case when $F \equiv 0$, the equations (1.2.2) are *homogeneous*, then the first equation is called *the Laplace equation*.

Here the first equation, being time independent, has a particular nature: besides the fact that it describes the steady-state heat transfer and the standing wave equations (loosely speaking, the time independent versions of the other two equations), the Laplace equation (the first equation in (1.2.2) with $F \equiv 0$) arises in describing several other physical phenomena such as *electrostatic potential* in regions with no electric charge, *gravitational potential* in the regions with no mass distribution, as well as problems in *elasticity*, etc.

A model for the stationary heat equation in one dimension
We model heat conduction in a thin heat-conducting wire stretched between the two endpoints of an interval $[a, b]$ that is subject to a heat source of intensity $f(x)$, see Figure 1.1. We are interested in the stationary distribution of temperature $u(x)$ in the wire.

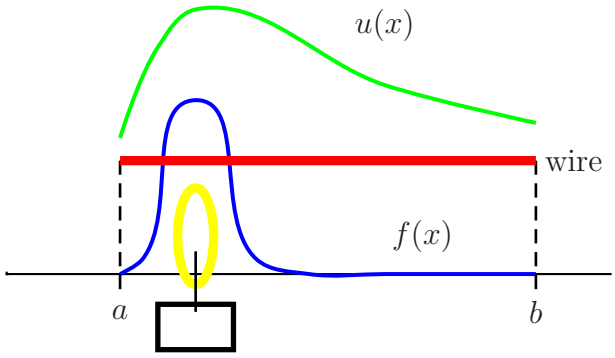


Figure 1.1: A heat-conducting 1 dimensional wire.

Let $q(x)$ denote the heat flux in the direction of the positive x-axis in the wire $a < x < b$. Conservation of energy in the stationary case requires that the net heat through the endpoints of an arbitrary subinterval (x_1, x_2) of (a, b) is equal to the heat produced in (x_1, x_2) per unit time:

$$q(x_2) - q(x_1) = \int_{x_1}^{x_2} f(x)\, dx.$$

By the *Fundamental Theorem of Calculus*,

$$q(x_2) - q(x_1) = \int_{x_1}^{x_2} q'(x)\, dx,$$

hence we conclude that

$$\int_{x_1}^{x_2} q'(x)\,dx = \int_{x_1}^{x_2} f(x)\,dx.$$

Since x_1 and x_2 are arbitrary, assuming that the integrands are continuous, yields

$$q'(x) = f(x), \qquad \text{for } \ a < x < b, \tag{1.2.3}$$

which expresses conservation of energy in differential equation form. We need an additional equation that relates the heat flux q to the temperature gradient u' called a *constitutive equation*. The simplest constitutive equation for heat flow is *Fourier's law:*

$$q(x) = -c(x)u'(x), \tag{1.2.4}$$

which states that heat flows from warm regions to cold regions at a rate proportional to the temperature gradient $u'(x)$. The constant of proportionality is the *coefficient of heat conductivity* $c(x)$, which we assume to be a positive function in $[a, b]$. Combining (1.2.3) and (1.2.4) gives the *stationary heat equation in one dimension*:

$$-(c(x)u'(x))' = f(x), \qquad \text{for } \ a < x < b. \tag{1.2.5}$$

To define a solution u uniquely, the differential equation is complemented by *boundary conditions* imposed at the boundaries $x = a$ and $x = b$. A common example is the homogeneous *Dirichlet conditions* $u(a) = u(b) = 0$, corresponding to keeping the temperature zero at the endpoints of the wire. The result is a *two-point boundary value problem:*

$$\begin{cases} -(c(x)u'(x))' = f(x), & \text{in } \ (a, b), \\ u(a) = u(b) = 0. \end{cases} \tag{1.2.6}$$

The boundary condition $u(a) = 0$ may be replaced by $-c(a)u'(a) = q(a) = 0$, corresponding to prescribing zero heat flux, or insulating the wire, at $x = a$. Later, we also consider non-homogeneous boundary conditions of the form $u(a) = u_a$ or $q(a) = g$ where u_a and g may be different from zero.

The time-dependent heat equation in (1.2.2) describes the diffusion of thermal energy in a homogeneous material where $u = u(\mathbf{x}, t)$ is the temperature at a position $\mathbf{x}$ at time t and $k(\mathbf{x})$ is called *thermal diffusivity* or *heat conductivity* (corresponding to $c(x)$ in (1.2.4)-(1.2.6)) of the material.

Remark 1.1. *The heat equation can be used to model the heat flow in solids and fluids, in the later case, however, it does not take into account the convection phenomenon; and provides a reasonable model only if phenomena such as macroscopic currents in the fluid are not present (or negligible). Further, the heat equation is not a fundamental law of physics, and it does not give reliable answers at very low or very high temperatures.*

Since temperature is related to heat, which is a form of energy, the basic idea in deriving the heat equation is to use the *law of conservation of energy*.

Fourier's law of heat conduction and derivation of the heat equation Let $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, be a fixed spatial domain with boundary $\partial\Omega$. The rate of change of thermal energy with respect to time in Ω is equal to the net flow of energy across the boundary of Ω plus the rate at which heat is generated within Ω.

Let $u(\mathbf{x}, t)$ denote the temperature at the position $\mathbf{x} = (x, y, z) \in \Omega$ and at time t. We assume that the solid is at rest and that it is rigid, so that the only energy present is thermal energy and the density $\rho(\mathbf{x})$ is independent of the time t and temperature u. Let $\mathcal{E}$ denote the energy per unit mass. Then the amount of thermal energy in Ω is given by

$$\int_\Omega \rho\, \mathcal{E}\, d\mathbf{x},$$

and the time rate (time derivative) of change of thermal energy in Ω is:

$$\frac{d}{dt} \int_\Omega \rho\, \mathcal{E}\, d\mathbf{x} = \int_\Omega \rho\, \mathcal{E}_t\, d\mathbf{x}.$$

Let $\mathbf{q} = (q_x, q_y, q_z)$ denote the heat flux vector and $\mathbf{n} = (n_x, n_y, n_z)$ denote the outward unit normal to the boundary $\partial\Omega$, at the point $\mathbf{x} \in \partial\Omega$. Then $\mathbf{q} \cdot \mathbf{n}$ represents the flow of heat per unit cross-sectional area per unit time crossing a surface element. Thus

$$-\int_{\partial\Omega} q \cdot \mathbf{n}\, dS$$

is the amount of heat per unit time flowing into Ω across the boundary $\partial\Omega$. Here dS represents the element of surface area. The minus sign reflects the fact that if more heat flows out of the domain Ω than in, the energy in Ω decreases. Finally, in general, the heat production is determined by external

sources that are independent of the temperature. In some cases (such as an air conditioner controlled by a thermostat) it depends on temperature itself but not on its derivatives. Hence in the presence of a source (or sink) we denote the corresponding rate at which heat is produced per unit volume by $f = f(\mathbf{x}, t, u)$ so that the source term becomes

$$\int_\Omega f(\mathbf{x}, t, u)\, d\mathbf{x}.$$

Now the law of conservation of energy takes the form

$$\int_\Omega \rho\, \mathcal{E}_t\, d\mathbf{x} + \int_{\partial\Omega} \mathbf{q}\cdot\mathbf{n}\, dS = \int_\Omega f(\mathbf{x}, t, u)\, d\mathbf{x}. \tag{1.2.7}$$

Applying the Gauss divergence theorem to the integral over $\partial\Omega$ we get

$$\int_\Omega (\rho\ \mathcal{E}_t\ + \nabla\cdot\mathbf{q} - f)\, d\mathbf{x} = 0, \tag{1.2.8}$$

where $\nabla\cdot$ denotes the divergence operator. In the sequel we shall use the following simple result:

Lemma 1.1. *Let h be a continuous function satisfying $\int_\Omega h(\mathbf{x})\, d\mathbf{x} = 0$ for every domain $\Omega \subset \mathbb{R}^d$. Then $h \equiv 0$.*

Proof. Let us assume to the contrary that there exists a point $\mathbf{x_0} \in \mathbb{R}^d$ where $h(\mathbf{x_0}) \neq 0$. Assume without loss of generality that $h(\mathbf{x_0}) > 0$. Since h is continuous, there exists a domain (maybe very small) Ω, containing $\mathbf{x_0}$, and $\varepsilon > 0$, such that $h(\mathbf{x}) > \varepsilon$, for all $\mathbf{x} \in \Omega$. Therefore we have $\int_\Omega h(\mathbf{x})\, d\mathbf{x} > \varepsilon \mathrm{Vol}(\Omega) > 0$, which contradicts the lemma's assumption. □

From (1.2.8), using the above lemma, we conclude that

$$\rho\, \mathcal{E}_t\ = -\nabla\cdot\mathbf{q} + f. \tag{1.2.9}$$

This is the basic form of our heat conduction law. The functions $\mathcal{E}$ and q are unknown and additional information of an empirical nature is needed to determine the equation for the temperature u. First, for many materials, over a fairly wide but not too large temperature range, the function $\mathcal{E} = \mathcal{E}(u)$ depends nearly linearly on u, so that

$$\mathcal{E}_t = \lambda u_t. \tag{1.2.10}$$

Here λ, called the *specific heat*, is assumed to be constant. Next, we relate the temperature u to the heat flux q. Here we use *Fourier's law* but, first, to be specific, we describe the simple facts supporting Fourier's law:
(i) Heat flows from regions of high temperature to regions of low temperature.
(ii) The rate of heat flow is small or large accordingly as temperature changes between neighboring regions are small or large.

To describe these quantitative properties of heat flow, we postulate a linear relationship between the rate of heat flow and the rate of temperature change. Recall that if $\mathbf{x}$ is a point in the heat conducting medium and $\mathbf{n}$ is a unit vector specifying a direction at $\mathbf{x}$, then the rate of heat flow at $\mathbf{x}$ in the direction $\mathbf{n}$ is $\mathbf{q} \cdot \mathbf{n}$ and the rate of change of the temperature is $\partial u/\partial \mathbf{n} = \nabla u \cdot \mathbf{n}$, the directional derivative of the temperature. Since $\mathbf{q} \cdot \mathbf{n} > 0$ requires $\nabla u \cdot \mathbf{n} < 0$, and vice versa, (from calculus the direction of maximal growth of a function is given by its gradient), our linear relation takes the form $\mathbf{q} \cdot \mathbf{n} = -\kappa \nabla u \cdot \mathbf{n}$, with $\kappa = \kappa(\mathbf{x}) > 0$. Since $\mathbf{n}$ specifies any direction at $\mathbf{x}$, this is equivalent to the assumption

$$\mathbf{q} = -\kappa \nabla u, \tag{1.2.11}$$

which is *Fourier's law*. The positive function κ is called the *heat conduction (or Fourier) coefficient*. Let now $\sigma = \kappa/\lambda\rho$ and $F = f/\lambda\rho$ and insert (1.2.10) and (1.2.11) into (1.2.9) to get the final form of the heat equation:

$$u_t = \nabla \cdot (\sigma \nabla u) + F. \tag{1.2.12}$$

The quantity σ is referred to as the *thermal diffusivity (or diffusion) coefficient*. If we assume that σ is constant, then the final form of the heat equation would be

$$u_t = \sigma \nabla^2 u + F, \qquad \text{or} \quad u_t = \sigma\, \Delta u + F. \tag{1.2.13}$$

Here $\Delta = \text{div}\nabla = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ denotes the Laplace operator in three dimensions.

The third equation in (1.2.2) is the wave equation: $u_{tt} - c^2 \nabla^2 u = F$. Here u represents a wave traveling through an n-dimensional medium; c is the speed of propagation of the wave in the medium and $u(\mathbf{x}, t)$ is the amplitude of the wave at position $\mathbf{x}$ and time t. The wave equation provides a mathematical model for a number of problems involving different physical processes as, e.g. in the following examples:

(i) Vibration of a stretched string, such as a violin string (1-dimensional).
(ii) Vibration of a column of air, such as a clarinet (1-dimensional).
(iii) Vibration of a stretched membrane, such as a drumhead (2-dimensional).
(iv) Waves in an incompressible fluid, such as water (2-dimensional).
(v) Sound waves in air or other elastic media (3-dimensional).
(vi) Electromagnetic waves, such as light waves and radio waves (3-dimensional).

Note that in (i), (iii) and (iv), u represents the transverse displacement of the string, membrane, or fluid surface; in (ii) and (v), u represents the longitudinal displacement of the air; and in (vi), u is any of the components of the electromagnetic field. For detailed discussions and a derivation of the equations modeling (i)-(vi), see, e.g. Folland [11], Strauss [23] and Taylor [24]. We should point out, however, that in most cases the derivation involves making some simplifying assumptions. Hence, the wave equation gives only an approximate description of the actual physical process, and the validity of the approximation will depend on whether certain physical conditions are satisfied. For instance, in example (i) the vibration should be small enough so that the string is not stretched beyond its limits of elasticity. In example (vi), it follows from Maxwell's equations, the fundamental equations of electromagnetism, that the wave equation is satisfied *exactly* in regions containing no electrical charges or current, which of course cannot be guaranteed under normal physical circumstances and can only be approximately justified in the real world. So an attempt to derive the wave equation corresponding to each of these examples from physical principles is beyond the scope of these notes. Nevertheless, to give an idea, below we shall derive the wave equation for a vibrating string which is, by the way, the most considered model.

The vibrating string, derivation of the wave equation in 1D
Consider a perfectly elastic and flexible string stretched along the segment $[0, L]$ of the x-axis, moving perpendicular to its equilibrium position. Let $\rho_0(x)$ denote the density of the string in the equilibrium position and $\rho(x,t)$ the density at time t. In an arbitrary small interval $[x, x+\Delta x]$ the mass will satisfy, see Figure 1.2.

$$\int_x^{x+\Delta x} \rho_0(x)\,dx = m = \int_x^{x+\Delta x} \rho(x,t)\sqrt{1+u_x^2}\,dx. \tag{1.2.14}$$

Thus, using Lemma 1.1, (1.2.14) gives the conservation of mass:

$$\rho_0(x) = \rho(x,t)\sqrt{1+u_x^2}. \tag{1.2.15}$$

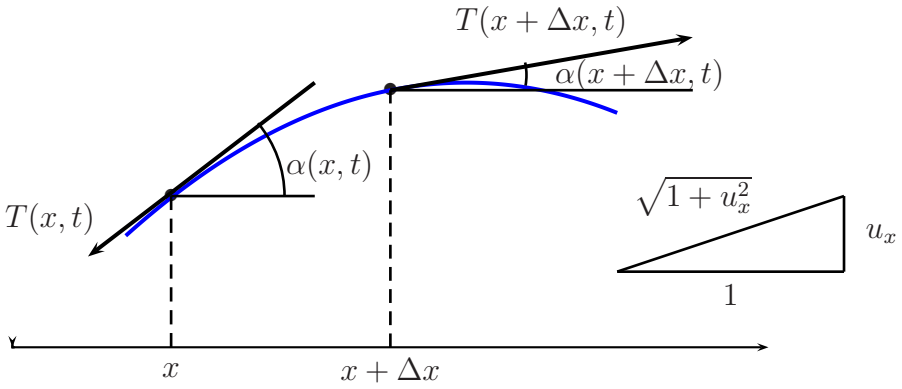


Figure 1.2: A vibrating string.

Now we use the tensions $T(x,t)$ and $T(x+\Delta x,t)$, at the endpoints of an element of the string and determine the forces acting on the interval $[x, x+\Delta x]$. Since we assumed that the string moves only vertically, the forces in the horizontal direction should be in balance: i.e.,

$$T(x+\Delta x,t)\cos\alpha(x+\Delta x,t) - T(x,t)\cos\alpha(x,t) = 0. \tag{1.2.16}$$

Dividing (1.2.16) by Δx and letting $\Delta x \to 0$, we thus obtain

$$\frac{\partial}{\partial x}\Big(T(x,t)\cos\alpha(x,t)\Big) = 0, \tag{1.2.17}$$

hence

$$T(x,t)\cos\alpha(x,t) = \tau(t), \tag{1.2.18}$$

where $\tau(t) > 0$ because it is the magnitude of the horizontal component of the tension.

On the other hand the vertical motion is determined by the fact that the time rate of change of linear momentum is given by the sum of the forces acting in the vertical direction. Hence, using (1.2.15), the momentum of the small element $[x, x+\Delta x]$ is given by

$$\int_x^{x+\Delta x} \rho_0(x) u_t \, dx = \int_x^{x+\Delta x} \rho(x,t)\sqrt{1+u_x^2}\, u_t \, dx, \tag{1.2.19}$$

with the time rate of change:

$$\frac{d}{dt}\int_x^{x+\Delta x} \rho_0 u_t\,dx = \int_x^{x+\Delta x} \rho_0 u_{tt}\,dx. \tag{1.2.20}$$

There are two kinds of forces acting on the segment $[x, x+\Delta x]$ of the string: (i) the forces due to tension that keep the string taut and whose horizontal components are in balance, and (ii) the forces acting along the whole length of the string, such as weight. Thus, using (1.2.18), the net tension force acting on the ends of the string element $[x, x+\Delta x]$ is

$$\begin{aligned} T(x+\Delta x,t)\sin\alpha(x+\Delta x,t) - T(x,t)\sin\alpha(x,t) & \\ = \tau\Big(\frac{\sin\alpha(x+\Delta x,t)}{\cos\alpha(x+\Delta x,t)} - \frac{\sin\alpha(x,t)}{\cos\alpha(x,t)}\Big) & \\ = \tau\Big(\tan\alpha(x+\Delta x,t) - \tan\alpha(x,t)\Big) & \\ = \tau\Big(u_x(x+\Delta x,t) - u_x(x,t)\Big). & \end{aligned} \tag{1.2.21}$$

Further, the weight of the string acting downward is

$$-\int \rho\, g\, dS = -\int_x^{x+\Delta x} \rho\, g\sqrt{1+u_x^2}\,dx = -\int_x^{x+\Delta x} \rho_0 g\,dx. \tag{1.2.22}$$

Next, for an external load, with density $f(x,t)$, acting on the string (e.g., when a violin string is bowed), we have

$$\int \rho\, f\, dS = \int_x^{x+\Delta x} \rho_0 f(x,t)\,dx. \tag{1.2.23}$$

Finally, one should model the friction forces acting on the string segment. We shall assume a linear law of friction of the form:

$$-\int \sigma\,\rho\,u_t\,dS = -\int_x^{x+\Delta x} \sigma\,\rho\,\sqrt{1+u_x^2}\,u_t\,dx = -\int_x^{x+\Delta x} \sigma\,\rho_0\,u_t\,dx. \tag{1.2.24}$$

Now applying Newton's second law yields

$$\begin{aligned} \int_x^{x+\Delta x} \rho_0 u_{tt}\,dx = \tau[u_x(x+\Delta x,t) - u_x(x,t)] & \\ - \int_x^{x+\Delta x} \sigma\,\rho_0\,u_t\,dx + \int_x^{x+\Delta x} \rho_0(f-g)\,dx. & \end{aligned} \tag{1.2.25}$$

Dividing (1.2.25) by Δx and letting $\Delta x \to 0$ we obtain the equation

$$\rho_0 u_{tt} = \tau u_{xx} - \sigma\, \rho_0\, u_t + \rho_0(f - g). \tag{1.2.26}$$

Letting $c^2 = \tau/\rho_0$ and $F = f - g$, we end up with the following concise form:

$$u_{tt} + \sigma\, u_t = c^2 u_{xx} + F. \tag{1.2.27}$$

Equation (1.2.27) describes the vibration of the considered string once it is set into motion. The *smallness* assumption here results in a single linear equation for u. Due to the presence of the friction term σu_t, equation (1.2.27) is often referred to as the *damped one-dimensional wave equation*. If friction is negligible, then we can let $\sigma = 0$ and get the *inhomogeneous wave equation*

$$u_{tt} = c^2 u_{xx} + F. \tag{1.2.28}$$

In the absence of external forces and when the weight of the string is negligible, we may take $F \equiv 0$ to get the *one-dimensional wave equation*:

$$u_{tt} = c^2 u_{xx}. \tag{1.2.29}$$

Note that since u has the unit of length ℓ, u_{tt} has the unit of acceleration and u_{xx} the unit of ℓ^{-1}, hence c has the unit of velocity.

1.2.1 Exercises

Problem 1.8. *Show that* $u(x, y) = \log(x^2 + y^2)$ *satisfies Laplace's equation* $u_{xx} + u_{yy} = 0$ *for* $(x, y) \neq (0, 0)$.

Problem 1.9. *Show that* $u(x, y, z) = (x^2 + y^2 + z^2)^{-1/2}$ *satisfies Laplace's equation* $u_{xx} + u_{yy} + u_{zz} = 0$, *for* $(x, y, z) \neq (0, 0, 0)$.

Problem 1.10. *Show that* $u(r, \theta) = Br^n \sin(n\theta)$ *satisfies the Laplace equation in polar coordinates:*

$$u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta} = 0.$$

Problem 1.11. *Verify that*

$$u = \frac{-2y}{x^2+y^2+2x+1}, \qquad v = \frac{x^2+y^2-1}{x^2+y^2+2x+1}$$

both satisfy the Laplace equation, and sketch the curves u = constant and v = constant. Show that

$$u + iv = \frac{i(z-1)}{z+1}, \qquad \text{where} \quad z = x + iy.$$

Problem 1.12. *Show that $u(x,t) = t^{-1/2}\exp(-x^2/4kt)$ satisfies the heat equation $u_t = ku_{xx}$, for $t > 0$.*

Problem 1.13. *Show that $u(x,y,t) = t^{-1}\exp[-(x^2+y^2)/4kt]$ satisfies the heat equation $u_t = k(u_{xx}+u_{yy})$, for $t > 0$.*

Problem 1.14. *The spherically symmetric form of the heat conduction equation is given by*

$$u_{rr} + \frac{2}{r}u_r = \frac{1}{\kappa}u_t.$$

Show that $v = ru$ satisfies the standard one-dimensional heat equation.

Problem 1.15. *Show that the equation*

$$\theta_t = \kappa\theta_{xx} - h(\theta - \theta_0)$$

can be reduced to the standard heat conduction equation by writing $u = e^{ht}(\theta - \theta_0)$. How do you interpret the term $h(\theta-\theta_0)$?

Problem 1.16. *Use the substitution $\xi = x - vt$, $\eta = t$ to transform the one-dimensional convection-diffusion equation*

$$u_t = ku_{xx} - vu_x,$$

into a heat equation for $\tilde{u}(\xi,\eta) = u(\xi + v\eta, \eta)$.

Problem 1.17. *If $f \in C[0,1]$, let $u(x,t)$ satisfy*

$$\begin{cases} u_t = u_{xx}, & 0 < x < 1, \quad t > 0, \\ u(0,t) = u(1,t) = 0, & t \geq 0, \\ u(x,0) = f(x), & 0 \leq x \leq 1. \end{cases}$$

Derive the identity $2u(u_t - u_{xx}) = (u^2)_t - (2uu_x)_x + 2u_x^2$.

Problem 1.18. *Find the possible values of a and b in the expression $u = \cos at \sin bx$, such that it satisfies the wave equation $u_{tt} = c^2 u_{xx}$.*

Problem 1.19. *Taking $u = f(x + \alpha t)$, where f is any function, find the values of α that will ensure u satisfies the wave equation $u_{tt} = c^2 u_{xx}$.*

Problem 1.20. *The spherically symmetric version of the wave equation $u_{tt} = c^2 u_{xx}$ takes the form*

$$u_{tt} = c^2(u_{rr} + 2u_r/r).$$

Show, by putting $v = ru$, that it has a solution of the form

$$v = f(ct - r) + g(ct + r).$$

Problem 1.21. *Let $\xi = x - ct$ and $\eta = x + ct$. Use the chain rule to show that*

$$u_{tt} - c^2 u_{xx} = -4u_{\xi\eta}.$$

Problem 1.22. *Show that the solution of the initial value problem*

$$u_{tt} = c^2 u_{xx}, \qquad u(x, 0) = f(x), \qquad u_t(x, 0) = g(x),$$

satisfies d'Alembert's formula:

$$u(x, t) = \frac{1}{2}\Big[f(x - ct) + f(x + ct)\Big] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(y)\, dy.$$

Chapter 2

Polynomial Approximation in 1d

Our objective is to present the finite element method (FEM) as an approximation technique for solution of differential equations using piecewise polynomials. This chapter is devoted to some necessary mathematical environments and tools, as well as a motivation for the unifying idea of using finite elements: A numerical strategy arising from the need of changing a continuous problem into a discrete one. The continuous problem will have infinitely many unknowns (if one asks for $u(x)$ at every x), and it cannot be solved exactly on a computer. Therefore it has to be approximated by a discrete problem with a finite number of unknowns. The more unknowns we keep, the better the accuracy of the approximation will be, but at a greater computational expense.

2.1 Overture

Below we shall introduce a few standard examples of classical differential equations and some regularity requirements.

Ordinary differential equations (ODEs)
An *initial value problem* (IVP), for instance a model in population dynamics where $u(t)$ is the size of the population at time t, can be written as

$$\dot{u}(t) = \lambda u(t), \quad 0 < t < T, \qquad u(0) = u_0, \tag{2.1.1}$$

where $\dot{u}(t) = \frac{du}{dt}$ and λ is a positive constant. For $u_0 > 0$ this problem has the increasing analytic solution $u(t) = u_0 e^{\lambda \cdot t}$, which blows up as $t \to \infty$.

Generally, we have $\dot{\mathbf{u}}(t) = F(\mathbf{u}(t), t)$, where $\mathbf{u}(t) \in \mathbb{R}^n$ is a time dependent vector in $\mathbb{R}^n$, with $\dot{\mathbf{u}} = \partial \mathbf{u}(t)/\partial t \in \mathbb{R}^n$ being its componentwise derivative with respect to $t \in \mathbb{R}^+$. Thus $\mathbf{u}(t) = [u_1(t), u_2(t), \ldots, u_n(t)]^T$, $\dot{\mathbf{u}}(t) = [\dot{u}_1(t), \dot{u}_2(t), \ldots, \dot{u}_n(t)]^T$ and

$$F : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n.$$

Partial differential equations (PDEs) in bounded domains
Let Ω be a bounded, convex, subset of the Euclidean space $\mathbb{R}^n$. Below is an example of a general *boundary value problem* in $\Omega \subset \mathbb{R}^n$ with the
• *Dirichlet boundary condition*,

$$\begin{cases} -\Delta u(\mathbf{x}) + \mathbf{b} \cdot \nabla u(\mathbf{x}) + \alpha u(\mathbf{x}) = f(\mathbf{x}), & \mathbf{x} \in \Omega \subset \mathbb{R}^n, \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega, \end{cases} \tag{2.1.2}$$

where $\alpha \in \mathbb{R}$, $\mathbf{b} = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$ and $u : \mathbb{R}^n \to \mathbb{R}$ *is a real-valued function* with $\nabla u := \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_n}\right)$, $\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2}$, and

$$\mathbf{b} \cdot \nabla u = b_1 \frac{\partial u}{\partial x_1} + b_2 \frac{\partial u}{\partial x_2} + \ldots + b_n \frac{\partial u}{\partial x_n}.$$

The *heat equation* is an example of an initial boundary value problem with
• *Neumann boundary condition*

$$\begin{cases} \frac{\partial u}{\partial t} = \Delta u, & \mathbf{x} \in \Omega \subset \mathbb{R}^k, \quad t > 0, \\ \frac{\partial u}{\partial \mathbf{n}} = 0, & \mathbf{x} \in \partial\Omega, \quad t > 0, \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \end{cases} \tag{2.1.3}$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_k)$ is the outward unit normal to the boundary $\partial\Omega$ at the point $\mathbf{x} \in \partial\Omega$, and $\partial u/\partial \mathbf{n}$ is the derivative in the direction of $\mathbf{n}$:

$$\frac{\partial u}{\partial \mathbf{n}} = \mathbf{n} \cdot \nabla u. \tag{2.1.4}$$

Regularity requirements for classical solutions

1) $\mathbf{u} \in \mathcal{C}^1$: every component of $\mathbf{u}$ has a continuous first order derivative.

2) $u \in \mathcal{C}^1$: all first order partial derivatives of u are continuous.

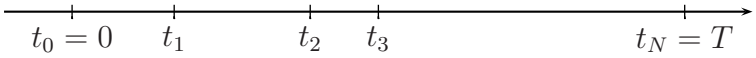
3) $u \in \mathcal{C}^2$: all second order partial derivatives of u are continuous.

4) $u \in \mathcal{C}^1\Big(\mathbb{R}^+; \mathcal{C}^2(\Omega)\Big)$: $\frac{\partial u}{\partial t}$ and $\frac{\partial^2 u}{\partial x_i \partial x_j}$, $i, j = 1, 2, \ldots, n$ are continuous.

Remark 2.1. *Above we mean that:* **u** *in 1) is a vector-valued function of a single variable as in the above example of general dynamical system, whereas u in 2)-4) is a scalar (real-valued) function of several variables.*

• Numerical solutions of (IVP)

Example 2.1. *Explicit (forward) Euler method (a finite difference method). We discretize the IVP* (2.1.1) *with the forward Euler method based on a partition of the interval* $[0, T]$ *into* N *subintervals, and an approximation of*



the derivative by a difference quotient at each subinterval $[t_k, t_{k+1}]$ *by* $\dot{u}(t) \approx \frac{u(t_{k+1}) - u(t_k)}{t_{k+1} - t_k}$. *Then an approximation of* (2.1.1) *is given by*

$$\frac{u(t_{k+1}) - u(t_k)}{t_{k+1} - t_k} = \lambda \cdot u(t_k), \qquad k = 0, \ldots, N-1, \quad \textit{and} \quad u(0) = u_0, \tag{2.1.5}$$

and thus, letting $\Delta t_k = t_{k+1} - t_k$,

$$u(t_{k+1}) = (1 + \lambda \Delta t_k) u(t_k). \tag{2.1.6}$$

Starting with $k = 0$ *and the data* $u(0) = u_0$, *the solution* $u(t_k)$ *would, iteratively, be computed at the subsequent points:* $t_1,\ t_2, \ldots, t_N = T$.
For a uniform partition, where all subintervals have the same length Δt, (2.1.6) *would be of the form*

$$u(t_{k+1}) = (1 + \lambda \Delta t) u(t_k), \quad k = 0, 1, \ldots, N-1. \tag{2.1.7}$$

Iterating we get

$$u(t_{k+1}) = (1 + \lambda \Delta t) u(t_k) = (1 + \lambda \Delta t)^2 u(t_{k-1}) = \ldots = (1 + \lambda \Delta t)^{k+1} u_0. \tag{2.1.8}$$

There are corresponding finite difference methods for PDE's. Our goal, however, is to study the *Galerkin finite element method*. To this approach we need to introduce some basic tools:

Finite dimensional linear space of functions defined on an interval Below we give a list of some examples of finite dimensional linear spaces. Some of these examples are studied in detail in Chapter 3: *the polynomial interpolation in 1D*.

I. $\mathcal{P}^{(q)}(a,b) := \{\text{The space of polynomials in } x \text{ of degree} \leq q,\ a \leq x \leq b\}$. A possible basis for $\mathcal{P}^{(q)}(a,b)$ would be $\{x^j\}_{j=0}^q = \{1, x, x^2, x^3, \ldots, x^q\}$. These are, in general, non-orthogonal polynomials and may be orthogonalized by the Gram-Schmidt procedure. The dimension of $\mathcal{P}^q$ is therefore $q+1$.

II. An example of orthogonal basis functions, on $(0,1)$ or $(-1,1)$, are the *Legendre polynomials*:

$$\tilde{P}_k(x) = \frac{(-1)^k}{k!}\frac{d^k}{dx^k}[x^k(1-x)^k] \quad \text{or} \quad P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n],$$

respectively. The first four Legendre orthogonal polynomials on $(-1,1)$ are:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}, \quad P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x.$$

III. Periodic orthogonal bases on $[0,T]$ are usually represented by trigonometric polynomials given by

$$T^N := \left\{ f(x) \middle| f(x) = \sum_{n=0}^{N} \left[a_n \cos\left(\frac{2\pi}{T} nx\right) + b_n \sin\left(\frac{2\pi}{T} nx\right) \right] \right\}$$

IV. A general form of basis functions on an interval are introduced in Chapter 3: the Lagrange basis $\{\lambda_i\}_{i=0}^q \subset \mathcal{P}^{(q)}(a,b)$ associated to a set of $(q+1)$ distinct points $\xi_0 < \xi_1 < \ldots < \xi_q$ in (a,b) determined by the requirement that

$$\lambda_i(\xi_j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j \end{cases} \qquad \text{or} \qquad \lambda_i(x) = \prod_{j=0,(j\neq i)}^{q} \frac{x - \xi_j}{\xi_i - \xi_j}.$$

A polynomial $P \in \mathcal{P}^{(q)}(a,b)$, that has the value $p_i = P(\xi_i)$ at the nodes $x = \xi_i$ for $i = 0, 1, \ldots, q$, expressed in terms of the corresponding Lagrange basis is then given by

$$P(x) = p_0\lambda_0(x) + p_1\lambda_1(x) + \ldots + p_q\lambda_q(x). \tag{2.1.9}$$

Note that for each node point $x = \xi_i$ we have associated a basis functions $\lambda_i(x)$, $i = 0, 1, \ldots, q$. Thus we have $q + 1$ basis functions.

Remark 2.2. *Our goal is to approximate general functions by polynomials of Lagrange type. Then, for a given function f, the Lagrange coefficients p_i, $0 \le i \le q$, in (2.1.9) will be replaced by $f(\xi_i)$, and $f(x)$ will be approximated by its Lagrange interpolant defined by*

$$f(x) \approx \sum_{i=0}^{q} f(\xi_i)\lambda_i(x) := \pi_q f(x). \tag{2.1.10}$$

We shall illustrate this in the next examples.

Example 2.2. *The linear Lagrange basis functions, $q = 1$, are given by (see Fig. 2.1.)*

$$\lambda_0(x) = \frac{\xi_1 - x}{\xi_1 - \xi_0} \qquad \textit{and} \qquad \lambda_1(x) = \frac{x - \xi_0}{\xi_1 - \xi_0}. \tag{2.1.11}$$

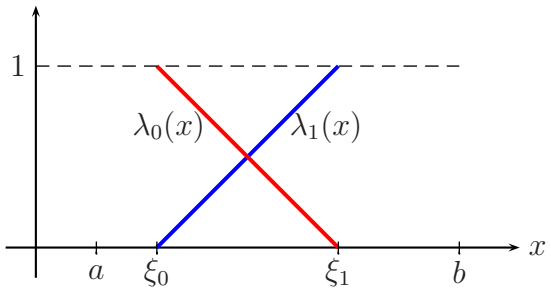


Figure 2.1: Linear Lagrange basis functions for $q = 1$.

Example 2.3. *A typical application of the Lagrange basis is in finding a polynomial interpolant $\pi_q f \in \mathcal{P}^q(a,b)$ of a continuous function $f(x)$ on an interval $[a,b]$. The procedure is as follows:*

Choose distinct interpolation nodes $\xi_i : a = \xi_0 < \xi_1 < \ldots < \xi_q = b$ and let $\pi_q f(\xi_i) = f(\xi_i)$. Then $\pi_q f \in \mathcal{P}^{(q)}(a,b)$, defined as the sum in (2.1.10), interpolates $f(x)$ at the nodes $\{\xi_i\}$, $i = 0, \ldots, q$, and using Lagrange's formula (2.1.9), with $p_i = f(\xi_i)$, $i = 0, 1, \ldots, q$, yields

$$\pi_q f(x) = f(\xi_0)\lambda_0(x) + f(\xi_1)\lambda_1(x) + \ldots + f(\xi_q)\lambda_q(x), \quad x \in [a,b].$$

For a linear interpolant, i.e. $q = 1$, we only need 2 nodes and 2 basis functions. Choosing $\xi_0 = a$ and $\xi_1 = b$ in (2.1.11), we get the linear interpolant

$$\pi_1 f(x) = f(a)\lambda_0(x) + f(b)\lambda_1(x),$$

where

$$\lambda_0(x) = \frac{b-x}{b-a} \quad \text{and} \quad \lambda_1(x) = \frac{x-a}{b-a},$$

i.e.,

$$\pi_1 f(x) = f(a)\frac{b-x}{b-a} + f(b)\frac{x-a}{b-a}.$$

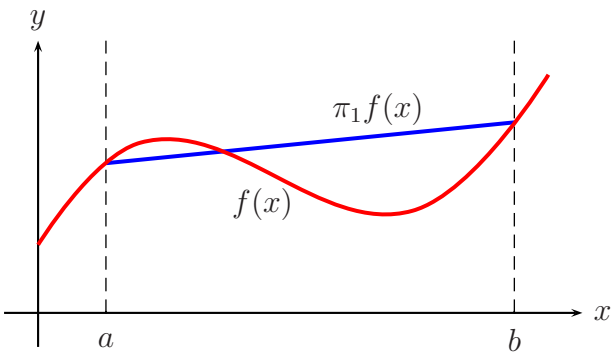


Figure 2.2: The linear interpolant $\pi_1 f(x)$ on a single interval.

V. We shall frequently use the space of *continuous piecewise polynomials* on a partition of an interval into a collection of subintervals. For example $\mathcal{T}_h : 0 = x_0 < x_1 < \ldots < x_M < x_{M+1} = 1$, with $h_j = x_j - x_{j-1}$ and $j = 1, \ldots, M+1$, is a partition of $[0,1]$ into $M+1$ subintervals.

Let $V_h^{(q)}$ denote the space of all continuous piecewise polynomial functions of degree $\leq q$ on $\mathcal{T}_h$. Let also

$$V_{0,h}^{(q)} = \{v : v \in V_h^{(q)}, \quad v(0) = v(1) = 0\}.$$

Our motivation in introducing these function spaces is due to the fact that these are function spaces, adequate in the numerical study of boundary value problems, using finite element methods for approximating solutions with piecewise polynomials.

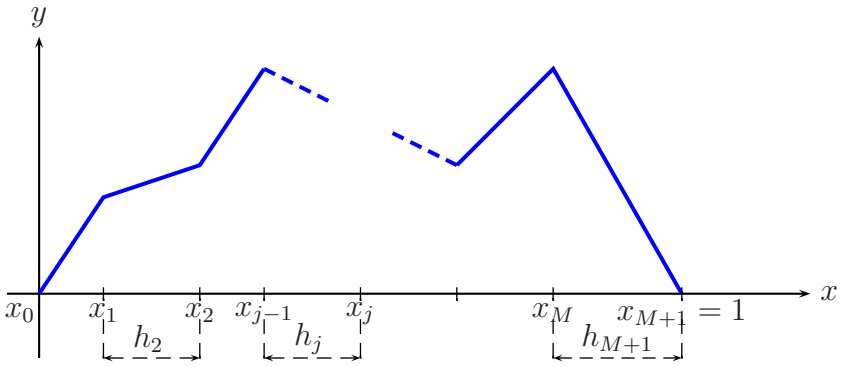


Figure 2.3: An example of a function in $V_{0,h}^{(1)}$.

The standard basis for piecewise linears: $V_h := V_h^{(1)}$ is given by the so called *hat-functions* $\varphi_j(x)$ with the property that $\varphi_j(x)$ is a piecewise linear function such that $\varphi_j(x_i) = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad \text{i.e.} \quad \varphi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h_j} & x_{j-1} \leq x \leq x_j \\ \frac{x_{j+1} - x}{h_{j+1}} & x_j \leq x \leq x_{j+1} \\ 0 & x \notin [x_{j-1}, x_{j+1}], \end{cases}$$

with obvious modifications for $j = 0$ and $j = M + 1$ (see Remark 2.3 and Figure 2.6).

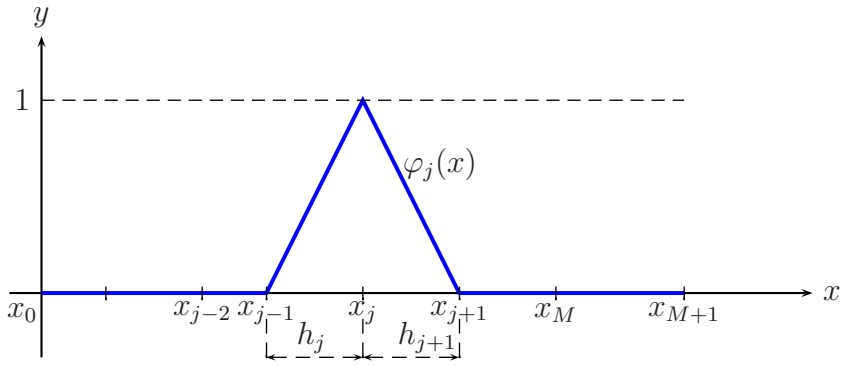


Figure 2.4: A general piecewise linear basis function $\varphi_j(x)$.

Vector spaces

To establish a framework we shall introduce some basic mathematical concepts:

Definition 2.1. *A set V of functions or vectors is called a linear space, or a vector space, if for all $u, v, w \in V$ and all $\alpha, \beta \in \mathbb{R}$ (real numbers), we have*

$$\begin{aligned}
&(i) && u + \alpha v \in V, \\
&(ii) && (u+v)+w = u+(v+w), \\
&(iii) && u+v = v+u, \\
&(iv) && \exists\ 0 \in V, \quad \textit{such that} \quad u+0 = 0+u = u, \\
&(v) && \forall\, u \in V,\ \exists\,(-u) \in V, \quad \textit{such that} \quad u+(-u) = 0, \\
&(vi) && (\alpha+\beta)u = \alpha u + \beta u, \\
&(vii) && \alpha(u+v) = \alpha u + \alpha v, \\
&(viii) && \alpha(\beta u) = (\alpha\beta)u.
\end{aligned} \tag{2.1.12}$$

Observe that (iii) and (i), with $\alpha = 1$ and $v = (-u)$ implies that 0 (zero vector) is an element of every vector space.

Definition 2.2 (Scalar product)**.** *A scalar product (inner product) is a real valued operator on $V \times V$, viz $\langle u, v\rangle : V \times V \to \mathbb{R}$ such that for all u, v, $w \in V$*

and all $\alpha \in \mathbb{R}$,

$$\begin{array}{ll} (i) & \langle u, v\rangle = \langle v, u\rangle, \qquad \textit{(symmetry)} \\ (ii) & \langle u + \alpha v, w\rangle = \langle u, w\rangle + \alpha\langle v, w\rangle, \quad \textit{(bi-linearity)} \\ (iii) & \langle v, v\rangle \geq 0, \quad \forall v \in V, \quad \textit{(positivity)} \\ (iv) & \langle v, v\rangle = 0, \Longleftrightarrow v = 0. \end{array} \tag{2.1.13}$$

Definition 2.3. *A vector space* W *is called an inner product space if* W *is associated with a scalar product* $\langle \cdot, \cdot \rangle$*, defined on* $W \times W$*.*

The function spaces $C([0,T])$, $C^k([0,T])$, $\mathcal{P}^q$, T^q and $V_h^{(q)}$ are examples of inner product spaces associated with the usual scalar product defined by

$$\langle u, v\rangle = \int_0^T u(x)v(x)dx. \tag{2.1.14}$$

Definition 2.4 (Orthogonality)**.** *Two, real-valued, functions* $u(x)$ *and* $v(x)$ *are called orthogonal if* $\langle u, v\rangle = 0$*. This orthogonality is also denoted by* $u \perp v$*.*

Definition 2.5 (Norm)**.** *If* $u \in V$ *then the norm of* u*, or the length of* u*, associated with the above scalar product is defined by*

$$\|u\| = \sqrt{\langle u, u\rangle} = \langle u, u\rangle^{1/2} = \Big(\int_0^T |u(x)|^2 dx\Big)^{1/2}. \tag{2.1.15}$$

This norm is known as the L_2*-norm of* $u(x)$*. There are other norms that we will introduce later on.*

We also recall one of the most useful tools that we shall frequently use throughout this note: *The Cauchy-Schwarz inequality*,

$$|\langle u, v\rangle| \leq \|u\|\|v\|. \tag{2.1.16}$$

A simple proof of (2.1.16) is given by using

$$\langle u - av, u - av\rangle \geq 0, \qquad \text{with} \qquad a = \langle u, v\rangle / \|v\|^2.$$

Then by the definition of the L_2-norm and the symmetry property of the scalar product we get

$$0 \leq \langle u - av, u - av\rangle = \|u\|^2 - 2a\langle u, v\rangle + a^2\|v\|^2.$$

Setting $a = \langle u, v\rangle / \|v\|^2$ and rearranging the terms we get

$$0 \le \|u\|^2 - \frac{\langle u, v\rangle^2}{\|v\|^4}\|v\|^2, \quad \text{and consequently} \quad \frac{\langle u, v\rangle^2}{\|v\|^2} \le \|u\|^2,$$

which yields the desired result.

• A Galerkin method for (IVP)
We multiply the initial value problem (2.1.1) with test functions v in a certain vector space V and integrate over $[0, T]$,

$$\int_0^T \dot{u}(t)v(t)\,dt = \lambda \int_0^T u(t)v(t)\,dt, \qquad \forall v \in V, \tag{2.1.17}$$

or equivalently

$$\int_0^T \Big(\dot{u}(t) - \lambda\, u(t)\Big)v(t)dt = 0, \qquad \forall v(t) \in V, \tag{2.1.18}$$

i.e.

$$\Big(\dot{u}(t) - \lambda\, u(t)\Big) \perp v(t), \quad \forall v(t) \in V. \tag{2.1.19}$$

We refer to (2.1.17) as the *variational problem* for (2.1.1).

For the variational problem (2.1.17) it is natural to seek a solution in $C([0, T])$, or in

$$V := H^1(0, T) := \Big\{ f : \int_0^T \Big(f(t)^2 + \dot{f}(t)^2\Big)\,dt < \infty \Big\}.$$

H^1 is consisting of all functions in $L_2(0, T)$ having also their derivatives in $L_2(0, T)$.

Definition 2.6. *If w is an approximation of u in the variational problem (2.1.17), then $\mathcal{R}\Big(w(t)\Big) := \dot{w}(t) - \lambda w(t)$ is called the residual error of $w(t)$*

In general for an approximate solution w we have $\dot{w}(t) - \lambda w(t) \neq 0$, otherwise w and u would satisfy the same equation and by uniqueness we would get the exact solution ($w = u$). Our requirement is instead that w satisfies equation (2.1.1) in average, or in other words we require that w satisfies (2.1.19):

$$\mathcal{R}\Big(w(t)\Big) \perp v(t), \quad \forall v(t) \in V. \tag{2.1.20}$$

We look for an approximate solution $U(t)$ in a finite dimensional subspace of V, say $\mathcal{P}^{(q)}$. More specifically, we want to look at an *approximate solution* $U(t)$, called a *trial function* for (2.1.1), in the space of polynomials of degree $\leq q$:

$$\mathcal{P}^{(q)} = \{U : U(t) = \xi_0 + \xi_1 t + \xi_2 t^2 + \ldots + \xi_q t^q\}. \tag{2.1.21}$$

Hence, to determine $U(t)$ we need to determine the coefficients $\xi_0, \xi_1, \ldots \xi_q$. We refer to $V^{(q)}$ as the *space of trial functions*. Note that $u(0) = u_0$ is given and therefore we may take $U(0) = \xi_0 = u_0$. It remains to find the real numbers $\xi_1, \ldots, \xi_q$. These are coefficients of the q linearly independent monomials t, t^2, $\ldots$, t^q. To this approach we define the *test function space*:

$$\mathcal{P}_0^{(q)} = \{v \in \mathcal{P}^{(q)} : v(0) = 0\}, \tag{2.1.22}$$

in other words v can be written as $v(t) = c_1 t + c_2 t^2 + \ldots + c_q t^q$. Note that

$$\mathcal{P}_0^{(q)} = span[t, t^2, \ldots, t^q]. \tag{2.1.23}$$

For an approximate solution U we require its residual $R(U)$ to satisfy the orthogonality condition (2.1.20):

$$\mathcal{R}\Big(U(t)\Big) \perp v(t), \qquad \forall v(t) \in \mathcal{P}_0^{(q)}.$$

Thus the *Galerkin finite element method for* (2.1.1) *is formulated* as follows: Given $u(0) = u_0$, find the approximate solution $U(t) \in \mathcal{P}^{(q)}$, for (2.1.1) such that (for simplicity we put $T \equiv 1$)

$$\int_0^1 \mathcal{R}\Big(U(t)\Big)v(t)dt = \int_0^1 (\dot{U}(t) - \lambda\, U(t))v(t)dt = 0,\ \forall v(t) \in \mathcal{P}_0^{(q)}. \tag{2.1.24}$$

Formally this can be obtained writing a *wrong!* equation by replacing u by $U \in \mathcal{P}^{(q)}$ in (2.1.1),

$$\begin{cases} \dot{U}(t) = \lambda\, U(t), & 0 < t < 1 \\ U(0) = u_0, \end{cases} \tag{2.1.25}$$

then, multiplying (2.1.25) by a function $v(t) \in \mathcal{P}_0^{(q)}$ from the test function space and integrating over $[0, 1]$.

Now since $U \in \mathcal{P}^{(q)}$, we can write $U(t) = u_0 + \sum_{j=1}^{q} \xi_j t^j$, then $\dot{U}(t) = \sum_{j=1}^{q} j\xi_j t^{j-1}$. Further we have $\mathcal{P}_0^{(q)}$ is spanned by $v_i(t) = t^i, i = 1, 2, \ldots, q$, it suffices to use these as test functions. Inserting these representations for U, $\dot{U}$ and $v = v_i$, $i = 1, 2, \ldots, q$ in (2.1.24) we get

$$\int_0^1 \Big(\sum_{j=1}^{q} j\xi_j t^{j-1} - \lambda u_0 - \lambda \sum_{j=1}^{q} \xi_j t^j \Big) \cdot t^i dt = 0, \quad i = 1, 2, \ldots, q, \tag{2.1.26}$$

which can be rewritten as

$$\int_0^1 \Big(\sum_{j=1}^{q} (j\xi_j t^{i+j-1} - \lambda\, \xi_j t^{i+j}) \Big) dt = \lambda u_0 \int_0^1 t^i dt, \quad i = 1, 2, \ldots, q. \tag{2.1.27}$$

Performing the integration (ξ_j:s are constants independent of t) we get

$$\sum_{j=1}^{q} \xi_j \Big[j \cdot \frac{t^{i+j}}{i+j} - \lambda \frac{t^{i+j+1}}{i+j+1} \Big]_{t=0}^{t=1} = \Big[\lambda \cdot u_0 \frac{t^{i+1}}{i+1} \Big]_{t=0}^{t=1}, \tag{2.1.28}$$

or equivalently

$$\sum_{j=1}^{q} \Big(\frac{j}{i+j} - \frac{\lambda}{i+j+1} \Big) \xi_j = \frac{\lambda}{i+1} \cdot u_0 \quad i = 1, 2, \ldots, q, \tag{2.1.29}$$

which is a linear system of equations with q equations and q unknowns (ξ_1, ξ_2, …, ξ_q); given in the matrix form as

$$\mathcal{A}\Xi = \mathbf{b}, \quad \text{with} \quad \mathcal{A} = (a_{ij}), \quad \Xi = (\xi_j)_{j=1}^{q}, \quad \text{and} \quad \mathbf{b} = (b_i)_{i=1}^{q}. \tag{2.1.30}$$

But the matrix $\mathcal{A}$ although invertible, is *ill-conditioned*, mostly because $\{t^i\}_{i=1}^{q}$ does not form an orthogonal basis. We observe that for large i and j the last two rows (columns) of $\mathcal{A}$ given by $a_{ij} = \dfrac{j}{i+j} - \dfrac{\lambda}{i+j+1}$, are very close to each other resulting in very small values for the determinant of $\mathcal{A}$. If we insist to use polynomial basis up to certain order, then instead of monomials, the use of Legendre orthogonal polynomials (see Chapter 3) would yield a diagonal (sparse) coefficient matrix and make the problem well conditioned. This however, is a rather tedious task. A better approach would be through the use of piecewise polynomial approximations (see Chapter 5) on

a partition of of $[0, T]$ into subintervals, where we use *low order* polynomial approximations on each subinterval.

Galerkin's method and orthogonal projection: L_2-projection

Let $\mathbf{u} = (u_1, u_2, u_3) \in \mathbb{R}^3$ and assume that for some reasons we only have u_1 and u_2 available. Letting $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$, the objective, then is to find $\mathbf{U} \in \{x : x_3 = 0\}$, such that $(\mathbf{u} - \mathbf{U})$ is as small as possible. Obviously in this case $\mathbf{U} = (u_1, u_2, 0)$ and we have $(\mathbf{u} - \mathbf{U}) \perp \mathbf{U}$, $\forall \mathbf{U}$ in the x_1x_2-plane, see Figure 2.5.

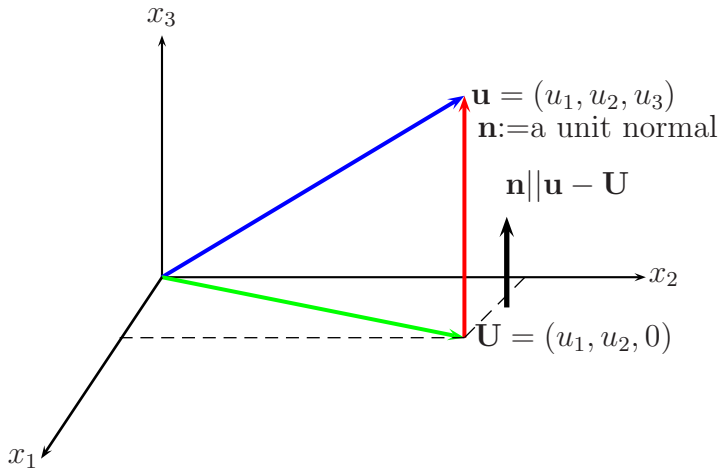


Figure 2.5: Example of a projection onto $\mathbb{R}^2$.

The L_2-projection onto a space of polynomials

A polynomial πf interpolating a given function $f(x)$ on an interval (a, b) agrees with point values of f at a certain discrete set of points $x_i \in (a, b)$: $\pi f(x_i) = f(x_i)$, $i = 1, \ldots, n$, for some integer n. Like Riemann sum for the integrals, this concept can be generalized to determine a polynomial πf so that certain averages (a kind of weighting) agree. These could include the usual average of f over $[a, b]$ defined by,

$$\frac{1}{b-a} \int_a^b f(x)\, dx,$$

or a *generalized average* of f with respect to a *weight function* w defined by

$$(f, w) = \int_a^b f(x)w(x)\,dx.$$

Definition 2.7. *The orthogonal projection, or L_2-projection, of the function f onto $\mathcal{P}^q(a,b)$ is the polynomial $Pf \in \mathcal{P}^q(a,b)$ such that*

$$(f, w) = (Pf, w) \iff (f - Pf, w) = 0 \quad \textit{for all } w \in \mathcal{P}^q(a,b). \tag{2.1.31}$$

We see that Pf is defined so that certain average values of Pf are the same as those of f. By the construction (2.1.31) is equivalent to a $(q+1) \times (q+1)$ system of equations.

We want to show that:

Lemma 2.1. *(i) Pf is uniquely defined by* (2.1.31)*.*
(ii) Pf is the best approximation of f in $\mathcal{P}^q(a,b)$ in the $L_2(a,b)$-norm, i.e.

$$||f - Pf||_{L_2(a,b)} \leq ||f - v||_{L_2(a,b)}, \qquad \textit{for all } v \in \mathcal{P}^q(a,b). \tag{2.1.32}$$

Proof. (i) Suppose that $P_1 f$ and $P_2 f$ are two polynomials in $\mathcal{P}^q(a,b)$ such that

$$(f - P_1 f, w) = 0 \quad \text{and} \quad (f - P_2 f, w) = 0 \qquad \text{for all } w \in \mathcal{P}^q(a,b).$$

Subtracting the two relations we conclude that

$$(P_2 f - P_1 f, w) = 0, \qquad \text{for all } w \in \mathcal{P}^q(a,b).$$

Now choosing $w = P_2 f - P_1 f$ we get

$$\int_a^b |P_2 f - P_1 f|^2\,dx = 0,$$

which yields $P_1 f = P_2 f$ since $|P_2 f - P_1 f|$ is a non-negative continuous functions.
(ii) Using Cauchy-Schwarz' inequality it follows that for all $v \in \mathcal{P}^q(a,b)$, since $(v - Pf) \in \mathcal{P}^q(a,b)$ and $(f - Pf) \perp \mathcal{P}^q(a,b)$,

$$\begin{aligned} ||f - Pf||^2_{L_2(a,b)} &= (f - Pf, f - Pf) = (f - Pf, f - v + v - Pf) \\ &= (f - Pf, f - v) + (f - Pf, v - Pf) \\ &= (f - Pf, f - v) \leq ||f - Pf||_{L_2(a,b)} ||f - v||_{L_2(a,b)}, \end{aligned}$$

which gives the desired result. □

• A Galerkin method for (BVP)

We consider the Galerkin method for the following stationary ($\dot{u} = du/dt = 0$) heat equation in one dimension:

$$\begin{cases} -\frac{d}{dx}\Big(c(x)\cdot\frac{d}{dx}u(x)\Big) = f(x), & 0 < x < 1; \\ u(0) = u(1) = 0. \end{cases} \tag{2.1.33}$$

Let $c(x) = 1$, then we have

$$-u''(x) = f(x), \quad 0 < x < 1; \qquad u(0) = u(1) = 0. \tag{2.1.34}$$

Now let $\mathcal{T}_h : 0 = x_0 < x_1 < \ldots < x_M < x_{M+1} = 1$ be a partition of the interval $(0,1)$ into the subintervals $I_j = (x_{j-1}, x_j)$, with length $|I_j| = h_j = x_j - x_{j-1}$, $j = 1, 2, \ldots, M+1$. We define the finite dimensional space

$$V_{0,h}^{(1)} = \{v \in \mathcal{C}(0,1) : v \,\text{is a piecewise linear function on}\, \mathcal{T}_h,\ v(0) = v(1) = 0\},$$

with the basis functions $\{\varphi_j\}_{j=1}^{M}$. Due to the fact that u is known at the boundary points 0 and 1; it is not necessary to supply test functions corresponding to the values at $x_0 = 0$ and $x_{M+1} = 1$. However, in the case of given non-homogeneous boundary data $u(0) = u_0 \neq 0$ and/or $u(1) = u_1 \neq 0$, to represent the trial function, one uses the basis functions to all internal nodes as well as those corresponding to the non-homogeneous data (i.e. at $x = 0$ and/or $x = 1$).

Remark 2.3. *If the Dirichlet boundary condition is given at only one of the boundary points; say $x_0 = 0$ and the other one satisfies, e.g. a Neumann condition as*

$$-u''(x) = f(x), \quad 0 < x < 1; \qquad u(0) = b_0, \quad u'(1) = b_1, \tag{2.1.35}$$

then the corresponding test function φ_0 will be unnecessary (no matter whether $b_0 = 0$ or $b_0 \neq 0$), whereas one needs to provide the half-base function φ_{M+1} at $x_{M+1} = 1$ (dashed in Figure 2.6). Again, φ_0 participates in representing the trial function U.

Now the *Galerkin method* for problem (2.1.34), is based on the variation formulation, where we multiply (2.1.34) by a test function in $V_0 = H_0^1$ and

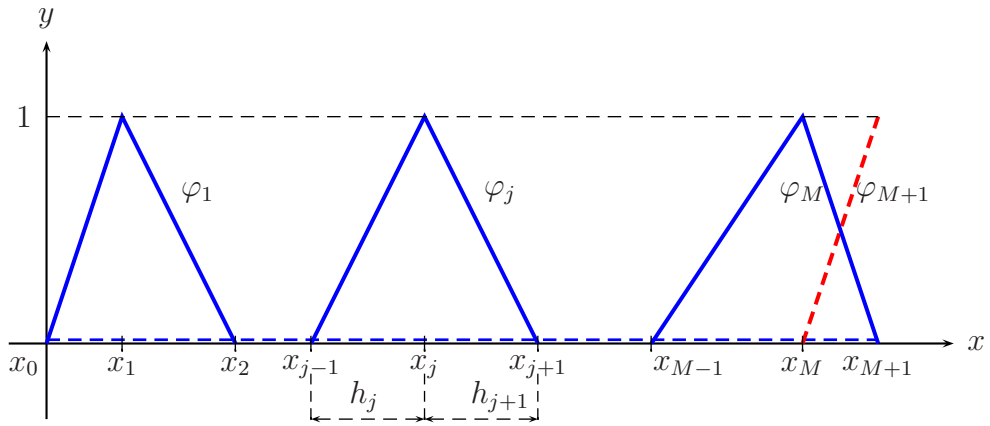


Figure 2.6: Piecewise linear basis functions

integrate over $[0,1)$: Find $u(x) \in V$ such that

$$\int_0^1 (-u''(x) - f(x))v(x)dx = 0, \qquad \forall v(x) \in V. \tag{2.1.36}$$

we perform partial integration in (2.1.36) to obtain

$$-\int_0^1 u''(x)v(x)dx = \int_0^1 u'(x)v'(x)dx - [u'(x)v(x)]_0^1 \tag{2.1.37}$$

and since $v(x) \in V_0$; $v(0) = v(1) = 0$, we get

$$-\int_0^1 u''(x)v(x)dx = \int_0^1 u'(x)v'(x)\,dx. \tag{2.1.38}$$

Thus the *variational formulation* is: Find $u \in V_0$ such that

$$\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)dx, \qquad \forall v \in V_0 \tag{2.1.39}$$

This is a justification of the *finite element formulation*:

The Galerkin finite element method (FEM) for the problem (2.1.34): Find $U(x) \in V_h^0$ such that

$$\int_0^1 U'(x)v'(x)\,dx = \int_0^1 f(x)v(x)dx, \qquad \forall v(x) \in V_h^0. \tag{2.1.40}$$

We shall determine $\xi_j = U(x_j)$ *the approximate* values of $u(x)$ at the node points x_j, $1 \le j \le M$. To this end using basis functions $\varphi_j(x)$, we may write

$$U(x) = \sum_{j=1}^{M} \xi_j \, \varphi_j(x) \quad \text{which implies that} \quad U'(x) = \sum_{j=1}^{M} \xi_j \varphi'(x). \quad (2.1.41)$$

Thus we can write (2.1.40) as

$$\sum_{j=1}^{M} \xi_j \int_0^1 \varphi_j'(x)\, v'(x) dx = \int_0^1 f(x) v(x) dx, \qquad \forall v(x) \in V_h^0. \quad (2.1.42)$$

Since every $v(x) \in V_h^0$ is a linear combination of the basis functions $\varphi_i(x)$, it suffices to try with $v(x) = \varphi_i(x)$, for $i = 1, 2, \ldots, M$: That is, to find ξ_j (constants), $1 \le j \le M$ such that

$$\sum_{j=1}^{M} \Big(\int_0^1 \varphi_i'(x) \varphi_j'(x) dx \Big) \xi_j = \int_0^1 f(x) \varphi_i(x) dx, \quad i = 1, 2, \ldots, M. \quad (2.1.43)$$

This equation can be written in the equivalent matrix form as

$$\mathbf{A}\xi = \mathbf{b}. \quad (2.1.44)$$

Here $\mathbf{A}$ is called the *stiffness matrix* and $\mathbf{b}$ the *load vector*:

$$\mathbf{A} = \{a_{ij}\}_{i,j=1}^{M}, \quad a_{ij} = \int_0^1 \varphi_i'(x) \varphi_j'(x) dx, \quad (2.1.45)$$

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \ldots \\ b_M \end{pmatrix}, \text{ with } \quad b_i = \int_0^1 f(x) \varphi_i(x) dx, \text{ and } \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \ldots \\ \xi_M \end{pmatrix}. \quad (2.1.46)$$

To compute the entries a_{ij} of the stiffness matrix $\mathbf{A}$, first we need to determine $\varphi_i'(x)$. Note that

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h_i} & x_{i-1} \le x \le x_i \\ \frac{x_{i+1} - x}{h_{i+1}} & x_i \le x \le x_{i+1} \\ 0 & \text{else} \end{cases} \quad \Longrightarrow \quad \varphi_i'(x) = \begin{cases} \frac{1}{h_i} & x_{i-1} < x < x_i \\ -\frac{1}{h_{i+1}} & x_i < x < x_{i+1} \\ 0 & \text{else} \end{cases}$$

Stiffness matrix A:
If $|i-j|>1$, then φ_i and φ_j have disjoint support, see Figure 2.7., and evidently we have

$$a_{ij}=\int_0^1 \varphi_i'(x)\varphi_j'(x)dx=0.$$

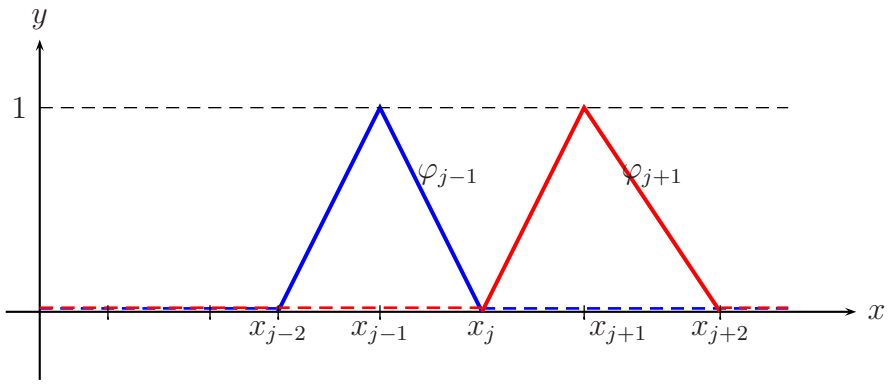


Figure 2.7: φ_{j-1} and φ_{j+1}.

As for $i=j$: we have that

$$a_{ii}=\int_{x_{i-1}}^{x_i}\Big(\frac{1}{h_i}\Big)^2dx+\int_{x_i}^{x_{i+1}}\Big(-\frac{1}{h_{i+1}}\Big)^2dx=\frac{\overbrace{x_i-x_{i-1}}^{h_i}}{h_i^2}+\frac{\overbrace{x_{i+1}-x_i}^{h_{i+1}}}{h_{i+1}^2}=\frac{1}{h_i}+\frac{1}{h_{i+1}}.$$

It remains to compute a_{ij} for the case of (applicable!) $j=i\pm1$: A straightforward calculation (see the fig below) yields

$$a_{i,i+1}=\int_{x_i}^{x_{i+1}}\Big(-\frac{1}{h_{i+1}}\Big)\cdot\frac{1}{h_{i+1}}dx=-\frac{x_{i+1}-x_i}{h_{i+1}^2}=-\frac{1}{h_{i+1}}. \tag{2.1.47}$$

Obviously $a_{i+1,i}=a_{i,i+1}=-\frac{1}{h_{i+1}}$.

To *summarize*, we have

$$\begin{cases} a_{ij}=0, & \text{if } |i-j|>1,\\ a_{ii}=\frac{1}{h_i}+\frac{1}{h_{i+1}}, & i=1,2,\ldots,M,\\ a_{i-1,i}=a_{i,i-1}=-\frac{1}{h_i}, & i=2,3,\ldots,M. \end{cases} \tag{2.1.48}$$

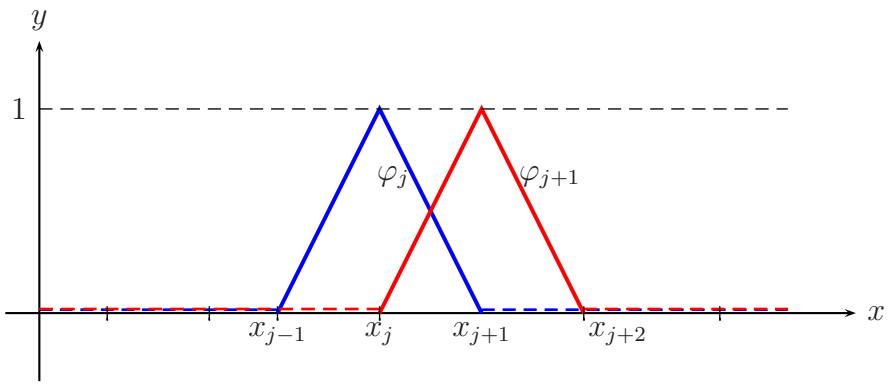


Figure 2.8: φ_j and φ_{j+1}.

By symmetry $a_{ij} = a_{ji}$, and we finally have the stiffness matrix for the stationary heat conduction as:

$$\mathbf{A} = \begin{bmatrix} \frac{1}{h_1} + \frac{1}{h_2} & -\frac{1}{h_2} & 0 & \dots & 0 \\ -\frac{1}{h_2} & \frac{1}{h_2} + \frac{1}{h_3} & -\frac{1}{h_3} & 0 & 0 \\ 0 & \dots & \dots & \dots & 0 \\ \dots & 0 & \dots & \dots & -\frac{1}{h_M} \\ 0 & \dots & 0 & -\frac{1}{h_M} & \frac{1}{h_M} + \frac{1}{h_{M+1}} \end{bmatrix}. \tag{2.1.49}$$

With a *uniform mesh*, i.e. $h_i = h$ we get that

$$\mathbf{A}_{unif} = \frac{1}{h} \cdot \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \dots \\ 0 & -1 & 2 & -1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix}. \tag{2.1.50}$$

As for the components of the load vector $\mathbf{b}$ we have

$$b_i = \int_0^1 f(x)\varphi_i(x)\,dx = \int_{x_i-1}^{x_i} f(x)\frac{x - x_{i-1}}{h_i}\,dx + \int_{x_i}^{x_{i+1}} f(x)\frac{x_{i+1} - x}{h_{i+1}}\,dx.$$

• **A finite difference approach** To illustrate a finite difference approach we reconsider the stationary heat equation (2.1.34):

$$-u''(x) = f(x), \qquad 0 < x < 1; \tag{2.1.51}$$

and motivate its boundary conditions. The equation (2.1.51) is linear in the unknown function u, with inhomogeneous source term f. There is some arbitrariness left in the problem, because any combination $C + Dx$ could be added to any solution. The sum would constitute another solution, since the second derivative of $C + Dx$ contributes nothing. Therefore the uncertainty left by these two arbitrary constants C and D will be removed by adding a *boundary condition* at each end point of the interval

$$u(0) = 0, \qquad u(1) = 0. \tag{2.1.52}$$

The result is a *two-point boundary-value problem*, describing not a transient but a steady-state phenomenon–the temperature distribution in a rod, for example with ends fixed at 0° and with a heat source $f(x)$.

As our goal is to solve a discrete problem, we cannot accept more than a finite amount of information about f, say its values at equally spaced points $x_1 = h$, $x_2 = 2h, \ldots, x_n = nh$. And what we compute will be approximate values $u_1, u_2, \ldots, u_n$ for the true solution u at these same points. At the end points $x_0 = 0$ and $x_{n+1} = 1 = (n+1)h$, we are already given the correct boundary values $u_0 = 0$, $u_{n+1} = 0$.

The first question is: How do we replace the derivative d^2u/dx^2? Since every derivative is a limit of difference quotients, it can be approximated by stopping at a finite step size, and not permitting h (or Δx) to approach zero. For du/dx there are several alternatives:

$$\frac{du}{dx} \approx \frac{u(x+h) - u(x)}{h} \quad \text{or} \quad \frac{u(x) - u(x-h)}{h} \quad \text{or} \quad \frac{u(x+h) - u(x-h)}{2h}.$$

The last, because it is symmetric about x, is the most accurate. For the second derivative we can write

$$\frac{d^2u}{dx^2} \approx \frac{u'(x) - u'(x-h)}{h}. \tag{2.1.53}$$

Replacing the approximations $u'(x) \approx \frac{u(x+h)-u(x)}{h}$ and $u'(x-h) \approx \frac{u(x)-u(x-h)}{h}$

in (2.1.53) we get

$$\begin{aligned}\frac{d^2u}{dx^2} &\approx \frac{(u(x+h)-u(x))/h-(u(x)-u(x-h))/h}{h} \\ &= \frac{u(x+h)-2u(x)+u(x-h)}{h^2},\end{aligned} \tag{2.1.54}$$

which has the merit of being symmetric about x. To repeat the right side approaches the true value of d^2u/dx^2 as $h \to 0$, but we have to stop at a positive h.

At a typical mesh point $x_j = jh$, the differential equation $-d^2u/dx^2 = f(x)$ is now replaced by this discrete analogue (2.1.54); after multiplying by h^2,

$$-u_{j+1} + 2u_j - u_{j-1} = h^2 f(jh). \tag{2.1.55}$$

There are n equations of exactly this form, for every value $j = 1, 2, \ldots, n$. The first and last equations include the expressions u_0 and u_{n+1}, which are not unknowns. Their values are the boundary conditions, and they are shifted to the right hand side of the equation and contribute to the inhomogeneous terms (or at least, they might, if they were not known to be equal zero). It is easy to understand (2.1.55) as a steady-state equation, in which the flows $(u_j - u_{j+1})$ coming from the right and $(u_j - u_{j-1})$ coming from the left are balanced by the source $h^2 f(jh)$ at the center.

The structure of the n equations (2.1.55) can be better visualized in matrix form $Au = b$ viz

$$\begin{bmatrix} 2 & -1 & 0 & \ldots & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & \ldots \\ 0 & -1 & 2 & -1 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ \ldots & \ldots & 0 & -1 & 2 & -1 \\ 0 & \ldots & \ldots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} = h^2 \begin{bmatrix} f(h) \\ f(2h) \\ f(3h) \\ \cdot \\ \cdot \\ f(nh) \end{bmatrix}, \tag{2.1.56}$$

which, once again, gives the structure of our uniform stiffness matrix A_{unif} given in (2.1.50).

So we conclude that, for this problem, the finite element and finite difference approximations are two equivalent approaches.

Remark 2.4. *Unlike the matrix $\mathcal{A}$ for polynomial approximation of IVP in* (2.1.29)*,* $\mathbf{A}$ *has a more desirable structure, e.g.* $\mathbf{A}$ *is a sparse, tridiagonal and symmetric matrix. This is due to the fact that the basis functions* $\{\varphi_j\}_{j=1}^M$ *are nearly orthogonal. The most important property of* $\mathbf{A}$ *is that it is positive definite.*

Definition 2.8. *A* $M \times M$ *matrix* $\mathbf{A}$ *is called positive definite if*

$$\forall \eta \in R^M,\ \eta \neq 0,\ \eta^T \mathbf{A} \eta > 0, \qquad i.e. \qquad \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j > 0. \tag{2.1.57}$$

We shall use the positive definiteness of $\mathbf{A}$ to argue that (2.1.44) is uniquely solvable. To this approach we prove the following well-known result:

Proposition 2.1. *If a square matrix* $\mathbf{A}$ *is positive definite then* $\mathbf{A}$ *is invertible and hence* $\mathbf{A}\xi = \mathbf{b}$ *has a unique solution.*

Proof. Suppose $\mathbf{Ax} = \mathbf{0}$ then $\mathbf{x}^T\mathbf{Ax} = \mathbf{0}$, and since $\mathbf{A}$ is positive definite, then $\mathbf{x} \equiv \mathbf{0}$. Thus $\mathbf{A}$ has full range and we conclude that $\mathbf{A}$ is invertible. Since $\mathbf{A}$ is invertible $\mathbf{A}\xi = \mathbf{b}$ has a unique solution: $\xi = \mathbf{A}^{-1}\mathbf{b}$. □

Note however, that it is a bad idea to invert a matrix to solve the linear system of equations. Finally we illustrate an example of the positive-definiteness argument for $\mathbf{A}_{unif}$.

Example 2.4. *Assume* $M = 2$ *and let* $\eta(x, y) = \begin{pmatrix} x \\ y \end{pmatrix}$*, then*

$$\begin{aligned} \eta^T \mathbf{A}_{unif} \eta &= (x, y) \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = (x, y) \begin{pmatrix} 2x - y \\ -x + 2y \end{pmatrix} \\ &= 2x^2 - xy - xy + 2y^2 = x^2 + y^2 + x^2 - 2xy + y^2 \\ &= x^2 + y^2 + (x - y)^2 \geq 0. \end{aligned} \tag{2.1.58}$$

Thus $\mathbf{A}_{unif}$ *is positive definite.*

Summary: Roughly speaking, a systematic procedure for approximate solution for a differential equation would involve the following steps:

1. We need to approximate functions by polynomials agreeing with the functional values at certain points (nodes). This is the matter of *Interpolation techniques* which we shall introduce in Chapter 3.

2. In the final system of equations: $\mathbf{A}\xi = \mathbf{b}$, entries of the coefficient matrix $\mathbf{A}$ as well as the components of the load vector $\mathbf{b}$ are integrals. For a more involved data function $f(x)$, and when approximating by higher order polynomials and/or solving equations with variable coefficients, these integrals are not easy to compute. Therefore we need to approximate different integrals over subintervals of a partition. This may be done using *quadrature rules*. In simple case one may use usual or composite *midpoint-*, *trapezoidal-*, or *Simpson's-rules*. In more involved cases one may employ *composite Gauss quadrature rules*. We shall briefly introduce the idea of quadrature rule in Chapter 3.

3. Finally we end up with linear systems of equations (LSE) of type (2.1.44). To solve LSE efficiently we may use exact *Gauss - elimination* or the iteration procedures such as *Gauss-Seidel*, *Gauss-Jacobi* or *Over-relaxation methods*. We discuss these concepts in Chapter 4.

2.2 Exercises

Problem 2.1. *Prove that* $V_0^{(q)} := \{v \in V^{(q)} : v(0) = 0\}$*, is a subspace of* $\mathcal{P}^{(q)}(0,1)$*.*

Problem 2.2. *Consider the ODE*

$$\dot{u}(t) = u(t), \quad 0 < t < 1; \qquad u(0) = 1.$$

Compute its Galerkin approximation in $\mathcal{P}^{(q)}(0,1)$*, for* $q = 1, 2, 3,$ *and* 4*.*

Problem 2.3. *Consider the ODE*

$$\dot{u}(t) = u(t), \quad 0 < t < 1; \qquad u(0) = 1.$$

Compute the $L_2(0,1)$ *projection of the exact solution* u *into* $\mathcal{P}^3(0,1)$*.*

Problem 2.4. *Compute the stiffness matrix and load vector in a finite element approximation of the boundary value problem*

$$-u''(x) = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0,$$

with $f(x) = x$ *and* $h = 1/4$*.*

Problem 2.5. *We want to find a solution approximation* $U(x)$ *to*

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

using the ansatz $U(x) = A \sin \pi x + B \sin 2\pi x$*.*

a. Calculate the exact solution $u(x)$*.*

b. Write down the residual $R(x) = -U''(x) - 1$

c. Use the orthogonality condition

$$\int_0^1 R(x) \sin \pi n x \, dx = 0, \quad n = 1, 2,$$

to determine the constants A *and* B*.*

d. Plot the error $e(x) = u(x) - U(x)$*.*

Problem 2.6. *Consider the boundary value problem*

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

a. Verify that the exact solution of the problem is given by

$$u(x) = x - \frac{\sinh x}{\sinh 1}.$$

b. Let $U(x)$ be a solution approximation defined by

$$U(x) = A \sin \pi x + B \sin 2\pi x + C \sin 3\pi x,$$

where A, B, and C are unknown constants. Compute the residual function

$$R(x) = -U''(x) + U(x) - x.$$

c. Use the orthogonality condition

$$\int_0^1 R(x) \sin \pi n x \, dx = 0, \quad n = 1, 2, 3,$$

to determine the constants A, B, and C.

Problem 2.7. *Let $U(x) = \xi_0 \phi_0(x) + \xi_1 \phi_1(x)$ be a solution approximation to*

$$-u''(x) = x - 1, \quad 0 < x < \pi, \quad u'(0) = u(\pi) = 0,$$

where ξ_i, $i = 0, 1$, are unknown coefficients and

$$\phi_0(x) = \cos \frac{x}{2}, \qquad \phi_1(x) = \cos \frac{3x}{2}.$$

a. Find the analytical solution $u(x)$.

b. Define the approximate solution residual $R(x)$.

c. Compute the constants ξ_i using the orthogonality condition

$$\int_0^\pi R(x) \, \phi_i(x) \, dx = 0, \quad i = 0, 1,$$

i.e., by projecting $R(x)$ onto the vector space spanned by $\phi_0(x)$ and $\phi_1(x)$.

Problem 2.8. *Use the projection technique of the previous exercises to solve*

$$-u''(x) = 0, \quad 0 < x < \pi, \quad u(0) = 0, \quad u(\pi) = 2,$$

assuming that $U(x) = A \sin x + B \sin 2x + C \sin 3x + \frac{2}{\pi^2} x^2$.

Problem 2.9. *Show that $(f - P_h f, v) = 0$, $\forall v \in V_h$, if and only if $(f - P_h f, \varphi_i) = 0$, $i = 0, \ldots, N$; where $\{\varphi_i\}_{i=1}^N \subset V_h$ is the basis of hat-functions.*

Chapter 3

Polynomial Interpolation and Numerical Integration in 1d

3.1 Preliminaries

In this chapter we give a detailed study of the interpolation concept, introduced in Chapter 2. To this approach, we consider a real-valued function f, defined on an interval $I = [a, b]$.

Definition 3.1. *A polynomial interpolant $\pi_q f$ of a function f, defined on an interval $I = [a, b]$, is a polynomial of degree $\leq q$ having the nodal values at $q + 1$ distinct points $x_j \in [a, b]$, $j = 0, 1, \ldots, q$, coinciding with those of f, i.e., $\pi_q f \in \mathcal{P}^q(a, b)$ and $\pi_q f(x_j) = f(x_j),\ j = 0, \ldots, q$.*

Below we illustrate this in some simple examples:

Linear interpolation on an interval. We start with the unit interval $I = [0, 1]$ and a function $f : [0, 1] \to \mathbb{R}$ which is continuous. We let $q = 1$ and seek the linear interpolant of f on I, i.e. $\pi_1 f \in \mathcal{P}^1$, such that $\pi_1 f(0) = f(0)$ and $\pi_1 f(1) = f(1)$. Thus we seek the constants C_0 and C_1 in the following representation of $\pi_1 f \in \mathcal{P}^1$,

$$\pi_1 f(x) = C_0 + C_1 x, \qquad x \in I, \tag{3.1.1}$$

where

$$\begin{array}{lll} \pi_1 f(0) = f(0) & \Longrightarrow & C_0 = f(0), \quad \text{and} \\ \pi_1 f(1) = f(1) & \Longrightarrow & C_0 + C_1 = f(1) \Longrightarrow C_1 = f(1) - f(0). \end{array} \tag{3.1.2}$$

Inserting C_0 and C_1 into (3.1.1) it follows that

$$\pi_1 f(x) = f(0) + \Big(f(1) - f(0)\Big)x = f(0)(1-x) + f(1)x := f(0)\lambda_0(x) + f(1)\lambda_1(x).$$

In other words $\pi_1 f(x)$ is represented in two different bases:

$\pi_1 f(x) = C_0 \cdot 1 + C_1 \cdot x$, with $\{1,\, x\}$ as the set of basis functions and

$\pi_1 f(x) = f(0)(1-x) + f(1)x$, with $\{1-x,\, x\}$ as the set of basis functions.

Note that the functions $\lambda_0(x) = 1-x$ and $\lambda_1(x) = x$ are linearly independent, since if

$$0 = \alpha_0(1-x) + \alpha_1 x = \alpha_0 + (\alpha_1 - \alpha_0)x, \qquad x \in I, \tag{3.1.3}$$

then

$$\left.\begin{array}{lll} x = 0 & \Longrightarrow & \alpha_0 = 0 \\ x = 1 & \Longrightarrow & \alpha_1 = 0 \end{array}\right\} \Longrightarrow \alpha_0 = \alpha_1 = 0. \tag{3.1.4}$$

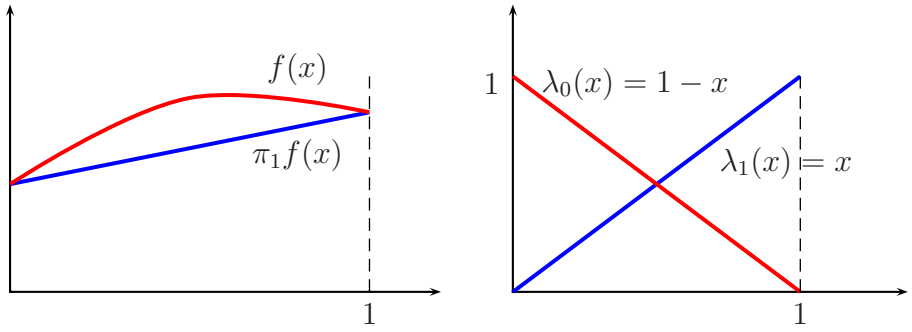


Figure 3.1: Linear interpolation and basis functions for $q = 1$.

Remark 3.1. *Note that if we define a scalar product on $\mathcal{P}^k(a,b)$ by*

$$(p,q) = \int_a^b p(x)q(x)\,dx, \qquad \forall p, q \in \mathcal{P}^k(a,b), \tag{3.1.5}$$

then we can easily verify that neither $\{1,x\}$ nor $\{1-x,x\}$ is an orthogonal basis for $\mathcal{P}^1(0,1)$, since $(1,x) := \int_0^1 1 \cdot x\,dx = [\frac{x^2}{2}] = \frac{1}{2} \neq 0$ and $(1-x,x) := \int_0^1 (1-x)x\,dx = \frac{1}{6} \neq 0$.

Now it is natural to ask the following question.

Question 3.1. *How well does $\pi_q f$ approximate f? In other words how large/small will the error be in approximating $f(x)$ by $\pi_q f(x)$?*

To answer this question we need to estimate the difference between $f(x)$ and $\pi_q f(x)$. For instance for $q = 1$, geometrically, the deviation of $f(x)$ from $\pi_1 f(x)$ (from being linear) depends on the *curvature* of $f(x)$, i.e. on how *curved* $f(x)$ is. In other words, on how *large* $f''(x)$ is, say, on an interval (a, b). To quantify the relationship between the size of the error $f - \pi_1 f$ and the size of f'', we need to introduce some measuring instrument for vectors and functions:

Definition 3.2. *Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ be two column vectors (T stands for transpose). We define the scalar product of $\mathbf{x}$ and $\mathbf{y}$ by*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = x_1 y_1 + \cdots + x_n y_n,$$

and the vector norm for $\mathbf{x}$ as the Euclidean length of $\mathbf{x}$:

$$\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{x_1^2 + \cdots + x_n^2}.$$

$L_p(a, b)$-norm: *Assume that f is a real valued function defined on the interval (a, b). Then we define the L_p-norm ($1 \le p \le \infty$) of f by*

L_p-norm $$\|f\|_{L_p(a,b)} := \Big(\int_a^b |f(x)|^p dx \Big)^{1/p}, \quad 1 \le p < \infty,$$

L_∞-norm $$\|f\|_{L_\infty(a,b)} := \max_{x \in [a,b]} |f(x)|.$$

For $1 \le p \le \infty$ we define the space

$$L_p(a, b) := \{f : \|f\|_{L_p(a,b)} < \infty\}.$$

Below we shall answer Question 3.1, first in the L_∞-norm, and then in the L_p-norm.

Theorem 3.1. *(L_∞-error estimates for linear interpolation in an interval) Assume that $f'' \in L_\infty(a, b)$. Then, for $q = 1$, i.e. only 2 interpolation nodes (the end-points of the interval), there are interpolation constants, c_i, independent of the function f and the interval $[a, b]$, such that*

(1) $\|\pi_1 f - f\|_{L_\infty(a,b)} \le c_i(b-a)^2\|f''\|_{L_\infty(a,b)}$

(2) $\|\pi_1 f - f\|_{L_\infty(a,b)} \le c_i(b-a)\|f'\|_{L_\infty(a,b)}$

(3) $\|(\pi_1 f)' - f'\|_{L_\infty(a,b)} \le c_i(b-a)\|f''\|_{L_\infty(a,b)}$.

Proof. Note that every linear function, $p(x)$ on $[a,b]$ can be written as a linear combination of the basis functions $\lambda_a(x)$ and $\lambda_b(x)$ where

$$\lambda_a(x) = \frac{b-x}{b-a} \quad \text{and} \quad \lambda_b(x) = \frac{x-a}{b-a}: \tag{3.1.6}$$

$$p(x) = p(a)\lambda_a(x) + p(b)\lambda_b(x). \tag{3.1.7}$$

For further reference, we point out the linear combinations of $\lambda_a(x)$ and $\lambda_b(x)$ that gives the basis functions $\{1, x\}$ for $\mathcal{P}^1$:

$$\lambda_a(x) + \lambda_b(x) = 1, \qquad a\lambda_a(x) + b\lambda_b(x) = x. \tag{3.1.8}$$

Note that $\pi_1 f(x)$ is a linear function connecting the two points $(a, f(a))$ and $(b, f(b))$,

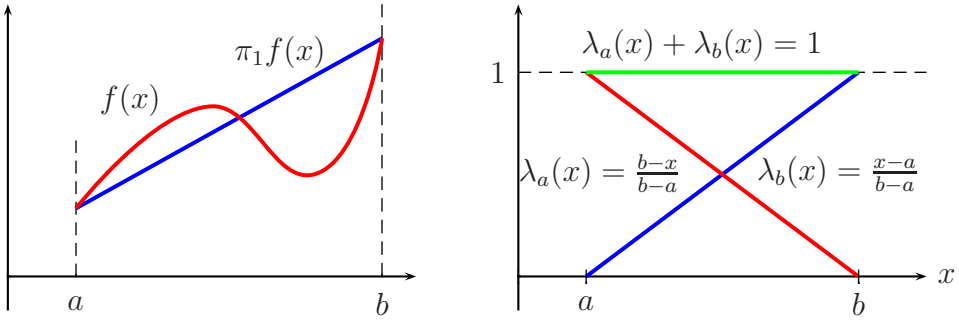


Figure 3.2: Linear Lagrange basis functions for $q = 1$.

and can be represented as

$$\pi_1 f(x) = f(a)\lambda_a(x) + f(b)\lambda_b(x). \tag{3.1.9}$$

By the Taylor expansion for f about $x \in (a,b)$ we can write

$$\begin{cases} f(a) = f(x) + (a-x)f'(x) + \dfrac{1}{2}(a-x)^2 f''(\eta_a), & \eta_a \in [a,x] \\ f(b) = f(x) + (b-x)f'(x) + \dfrac{1}{2}(b-x)^2 f''(\eta_b), & \eta_b \in [x,b]. \end{cases} \tag{3.1.10}$$

Inserting $f(a)$ and $f(b)$ from (3.1.10) into (3.1.9), it follows that

$$\begin{aligned}\pi_1 f(x) =&[f(x) + (a-x)f'(x) + \frac{1}{2}(a-x)^2 f''(\eta_a)]\lambda_a(x) + \\ &+[f(x) + (b-x)f'(x) + \frac{1}{2}(b-x)^2 f''(\eta_b)]\lambda_b(x).\end{aligned}$$

Rearranging the terms, using (3.1.8) and the identity (which also follows from (3.1.8)) $(a-x)\lambda_a(x) + (b-x)\lambda_b(x) = 0$ we get

$$\begin{aligned}\pi_1 f(x) &= f(x)[\lambda_a(x) + \lambda_b(x)] + f'(x)[(a-x)\lambda_a(x) + (b-x)\lambda_b(x)] + \\ &\quad + \frac{1}{2}(a-x)^2 f''(\eta_a)\lambda_a(x) + \frac{1}{2}(b-x)^2 f''(\eta_b)\lambda_b(x) = \\ &= f(x) + \frac{1}{2}(a-x)^2 f''(\eta_a)\lambda_a(x) + \frac{1}{2}(b-x)^2 f''(\eta_b)\lambda_b(x).\end{aligned}$$

Consequently

$$|\pi_1 f(x) - f(x)| = \Big|\frac{1}{2}(a-x)^2 f''(\eta_a)\lambda_a(x) + \frac{1}{2}(b-x)^2 f''(\eta_b)\lambda_b(x)\Big|. \quad (3.1.11)$$

To proceed, we note that for $a \le x \le b$ both $(a-x)^2 \le (a-b)^2$ and $(b-x)^2 \le (a-b)^2$, furthermore $\lambda_a(x) \le 1$ and $\lambda_b(x) \le 1$, $\forall\, x \in (a,b)$. Moreover, by the definition of the maximum norm $|f''(\eta_a)| \le \|f''\|_{L_\infty(a,b)}$, $|f''(\eta_b)| \le \|f''\|_{L_\infty(a,b)}$. Thus we may estimate (3.1.11)

$$|\pi_1 f(x) - f(x)| \le \frac{1}{2}(a-b)^2 \cdot 1 \cdot \|f''\|_{L_\infty(a,b)} + \frac{1}{2}(a-b)^2 \cdot 1 \cdot \|f''\|_{L_\infty(a,b)}, \quad (3.1.12)$$

and hence

$$|\pi_1 f(x) - f(x)| \le (a-b)^2 \|f''\|_{L_\infty(a,b)} \quad \text{corresponding to} \quad c_i = 1. \quad (3.1.13)$$

The other two estimates (2) and (3) are proved similarly. □

Remark 3.2. *It can be shown that the optimal value of* $c_i = \frac{1}{8}$ *(cf Problem 3.10).*

An analogue to Theorem 3.1 can be proved in the L_p-norm, $1 \le p \le \infty$. This general version (concisely stated below as Theorem 3.2) is the frequently used L_p-interpolation error estimate.

Theorem 3.2. *Let $\pi_1 v(x)$ be the linear interpolant of the function $v(x)$ on (a,b). Then, assuming that v is sufficiently regular ($v \in \mathcal{C}^2(a,b)$), there are interpolation constants c_i such that for $1 \le p \le \infty$,*

$$\|\pi_1 v - v\|_{L_p(a,b)} \le c_i (b-a)^2 \|v''\|_{L_p(a,b)}, \tag{3.1.14}$$
$$\|(\pi_1 v)' - v'\|_{L_p(a,b)} \le c_i (b-a) \|v''\|_{L_p(a,b)}, \tag{3.1.15}$$
$$\|\pi_1 v - v\|_{L_p(a,b)} \le c_i (b-a) \|v'\|_{L_p(a,b)}. \tag{3.1.16}$$

For $p = \infty$ this is just the previous Theorem 3.1.

Proof. The proof is similar to that of Theorem 3.1 and is left as an exercise. □

Below we review a simple piecewise linear interpolation procedure on a partition of an interval:

Vector space of piecewise linear functions on an interval. Given $I = [a,b]$, let $\mathcal{T}_h : a = x_0 < x_1 < x_2 < \ldots < x_{N-1} < x_N = b$ be a partition of I into subintervals $I_j = [x_{j-1}, x_j]$ of length $h_j = |I_j| := x_j - x_{j-1}$; $j = 1, 2, \ldots, N$. Let

$$V_h := \{v | v \text{ is a continuous, piecewise linear function on } \mathcal{T}_h\}, \tag{3.1.17}$$

then V_h is a vector space with the previously introduced *hat functions*: $\{\varphi_j\}_{j=0}^N$ as basis functions. Note that $\varphi_0(x)$ and $\varphi_N(x)$ are left and right *half-hat* functions, respectively. We know show that every function in V_h is a linear combination of φ_j:s.

Lemma 3.1. *We have that*

$$\forall v \in V_h; \qquad v(x) = \sum_{j=0}^{N} v(x_j)\varphi_j(x), \quad \Longrightarrow \Big(dim V_h = N+1 \Big). \tag{3.1.18}$$

Proof. Both the left and right hand side are continuous piecewise linear functions. Thus it suffices to show that they have the same nodal values: Let $x = x_j$, then since $\varphi_i(x_j) = \delta_{ij}$,

$$\begin{aligned} RHS|_{x_j} =& v(x_0)\varphi_0(x_j) + v(x_1)\varphi_1(x_j) + \ldots + v(x_{j-1})\varphi_{j-1}(x_j) \\ &+ v(x_j)\varphi_j(x_j) + v(x_{j+1})\varphi_{j+1}(x_j) + \ldots + v(x_N)\varphi_N(x_j) \\ =& v(x_j) = LHS|_{x_j}. \end{aligned} \tag{3.1.19}$$

□

Definition 3.3. *For a partition* $\mathcal{T}_h : a = x_0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = b$ *of the interval* $[a, b]$ *we define the mesh function* $h(x)$ *as the piecewise constant function* $h(x) := h_j = x_j - x_{j-1}$ *for* $x \in I_j = (x_{j-1}, x_j)$, $j = 1, 2, \ldots, N+1$.

Definition 3.4. *Assume that* f *is a continuous function in* $[a, b]$. *Then the continuous piecewise linear interpolant of* f *is defined by*

$$\pi_h f(x) = \sum_{j=0}^{N} f(x_j)\varphi_j(x), \qquad x \in [a, b],$$

where

$$\pi_h f(x_j) = f(x_j), \qquad j = 0, 1, \ldots, N. \tag{3.1.20}$$

Here the sub-index h *refers to the mesh function* $h(x)$.

Remark 3.3. *Note that we denote the linear interpolant, defined for a single interval* $[a, b]$, *by* $\pi_1 f$ *which is a polynomial of degree 1, whereas the piecewise linear interpolant* $\pi_h f$ *is defined for a partition* $\mathcal{T}_h$ *of* $[a, b]$ *and is a piecewise linear function. For the piecewise polynomial interpolants of (higher) degree* q *we shall use the notation for Cardinal functions of Lagrange interpolation (see Section 3.2).*

Note that for each interval I_j, $j = 1, \ldots, N$, we have that

(i) $\pi_h f(x)$ is linear on $I_j \implies \pi_h f(x) = c_0 + c_1 x$ for $x \in I_j$.

(ii) $\pi_h f(x_{j-1}) = f(x_{j-1})$ and $\pi_h f(x_j) = f(x_j)$.

Now using (i) and (ii) we get the equation system

$$\begin{cases} \pi_h f(x_{j-1}) = c_0 + c_1 x_{j-1} = f(x_{j-1}) \\ \pi_h f(x_j) = c_0 + c_1 x_j = f(x_j) \end{cases} \implies \begin{cases} c_1 = \frac{f(x_j) - f(x_{j-1})}{x_j - x_{j-1}} \\ c_0 = \frac{-x_{j-1} f(x_j) + x_j f(x_{j-1})}{x_j - x_{j-1}}, \end{cases}$$

Consequently we may write

$$\begin{cases} c_0 = f(x_{j-1})\frac{x_j}{x_j - x_{j-1}} + f(x_j)\frac{-x_{j-1}}{x_j - x_{j-1}} \\ c_1 x = f(x_{j-1})\frac{-x}{x_j - x_{j-1}} + f(x_j)\frac{x}{x_j - x_{j-1}}. \end{cases} \tag{3.1.21}$$

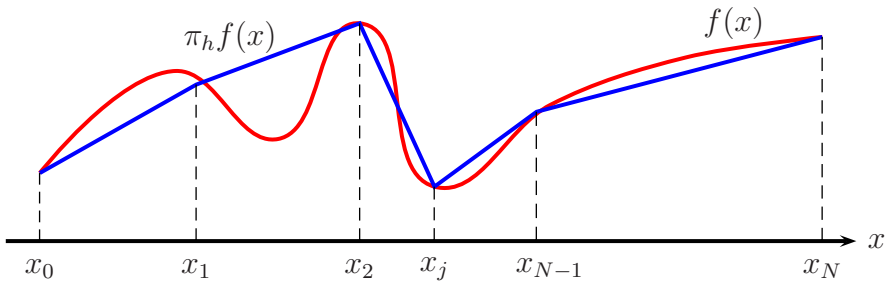


Figure 3.3: Piecewise linear interpolant $\pi_h f(x)$ of $f(x)$.

Hence for $x_{j-1} \leq x \leq x_j, \quad j = 1, 2, \ldots, N,$

$$\begin{aligned}\pi_h f(x) = c_0 + c_1 x &= f(x_{j-1})\frac{x_j - x}{x_j - x_{j-1}} + f(x_j)\frac{x - x_{j-1}}{x_j - x_{j-1}} \\ &= f(x_{j-1})\lambda_{j-1}(x) + f(x_j)\lambda_j(x),\end{aligned}$$

where $\lambda_{j-1}(x)$ and $\lambda_j(x)$ are the restrictions of the piecewise linear basis functions $\varphi_{j-1}(x)$ and $\varphi_j(x)$ to I_j.

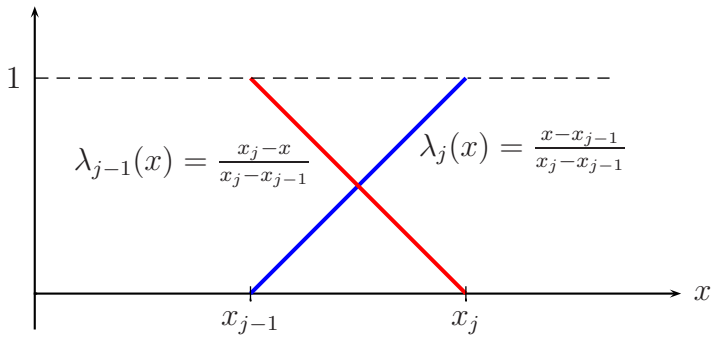


Figure 3.4: Linear Lagrange basis functions for $q = 1$ on the subinterval I_j.

In the next section we generalize the above procedure and introduce Lagrange interpolation basis functions.

The main result of this section can be stated in the following theorem:

Theorem 3.3. *Let $\pi_h v(x)$ be the piecewise linear interpolant of the function $v(x)$ on the partition $\mathcal{T}_h$. Then assuming that is sufficiently regular ($v \in \mathcal{C}^2(a,b)$), there are interpolation constants c_i such that for $1 \le p \le \infty$,*

$$\|\pi_h v - v\|_{L_p(a,b)} \le c_i \|h^2 v''\|_{L_p(a,b)}, \tag{3.1.22}$$

$$\|(\pi_h v)' - v'\|_{L_p(a,b)} \le c_i \|h v''\|_{L_p(a,b)}, \tag{3.1.23}$$

$$\|\pi_h v - v\|_{L_p(a,b)} \le c_i \|h v'\|_{L_p(a,b)}. \tag{3.1.24}$$

Proof. Recalling the definition of the partition $\mathcal{T}_h$, we may write

$$\begin{aligned} \|\pi_h v - v\|^p_{L_p(a,b)} &= \sum_{j=0}^{N+1} \|\pi_h v - v\|^p_{L_p(I_j)} \le \sum_{j=0}^{N+1} c_i^p \|h_j^2 v''\|_{L_p(I_j)} \\ &\le c_i^p \|h^2 v''\|_{L_p(a,b)}, \end{aligned} \tag{3.1.25}$$

where in the first inequality we apply Theorem 3.2 to an arbitrary partition interval I_j and them sum over j. The other two estimates can be proved similarly. □

3.2 Lagrange interpolation

Consider $\mathcal{P}^q(a,b)$; the vector space of all polynomials of degree $\le q$ on the interval (a,b), with the basis functions $1, x, x^2, \ldots, x^q$. We have seen, in Chapter 2, that this is a non-orthogonal basis that leads to ill-conditioned coefficient matrices. We will now introduce a new set of basis functions, which being *almost orthogonal* have some useful properties.

Definition 3.5 (Cardinal functions)**.** *Lagrange basis is the set of polynomials $\{\lambda_i\}_{i=0}^q \subset P^q(a,b)$ associated with the $(q+1)$ distinct points, $a = x_0 < x_1 < \ldots < x_q = b$ in $[a,b]$ and determined by the requirement that: at the nodes, $\lambda_i(x_j) = 1$ for $i = j$, and 0 otherwise ($\lambda_i(x_j) = 0$ for $i \ne j$), i.e. for $x \in [a,b]$,*

$$\lambda_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1}) \downarrow (x-x_{i+1})\ldots(x-x_q)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1}) \uparrow (x_i-x_{i+1})\ldots(x_i-x_q)}. \tag{3.2.1}$$

By the arrows $\downarrow, \uparrow$ in (3.2.1) we want to emphasize that $\lambda_i(x) = \prod_{j \ne i} \left(\frac{x-x_j}{x_i-x_j}\right)$ does not contain the singular factor $\frac{x-x_i}{x_i-x_i}$. Hence

$$\lambda_i(x_j) = \frac{(x_j-x_0)(x_j-x_1)\ldots(x_j-x_{i-1})(x_j-x_{i+1})\ldots(x_j-x_q)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_q)} = \delta_{ij},$$

and $\lambda_i(x)$, $i = 0, 1, \ldots, q$, *is a polynomial of degree* q *on* (a, b) *with*

$$\lambda_i(x_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{3.2.2}$$

Example 3.1. *Let* $q = 2$, *then we have* $a = x_0 < x_1 < x_2 = b$, *where*

$$i = 1, j = 2 \Rightarrow \delta_{12} = \lambda_1(x_2) = \frac{(x_2 - x_0)(x_2 - x_2)}{(x_1 - x_0)(x_1 - x_2)} = 0$$

$$i = j = 1 \Rightarrow \delta_{11} = \lambda_1(x_1) = \frac{(x_1 - x_0)(x_1 - x_2)}{(x_1 - x_0)(x_1 - x_2)} = 1.$$

A polynomial $P(x) \in \mathcal{P}^q(a, b)$ with the values $p_i = P(x_i)$ at the nodes x_i, $i = 0, 1, \ldots, q$, can be expressed in terms of the above Lagrange basis as

$$P(x) = p_0\lambda_0(x) + p_1\lambda_1(x) + \ldots + p_q\lambda_q(x). \tag{3.2.3}$$

Using (3.2.2), $P(x_i) = p_0\lambda_0(x_i) + p_1\lambda_1(x_i) + \ldots + p_i\lambda_i(x_i) + \ldots + p_q\lambda_q(x_i) = p_i$. Recall that in the previous chapter, introducing examples of finite dimensional linear spaces, we did construct Lagrange basis functions for $q = 1$: $\lambda_0(x) = (x - \xi_1)/(\xi_0 - \xi_1)$ and $\lambda_1(x) = (x - \xi_0)/(\xi_1 - \xi_0)$, for an arbitrary subinterval $(\xi_0, \xi_1) \subset (a, b)$.

For a continuous function $f(x)$ on (a, b), we define the *Lagrange interpolation polynomial* $\pi_q f \in P^q(a, b)$.

Definition 3.6. *Let* $a \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq b$, *be* $q + 1$ *distinct interpolation nodes on* $[a, b]$. *Then* $\pi_q f \in \mathcal{P}^q(a, b)$ *interpolates* $f(x)$ *at the nodes* ξ_i, *if*

$$\pi_q f(\xi_i) = f(\xi_i), \qquad i = 0, 1, \ldots, q \tag{3.2.4}$$

and the Lagrange formula (3.2.3) *for* $\pi_q f(x)$ *reads as*

$$\pi_q f(x) = f(\xi_0)\lambda_0(x) + f(\xi_1)\lambda_1(x) + \ldots + f(\xi_q)\lambda_q(x), \quad a \leq x \leq b.$$

Example 3.2. *For* $q = 1$, *we have only the nodes* a *and* b. *Recall that* $\lambda_a(x) = \dfrac{b - x}{b - a}$ *and* $\lambda_b(x) = \dfrac{x - a}{b - a}$, *thus as in the introduction to this chapter we have that*

$$\pi_1 f(x) = f(a)\lambda_a(x) + f(b)\lambda_b(x). \tag{3.2.5}$$

Below we want to compare the Lagrange polynomial of degree q with another well-known polynomial: namely the *Taylor polynomial* of degree q.

Definition 3.7 (Taylor's Theorem)**.** *Suppose that the function f is $q+1$-times continuously differentiable at the point $x_0 \in (a,b)$. Then, f can be expressed by a Taylor expansion about x_0 as*

$$f(x) = T_q f(x) + R_q f(x), \tag{3.2.6}$$

where

$$T_q f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)(x-x_0)^2 + \ldots + \frac{1}{q!}f^{(q)}(x_0)(x-x_0)^q,$$

is the Taylor polynomial of degree q, approximating f and

$$R_q f(x) = \frac{1}{(q+1)!}f^{(q+1)}(\xi)(x-x_0)^{q+1}, \tag{3.2.7}$$

is the remainder term, where ξ is a point between x_0 and x.

Theorem 3.4. *We have the following error estimate*

$$|f(x) - T_q f(x)| = |R_q(f)| \leq \frac{1}{(q+1)!}|x-x_0|^{q+1} \cdot \max_{x\in[a,b]} |f^{(q+1)}(x)|,$$

for the Taylor polynomial which is of order $q+1$ near $x = x_0$; and

$$|f(x) - \pi_q f(x)| \leq \frac{1}{(q+1)!}\prod_{i=0}^{q}|x-x_i| \cdot \max_{x\in[a,b]} |f^{(q+1)}(x)|,$$

for the Lagrange interpolant which is of order 1 (the convergence rate is of order 1) at each interpolation node $x_0, x_1, \ldots, x_q$.

Proof. The *Taylor polynomial error* follows immediately from (3.2.7). As for the *Lagrange* interpolation error we note that at the node points x_i we have $f(x_i) - \pi_q f(x_i) = 0$, for $i = 0, 1, \ldots, q$. Thus, $f(x) - \pi_q f(x)$ has $q+1$ zeros in $[a,b]$ at the interpolation nodes x_i and hence there is a function $g(x)$ defined on $[a,b]$ such that

$$f(x) - \pi_q f(x) = (x-x_0)(x-x_1)\ldots(x-x_q)g(x). \tag{3.2.8}$$

To determine $g(x)$ we define an auxiliary function φ by

$$\varphi(t) := f(t) - \pi_q f(t) - (t - x_0)(t - x_1)\dots(t - x_q)g(x). \tag{3.2.9}$$

Note that $g(x)$ is independent of t. $\varphi(t)$ vanishes at the nodes $x_i,\ i = 0, \dots, q$ as well as for $t = x$, i.e. $\varphi(x_0) = \varphi(x_1) = \dots = \varphi(x_q) = \varphi(x) = 0$. Thus $\varphi(t)$ has $(q + 2)$ roots in the interval $[a, b]$. Now by the *Generalized Rolle's theorem* (see below), there exists a point $\xi \in (a, b)$ such that $\varphi^{(q+1)}(\xi) = 0$. Taking the $(q + 1)$-th derivative of the function $\varphi(t)$, with respect to t, we get using the fact that $\deg(\pi_q f(x)) = q$,

$$\varphi^{(q+1)}(t) = f^{(q+1)}(t) - 0 - (q + 1)!g(x), \tag{3.2.10}$$

where the last term is due to $(t - x_0)(t - x_1)\dots(t - x_q) = t^{q+1} + \alpha t^q + \dots$, (for some constant α), and $g(x)$ is independent of t. Thus

$$0 = \varphi^{(q+1)}(\xi) = f^{(q+1)}(\xi) - (q + 1)!g(x), \tag{3.2.11}$$

which yields

$$g(x) = \frac{f^{(q+1)}(\xi)}{(q + 1)!}. \tag{3.2.12}$$

Inserting $g(x)$ in (3.2.8) we get the Lagrange interpolation error

$$E(x) = f(x) - \pi_q f(x) = \frac{f^{(q+1)}(\xi)}{(q + 1)!} \prod_{i=0}^{q} (x - x_i), \tag{3.2.13}$$

and the proof is complete. □

Theorem 3.5 (Generalized Rolle's theorem)**.** *If a function $u(x) \in \mathcal{C}^{q+1}(a, b)$ has $(q + 2)$ roots, $x_0, x_1, \dots, x_q, x$, in a closed interval $[a, b]$, then there is a point $\xi \in (a, b)$, generated by $x_0, x_1, \dots, x_q, x$, such that $u^{(q+1)}(\xi) = 0$.*

In the finite element approximation procedure of solving differential equations, with a given source term (data) $f(x)$, we need to evaluate integrals of the form $\int f(x)\varphi_i(x)$, with $\varphi_i(x)$ being a finite element basis function. Such integrals are not easily computable for higher order approximations and more involved data. Further, we encounter matrices with entries being the integrals of products of basis functions, their derivatives and variable coefficients (in case they appear) in the equation. Except some special cases (see calculations for $\mathbf{A}$ and $\mathbf{A}_{unif}$ in the previous chapter), these integrals are not elementary and can only be computed approximately by using numerical integration. Below we briefly review some of these numerical integration techniques.

3.3 Numerical integration, Quadrature rule

We approximate the integral $I = \int_a^b f(x)dx$ using a partition of I into subintervals, where on each subinterval f is approximated by polynomials of a certain degree d. We shall denote the approximate value of the integral I by I_d. To proceed we assume, without loss of generality, that $f(x) > 0$ on (a, b) and that f is continuous on (a, b). Then the integral $I = \int_a^b f(x)dx$ is interpreted as the area of the domain under the curve $y = f(x)$; limited by the x-axis and the lines $x = a$ and $x = b$. We shall approximate this area using the values of f at certain points as follows.

We start by approximating the integral over a single interval $[a, b]$. These rules are referred to as *simple rules*.

i) Simple midpoint rule uses the value of f at the midpoint $\bar{x} := \frac{a+b}{2}$ of $[a, b]$, i.e. $f\Big(\frac{a+b}{2}\Big)$. This means that f is approximated by the constant function (polynomial of degree 0) $P_0(x) = f\Big(\frac{a+b}{2}\Big)$ and the area under the curve $y = f(x)$ by

$$I = \int_a^b f(x)dx \approx (b-a)f\Big(\frac{a+b}{2}\Big). \tag{3.3.1}$$

This is the general idea of the *simple midpoint rule*. To prepare for generalizations, if we let $x_0 = a$ and $x_1 = b$ and assume that the length of the interval is h, then

$$I \approx I_0 = hf\Big(a + \frac{h}{2}\Big) = hf(\bar{x}) \tag{3.3.2}$$

ii) Simple trapezoidal rule uses the values of f at two endpoints a and b, i.e. $f(a)$ and $f(b)$. Here f is approximated by the linear function (polynomial of degree 1) $P_1(x)$ passing through the two points $\Big(a, f(a)\Big)$ and $\Big(b, f(b)\Big)$, consequently, the area under the curve $y = f(x)$ is approximated as

$$I = \int_a^b f(x)dx \approx (b-a)\frac{f(a)+f(b)}{2}. \tag{3.3.3}$$

This is the area of the trapezoidal between the lines $y = 0$, $x = a$ and $x = b$ and under the graph of $P_1(x)$, and therefore is referred to as the *simple trapezoidal rule*. Once again, for the purpose of generalization, we let $x_0 = a$,

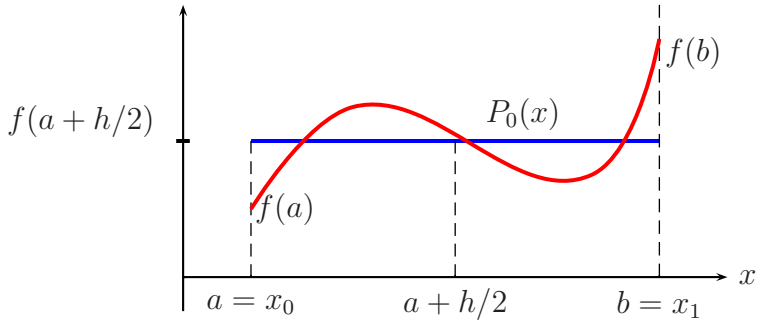


Figure 3.5: Midpoint approximation I_0 of the integral $I = \int_{x_0}^{x_1} f(x)dx$.

$x_1 = b$ and assume that the length of the interval is h, then (3.3.3) can be written as

$$\begin{aligned} I \approx I_1 =& hf(a) + \frac{h[f(a+h) - f(a)]}{2} = h\frac{f(a) + f(a+h)}{2} \\ =& \frac{h}{2}[f(x_0) + f(x_1)]. \end{aligned} \tag{3.3.4}$$

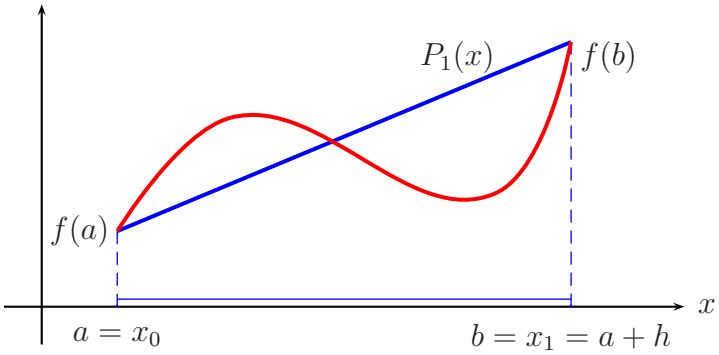


Figure 3.6: Trapezoidal approximation I_1 of the integral $I = \int_{x_0}^{x_1} f(x)dx$.

iii) Simple Simpson's rule uses the values of f at the two endpoints a and b, and the midpoint $\frac{a+b}{2}$ of the interval $[a, b]$, i.e. $f(a)$, $f(b)$, and $f\left(\frac{a+b}{2}\right)$. In this

case the area under $y = f(x)$ is approximated by the area under the graph of the second degree polynomial $P_2(x)$; with $P_2(a) = f(a)$, $P_2\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right)$, and $P_2(b) = f(b)$. To determine $P_2(x)$ we may use Lagrange interpolation for $q = 2$: let $x_0 = a$, $x_1 = (a+b)/2$ and $x_2 = b$, then

$$P_2(x) = f(x_0)\lambda_0(x) + f(x_1)\lambda_1(x) + f(x_2)\lambda_2(x), \tag{3.3.5}$$

where

$$\begin{cases} \lambda_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, \\ \lambda_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}, \\ \lambda_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}. \end{cases} \tag{3.3.6}$$

Thus

$$I = \int_a^b f(x)dx \approx \int_a^b P_2(x)\,dx = \sum_{i=0}^{2} f(x_i) \int_a^b \lambda_i(x)\,dx. \tag{3.3.7}$$

Now we can easily compute the integrals

$$\int_a^b \lambda_0(x)\,dx = \int_a^b \lambda_2(x)\,dx = \frac{b-a}{6}, \quad \int_a^b \lambda_1(x)\,dx = \frac{4(b-a)}{6}. \tag{3.3.8}$$

Hence

$$I = \int_a^b f(x)dx \approx \frac{b-a}{6}[f(x_0) + 4f(x_1) + f(x_2)]. \tag{3.3.9}$$

This is the basic idea behind the *simple Simpson's rule*.

Obviously these approximations are less accurate for large intervals, $[a, b]$ and/or oscillatory functions f. Following Riemann's idea we can use these rules, instead of on the whole interval $[a, b]$, for the subintervals in an appropriate partition of $[a, b]$. Then we get the generalized versions, namely:

3.3.1 Composite rules for uniform partitions

The composite rules are based on the following *General algorithm* used to approximate the integral

$$I = \int_a^b f(x)dx.$$

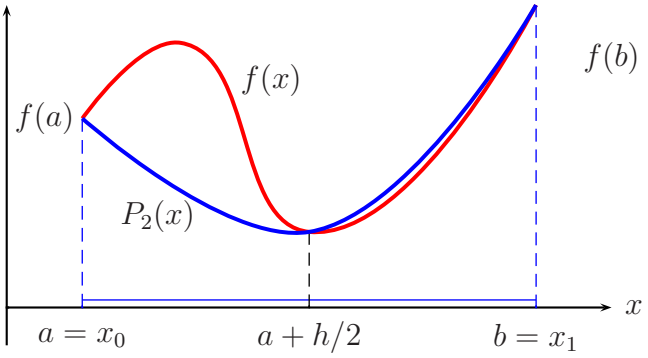


Figure 3.7: Simpson's rule approximation I_2 of the integral $I = \int_{x_0}^{x_1} f(x)dx$.

(1) Divide the interval $[a, b]$, uniformly, into N subintervals

$$a = x_0 < x_1 < x_2 < \ldots < x_{N-1} < x_N = b. \tag{3.3.10}$$

(2) Write the integral as

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)\,dx + \ldots + \int_{x_{N-1}}^{x_N} f(x)\,dx = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} f(x)\,dx. \tag{3.3.11}$$

(3) For each subinterval $I_k := [x_{k-1}, x_k]$, $k = 1, 2, \ldots, N$, apply the *same* integration rule $(i) - (iii)$. Then we get the following generalizations.

(M) *Composite midpoint rule:* approximates f by constants (the values of f at the midpoint of the subinterval) on each subinterval. Let

$$h = |I_k| = \frac{b-a}{N}, \qquad \text{and} \quad \bar{x}_k = \frac{x_{k-1} + x_k}{2}, \quad k = 1, 2, \ldots, N.$$

Then, using the simple midpoint rule for the interval $I_k := [x_{k-1}, x_k]$,

$$\int_{x_{k-1}}^{x_k} f(x)\,dx \approx \int_{x_{k-1}}^{x_k} f(\bar{x}_k)\,dx = hf(\bar{x}_k). \tag{3.3.12}$$

Summing over k, we get the Composite midpoint rule as:

$$\int_a^b f(x)dx \approx \sum_{k=1}^{N} hf(\bar{x}_k) = h[f(\bar{x}_1) + \ldots + f(\bar{x}_N)] := M_N. \tag{3.3.13}$$

(T) *Composite trapezoidal rule*: approximates f by simple trapezoidal rule on each subinterval I_k,

$$\int_{x_{k-1}}^{x_k} f(x)\,dx \approx \frac{h}{2}[f(x_{k-1}) + f(x_k)]. \tag{3.3.14}$$

Summing over k yields the composite trapezoidal rule

$$\begin{aligned}\int_a^b f(x)dx &\approx \sum_{k=1}^{N} \frac{h}{2}[f(x_{k-1}) + f(x_k)] \\ &= \frac{h}{2}[f(x_0) + 2f(x_1) + \ldots + 2f(x_{k-1}) + f(x_k)] := T_N.\end{aligned} \tag{3.3.15}$$

(S) *Composite Simpson's rule*: approximates f by simple Simpson's rule on each subinterval I_k,

$$\int_{x_{k-1}}^{x_k} f(x)\,dx \approx \frac{h}{6}\Big[f(x_{k-1}) + 4f\Big(\frac{x_{k-1}+x_k}{2}\Big) + f(x_k)\Big]. \tag{3.3.16}$$

To simplify, we introduce the following identification on each I_k:

$$z_{2k-2} = x_{k-1}, \quad z_{2k-1} = \frac{x_{k-1}+x_k}{2} := \bar{x}_k, \quad z_{2k} = x_k, \quad h_z = \frac{h}{2}. \tag{3.3.17}$$

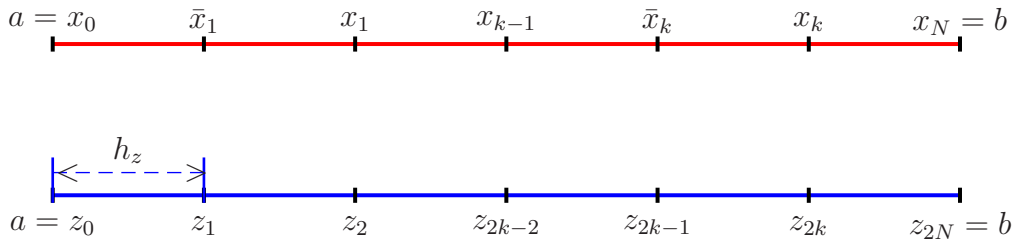


Figure 3.8: Identification of subintervals for composite Simpson's rule

Then, summing (3.3.16) over k and using the above identification, we obtain the composite Simpson's rule viz,

$$\begin{aligned}\int_a^b f(x)dx &\approx \sum_{k=1}^{N} \frac{h}{6}\Big[f(x_{k-1}) + 4f\Big(\frac{x_{k-1}+x_k}{2}\Big) + f(x_k)\Big] \\ &= \sum_{k=1}^{N} \frac{h_z}{3}\Big[f(z_{2k-2}) + 4f(z_{2k-1}) + f(z_{2k})\Big] \\ &= \frac{h_z}{3}\Big[f(z_0) + 4f(z_1) + 2f(z_2) + 4f(z_3) + 2f(z_4) \\ &\quad + \ldots + 2f(z_{2N-2}) + 4f(z_{2N-1}) + f(z_{2N})\Big] := S_N.\end{aligned} \tag{3.3.18}$$

The figure below illustrates the starting procedure for the composite Simpson's rule. The numbers in the brackets indicate the actual coefficient on each subinterval. For instance the end of interval 1: $x_1 = z_2$ coincides with the start of interval 2, yielding the add-up $[1] + [1] = 2$ as the coefficient of $f(z_2)$. A *resonance* which is repeated for each interior node x_k which are z_{2k}:s; $k = 1, \ldots, N-1$.

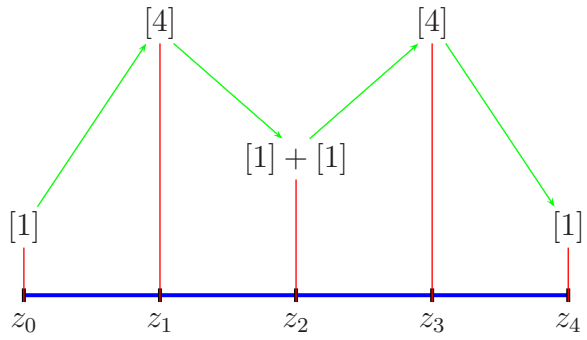


Figure 3.9: Coefficients for composite Simpson's rule

Remark 3.4. *The rules (M), (T) and (S) use values of the function at equally spaced points. These are not always the best approximation methods. Below we introduce an optimal method.*

3.3.2 Gauss quadrature rule

This is an approximate integration rule aimed to choose the points of evaluation of (a non-polynomial) integrand f in an *optimal* manner, not necessarily at equally spaced points. We illustrate this rule by an example:

Problem: Choose the nodes $x_i \in [a,b]$, and coefficients c_i, $1 \le i \le n$ such that, for an arbitrary integrable function f, the following error is minimal:

$$\int_a^b f(x)dx - \sum_{i=1}^{n} c_i f(x_i). \tag{3.3.19}$$

Solution. The relation (3.3.19) contains $2n$ unknowns consisting of n nodes x_i and n coefficients c_i. Therefore we need $2n$ equations. Thus if we replace f by a polynomial, then an optimal choice of these $2n$ parameters yields a quadrature rule (3.3.19) which is exact for *polynomials*, f, of *degree* $\le 2n-1$.

Example 3.3. *Let $n = 2$ and $[a,b] = [-1,1]$. Then the coefficients are c_1 and c_2 and the nodes are x_1 and x_2. Thus optimal choice of these 4 parameters should yield that the approximation*

$$\int_{-1}^{1} f(x)dx \approx c_1 f(x_1) + c_2 f(x_2), \tag{3.3.20}$$

is indeed exact for $f(x)$ replaced by any polynomial of degree ≤ 3. So, we replace f by a polynomial of the form $f(x) = Ax^3 + Bx^2 + Cx + D$ and require equality in (3.3.20)*. Thus, to determine the coefficients c_1, c_2 and the nodes x_1, x_2, in an optimal way, it suffices to change the above approximation to equality when f is replaced by the basis functions for polynomials of degree ≤ 3: i.e. $1, x, x^2$ and x^3. Consequently we get the equation system*

$$\begin{aligned}
&\int_{-1}^{1} 1dx = c_1 + c_2 \text{ and we get } [x]_{-1}^{1} = 2 = c_1 + c_2 \\
&\int_{-1}^{1} xdx = c_1 \cdot x_1 + c_2 \cdot x_2 \text{ and } \left[\frac{x^2}{2}\right]_{-1}^{1} = 0 = c_1 \cdot x_1 + c_2 \cdot x_2 \\
&\int_{-1}^{1} x^2dx = c_1 \cdot x_1^2 + c_2 \cdot x_2^2 \text{ and } \left[\frac{x^3}{3}\right]_{-1}^{1} = \frac{2}{3} = c_1 \cdot x_1^2 + c_2 \cdot x_2^2 \\
&\int_{-1}^{1} x^3dx = c_1 \cdot x_1^3 + c_2 \cdot x_2^3 \text{ and } \left[\frac{x^4}{4}\right]_{-1}^{1} = 0 = c_1 \cdot x_1^3 + c_2 \cdot x_2^3,
\end{aligned} \tag{3.3.21}$$

which, although nonlinear, has the unique solution presented below:

$$\begin{cases} c_1 + c_2 = 2 \\ c_1x_1 + c_2x_2 = 0 \\ c_1x_1^2 + c_2x_2^2 = \frac{2}{3} \\ c_1x_1^3 + c_2x_2^3 = 0 \end{cases} \implies \begin{cases} c_1 = 1 \\ c_2 = 1 \\ x_1 = -\frac{\sqrt{3}}{3} \\ x_2 = \frac{\sqrt{3}}{3} \end{cases} \tag{3.3.22}$$

Hence, the approximation

$$\int_{-1}^{1} f(x)dx \approx c_1f(x_1) + c_2f(x_2) = f\Big(-\frac{\sqrt{3}}{3}\Big) + f\Big(\frac{\sqrt{3}}{3}\Big), \tag{3.3.23}$$

is exact for all polynomials of degree ≤ 3.

Example 3.4. *Let* $f(x) = 3x^2 + 2x + 1$. *Then* $\int_{-1}^{1}(3x^2 + 2x + 1)dx = [x^3 + x^2 + x]_{-1}^{1} = 4$, *and we can easily check that* $f(-\sqrt{3}/3) + f(\sqrt{3}/3) = 4$, *which is also the exact value of the integral.*

Higher order Gauss quadrature. To generalize the Gauss quadrature rule to $n > 2$ Legendre polynomials are used. To illustrate, we choose $\{P_n\}_{n=0}^{\infty}$ such that

(1) For each n, P_n is a polynomial of degree n.

(2) $P_n \perp P_m$ if $m \neq n \iff \int_{-1}^{1} P_n(x)P_m(x)dx = 0$.

The *Legendre polynomials*, on $[-1, 1]$, can be obtained through the formula; (see Chapter 2, Overture),

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\Big((x^2 - 1)^n\Big).$$

Here are the few first Legendre polynomials:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}, \quad P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x, \ldots,$$

The roots of Legendre polynomials are *distinct*, *symmetric*, with respect to $x = 0$, and the *correct choices* as *quadrature* points, i.e. the roots of the Legendre polynomial P_n ($P_0 = 1$ is an exception) are our optimal quadrature points $x_i, 1 \leq i \leq n$ (viz, Theorem 3.5 below).

Example 3.5. *Roots of the Legendre polynomial as quadrature points:*

$$P_1(x) = x = 0.$$

$$P_2(x) = \frac{3}{2}x^2 - \frac{1}{2} = 0, \quad \textit{gives} \quad x_{1,2} = \pm\frac{\sqrt{3}}{3}. \quad \textit{(compare with the result above).}$$

$$P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x = 0, \quad \textit{gives} \quad x_1 = 0,\ x_{2,3} = \pm\sqrt{\frac{3}{5}}.$$

Theorem 3.6. *Suppose that $x_i, i = 1, 2, \ldots, n$, are the roots of the n-th Legendre polynomial P_n and that*

$$c_i = \int_{-1}^{1} \prod_{\substack{j=1 \\ j\neq i}}^{n} \Big(\frac{x - x_j}{x_i - x_j}\Big) dx,$$

where

$$\prod_{\substack{j=1 \\ j\neq i}}^{n} \Big(\frac{x - x_j}{x_i - x_j}\Big),$$

are the Lagrange basis functions with interpolation nodes x_i. If $f(x)$ is a polynomial of degree $< 2n$, then $\displaystyle\int_{-1}^{1} f(x)dx = \sum_{i=1}^{n} c_i f(x_i)$.

Proof. Consider a polynomial $R(x)$ of degree $< n$. We can rewrite $R(x)$ as a $(n-1)$-th degree Lagrange interpolation polynomial with its nodes at the roots of the n-th Legendre polynomial P_n:

$$R(x) = \sum_{i=1}^{n} \prod_{\substack{j=1 \\ j\neq i}}^{n} \Big(\frac{x - x_j}{x_i - x_j}\Big) R(x_i).$$

This representation of $R(x)$ is exact, since for the error

$$E(x) = \frac{1}{n!}(x - x_1)(x - x_2)\ldots(x - x_n)R^{(n)}(\xi), \quad \text{we have } R^{(n)}(\xi) \equiv 0. \tag{3.3.24}$$

Integrating we get

$$\begin{aligned}\int_{-1}^{1} R(x)dx &= \int_{-1}^{1} \Big[\sum_{i=1}^{n}\prod_{\substack{j=1\\ j\neq i}}^{n}\Big(\frac{x-x_j}{x_i-x_j}\Big)R(x_i)\Big]dx \\ &= \sum_{i=1}^{n}\Big[\int_{-1}^{1}\prod_{\substack{j=1\\ j\neq i}}^{n}\Big(\frac{x-x_j}{x_i-x_j}\Big)dx\Big]R(x_i).\end{aligned} \tag{3.3.25}$$

Hence,

$$\int_{-1}^{1} R(x)dx = \sum_{i=1}^{n} c_i R(x_i), \tag{3.3.26}$$

which shoes the result for polynomials of degree $< n$.

Now consider a polynomial, $P(x)$, of degree $< 2n$. Dividing $P(x)$ by the n-th Legendre polynomial $P_n(x)$, we get

$$P(x) = Q(x) \times P_n(x) + R(x), \qquad \deg Q(x) < n, \quad \deg R(x) < n, \tag{3.3.27}$$

and

$$\int_{-1}^{1} P(x)dx = \int_{-1}^{1} Q(x)P_n(x)dx + \int_{-1}^{1} R(x)dx. \tag{3.3.28}$$

Since $P_0(x), P_1(x), \ldots, P_{n-1}(x)$ are basis functions for $\mathcal{P}_{n-1}(-1,1)$, $Q(x) \perp P_n(x)$, $\forall Q(x)$ with $\deg Q < n$. Using (3.3.27) it follows that

$$\int_{-1}^{1} Q(x)P_n(x)dx = 0 \Longrightarrow \int_{-1}^{1} P(x)dx = \int_{-1}^{1} R(x)dx. \tag{3.3.29}$$

Recall that, the x_i's are the roots of $P_n(x)$, i.e. $P_n(x_i) = 0$. Thus using (3.3.27),

$$P(x_i) = Q(x_i)P_n(x_i) + R(x_i) = R(x_i). \tag{3.3.30}$$

Hence, by (3.3.29), (3.3.26) and (3.3.30),

$$\int_{-1}^{1} P(x)dx = \int_{-1}^{1} R(x)dx = \sum_{i=1}^{n} c_i R(x_i) = \sum_{i=1}^{n} c_i P(x_i). \tag{3.3.31}$$

Summing up:

$$\int_{-1}^{1} P(x)dx = \sum_{i=1}^{n} c_i P(x_i), \tag{3.3.32}$$

and the proof is complete. □

Remark 3.5. *Error estimates for both simple and composite quadrature rules can be found in any elementary book in numerical linear algebra and/or numerical analysis.*

3.4 Exercises

Problem 3.1. *Use the expressions* $\lambda_a(x) = \frac{b-x}{b-a}$ *and* $\lambda_b(x) = \frac{x-a}{b-a}$ *to show that*

$$\lambda_a(x) + \lambda_b(x) = 1, \qquad a\lambda_a(x) + b\lambda_b(x) = x.$$

Give a geometric interpretation by plotting, $\lambda_a(x)$, $\lambda_b(x)$, $\lambda_a(x) + \lambda_b(x)$, $a\lambda_a(x)$, $b\lambda_b(x)$ *and* $a\lambda_a(x) + b\lambda_b(x)$.

Problem 3.2. *Let* $f : [0.1] \to \mathbb{R}$ *be a Lipschitz continuous function. Determine the linear interpolant* $\pi f \in \mathcal{P}(0,1)$ *and plot* f *and* πf *in the same figure, when*

(a) $f(x) = x^2$, *(b)* $f(x) = \sin(\pi x)$.

Problem 3.3. *Determine the linear interpolation of the function*

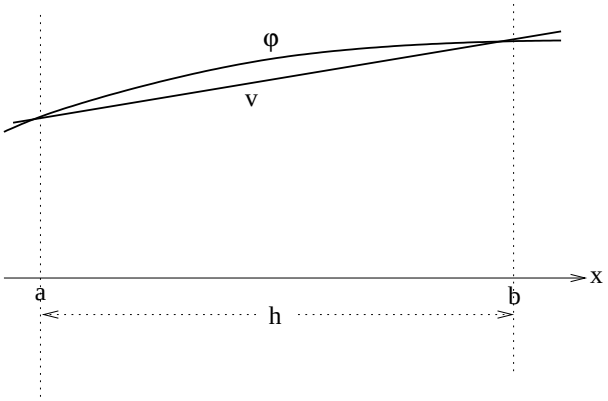
$$f(x) = \frac{1}{\pi^2}(x-\pi)^2 - \cos^2(x - \frac{\pi}{2}), \qquad -\pi \le x \le \pi.$$

where the interval $[-\pi, \pi]$ *is divided into 4 equal subintervals.*

Problem 3.4. *Assume that* $w' \in L_1(I)$. *Let* $x, \bar{x} \in I = [a, b]$ *and* $w(\bar{x}) = 0$. *Show that*

$$|w(x)| \le \int_I |w'| dx. \tag{3.4.1}$$

Problem 3.5. *Assume that* v *interpolates* φ, *at the points* a *and* b.



Show, using (3.4.1) *that*

$$\text{(i)} \qquad |(\varphi - v)(x)| \le \int_I |(\varphi - v)'|\, dx,$$

$$\text{(ii)} \qquad |(\varphi - v)'(x)| \le \int_I |(\varphi - v)''|\, dx = \int_I |\varphi''|\, dx,$$

$$\text{(iii)} \qquad \max_I |\varphi - v| \le \max_I |h^2 \varphi''|,$$

$$\text{(iv)} \qquad \int_I |\varphi - v|\, dx \le \int_I |h^2 \varphi''|\, dx,$$

$$\text{(v)} \qquad \|\varphi - v\|_I \le \|h^2 \varphi''\|_I \quad \textit{and} \quad \|h^{-2}(\varphi - v)\|_I \le \|\varphi''\|_I,$$

where $\|w\|_I = \Big(\int_I w^2\, dx \Big)^{1/2}$ *is the* $L_2(I)$*-norm.*

Problem 3.6. *Use, in the above problem*

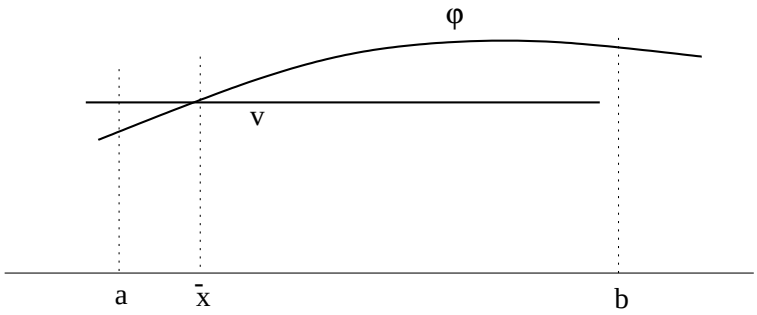
$$v' = \frac{\varphi(b) - \varphi(a)}{h} = \frac{1}{h} \int_a^b \varphi' dx \ \ (\varphi' \textit{ is constant on } I),$$

and show that

$$\text{(vi)} \qquad |(\varphi - v)(x)| \le 2 \int_I |\varphi'|\, dx,$$

$$\text{(vii)} \qquad \int_I h^{-1} |\varphi - v|\, dx \le 2 \int_I |\varphi'|\, dx \quad \textit{and} \quad \|h^{-1}(\varphi - v)\| \le 2 \|\varphi'\|_I.$$

Problem 3.7. *Let now* $v(t)$ *be the constant interpolant of* φ *on* I*.*



Show that

$$\int_I h^{-1}|\varphi - v|\,dx \leq \int_I |\varphi'|\,dx. \tag{3.4.2}$$

Problem 3.8. *Show that*

$$\mathcal{P}^q(a,b) := \{p(x)|p(x) \text{ is a polynomial of degree } \leq q\},$$

is a vector space but

$$P^q(a,b) := \{p(x)|p(x) \text{ is a polynomial of degree } = q\},$$

is not! a vector space.

Problem 3.9. *Compute formulas for the linear interpolant of a continuous function f through the points a and $(b+a)/2$. Plot the corresponding Lagrange basis functions.*

Problem 3.10. *Prove the following interpolation error estimate:*

$$||\pi_1 f - f||_{L_\infty(a,b)} \leq \frac{1}{8}(b-a)^2||f''||_{L_\infty(a,b)}.$$

Hint: Using an scaling argument, it suffices to prove the inequality for $[a,b] \equiv [0,1]$.

Problem 3.11. *Prove that any value of f on the sub-intervals, in a partition of (a,b), can be used to define $\pi_h f$ satisfying the error bound*

$$||f - \pi_h f||_{L_\infty(a,b)} \leq \max_{1\leq i\leq m+1} h_i||f'||_{L_\infty(I_i)} = ||hf'||_{L_\infty(a,b)}.$$

Prove that choosing the midpoint improves the bound by an extra factor $1/2$.

Problem 3.12. *Compute and graph* $\pi_4\Big(e^{-8x^2}\Big)$ *on* $[-2,2]$*, which interpolates* e^{-8x^2} *at 5 equally spaced points in* $[-2,2]$.

Problem 3.13. *Write down a basis for the set of piecewise quadratic polynomials* $W_h^{(2)}$ *on a partition* $a = x_0 < x_1 < x_2 < \ldots < x_{m+1} = b$ *of* (a,b) *into subintervals* $I_i = (x_{i-1}, x_i)$*, where*

$$W_h^{(q)} = \{v : v|_{I_i} \in \mathcal{P}^q(I_i), i = 1, \ldots, m+1\}.$$

Problem 3.14. *Determine a set of basis functions for the space of continuous piecewise quadratic functions* $V_h^{(2)}$ *on* $I = (a,b)$*, where*

$$V_h^{(q)} = \{v \in W_h^{(q)} : v \text{ is continuous on } I\}.$$

Problem 3.15. *Prove that*

$$\int_{x_0}^{x_1} f'\Big(\frac{x_1+x_0}{2}\Big)\Big(x - \frac{x_1+x_0}{2}\Big)\,dx = 0.$$

Problem 3.16. *Prove that*

$$\Big|\int_{x_0}^{x_1} f(x)\,dx - f\Big(\frac{x_1+x_0}{2}\Big)(x-x_0)\Big| \\ \leq \frac{1}{2}\max_{[x_0,x_1]}|f''|\int_{x_0}^{x_1}\Big(x - \frac{x_1+x_0}{2}\Big)^2 dx \leq \frac{1}{24}(x_1-x_0)^3 \max_{[x_0,x_1]}|f''|.$$

Hint: Use Taylor expansion of f *about* $x = \frac{x_1+x_2}{2}$.

Chapter 4

Linear Systems of Equations

We have seen that the numerical solution of a differential equation, e.g. using the Galerkin finite element method, is an approximation of the exact solution in a finite dimensional vector space. If the differential equation is linear, then the procedure ends solving a linear system of equations of the form $A\mathbf{x} = \mathbf{b}$, where the coefficient matrix A is a square matrix of the same order as the dimension of the approximation space, and is related to the differential operator in the equation, $\mathbf{x}$ is a vector in the approximation space with entries consisting of certain nodal values of the approximate solution, and the vector $\mathbf{b}$ is related to the data and basis functions of the approximation space.

The criterion for the quality of a numerical method, to solve a certain problem, lies in its potential of convergence of the approximate solution to the exact one in an adequate measuring environment (norm). Quantitatively, this is expressed by "how fast" the approximate solution would converge to the exact solution by increasing the approximation degree (the dimension of the approximation space), which is a theoretical procedure. In computations however, with an already justified convergence, it is important to have a numerical algorithm that solves the approximate problem reasonably fast (takes shorter time which may be achieved, e.g. by taking a fewer number of approximation points).

This chapter is devoted to the numerical solution of linear systems of equations of type $A\mathbf{x} = \mathbf{b}$. *Throughout this chapter we assume that* $detA \neq 0$, *i.e. the matrix* A *is invertible.* Then $A\mathbf{x} = \mathbf{b} \Leftrightarrow \mathbf{x} = A^{-1}\mathbf{b}$ where we circumvent to invert the matrix A. To this approach we shall review the well-known direct method of Gauss elimination and then continue with some

more efficient iterative methods. A thorough study of this type is in the realm of numerical linear algebra where the solution of linear systems of equations is undoubtedly one of the most applied tools.

4.1 Direct methods

Consider the general form of an $n \times n$ linear system of equations given by

$$A\mathbf{x} = \mathbf{b} \Leftrightarrow \begin{array}{l} \sum_{j=1}^{n} a_{ij}x_j = b_i, \\ i = 1, \ldots, n, \end{array} \quad \text{or} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{cases}$$

We introduce the extended $n \times (n+1)$ coefficient matrix $\mathcal{E}$ consisting of the coefficient matrix A enlarged by putting the right hand side $\mathbf{b}$ as the additional last column:

$$\mathcal{E} := \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & b_n \end{pmatrix}. \tag{4.1.1}$$

Note that to solve the equation system $A\mathbf{x} = \mathbf{b}$, it is a bad idea to calculate A^{-1} and then multiply by $\mathbf{b}$. However, if A is an upper (or lower) triangular matrix, i.e., if $a_{ij} = 0$ for $i > j$ (or $i < j$), and A is invertible, then we can solve $\mathbf{x}$ using the *back substitution method*:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ \qquad\qquad a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \qquad\qquad\qquad \ldots \quad \ldots \quad \ldots \\ \qquad\qquad\qquad \ldots \quad \ldots \quad \ldots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ \qquad\qquad\qquad a_{nn}x_n = b_n, \end{cases} \tag{4.1.2}$$

which yields

$$\begin{cases} x_1 = \dfrac{1}{a_{11}}[b_1 - a_{12}x_2 - \ldots - a_{1n}x_n] \\ \qquad\qquad \ldots \qquad \ldots \qquad \ldots \\ \qquad\qquad \ldots \qquad \ldots \qquad \ldots \\ x_{n-1} = \dfrac{1}{a_{n-1,n-1}}[b_{n-1} - a_{n-1,n}x_n] \\ \qquad\qquad\qquad\qquad x_n = \dfrac{b_n}{a_{nn}}. \end{cases} \tag{4.1.3}$$

Number of operations. The speed of convergence depends on the number of operations performed during the computations. In the above procedure additions and subtractions are not considered as time consuming operations, therefore we shall count only the number of multiplications and divisions.
• The number of multiplications to solve for x_n from (4.1.3) is zero and the number of divisions is one.
• To solve for x_{n-1} we need one multiplication and one division.
• To solve for x_1 we need $(n-1)$ multiplication and one division.
Thus to solve the triangular linear system of equations given by (4.1.2) we shall need

$$1 + 2 + \ldots + (n-1) = \frac{n(n-1)}{2} := \frac{n^2}{2} + Q(n),$$

multiplications, where $Q(n)$ is a remainder of order n, and n divisions.

Gaussian elimination method. This method is based on the following obvious facts expressing that: a solution to a linear system of equations is not changed under *elementary row operations*. These are

(i) interchanging two equations

(ii) adding a multiple of one equation to another

(iii) multiplying an equation by a nonzero constant.

The Gauss elimination procedure is based on splitting the coefficient matrix A to two factors, an upper triangular matrix U and a lower triangular matrix

L, known as LU factorization. Below are examples of 3×3 dimensional *upper triangular matrix* U, *lower triangular matrix* L and *diagonal matrix* D,

$$U = \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}, \qquad L = \begin{pmatrix} a & 0 & 0 \\ g & d & 0 \\ h & i & f \end{pmatrix}, \qquad D = \begin{pmatrix} a & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & f \end{pmatrix}.$$

Of course some of the entries $a, b, \ldots, i$ above may also be zero. To perform the Gauss *elimination procedure*, we start from the first row of the coefficient matrix of the equation system and use elementary row operations to eliminate the elements $a_{i1}, i > 1$, under a_{11} (make $a_{i1} = 0$).

Remark 4.1. *If $a_{11} = 0$, then we interchange equations, and replace the first equation by an equation in the system with $a_{i1} \neq 0$.*

The equation system corresponding to this newly obtained matrix $\tilde{A}$ with elements $\tilde{a}_{ij}$, and where $\tilde{a}_{i1} = 0, i > 1$, has the same solution as the original one. We repeat the same procedure of elementary row operations to eliminate the elements $\tilde{a}_{i2}, i > 2$, from the matrix $\tilde{A}$. Continuing in this way, we obtain an upper triangular matrix U with a corresponding equation system which is equivalent to the original system (has the same solution). Below we shall illustrate this procedure through an *example*.

Example 4.1. *Solve the linear system of equations*

$$\begin{cases} 2x_1 + x_2 + x_3 = 2 \\ 4x_1 - x_2 + 3x_3 = 0 \\ 2x_1 + 6x_2 - 2x_3 = 10. \end{cases} \tag{4.1.4}$$

In the extended coefficient matrix:

$$\mathcal{E} = \left(\begin{array}{ccc|c} 2 & 1 & 1 & 2 \\ 4 & -1 & 3 & 0 \\ 2 & 6 & -2 & 10 \end{array}\right), \tag{4.1.5}$$

we have that $a_{11} = 2$, $a_{21} = 4$, and $a_{31} = 2$. We introduce the multipliers $m_{i1}, i > 1$ by letting

$$m_{21} = \frac{a_{21}}{a_{11}} = \frac{4}{2} = 2 \qquad m_{31} = \frac{a_{31}}{a_{11}} = \frac{2}{2} = 1. \tag{4.1.6}$$

Now we multiply the first row by m_{21} and then subtract it from row 2 and replace the result in row 2:

$$\left(\begin{array}{ccc|c} 2 & 1 & 1 & 2 \\ 4 & -1 & 3 & 0 \\ 2 & 6 & -2 & 10 \end{array}\right) \cdot(-2) \quad \Longrightarrow \left(\begin{array}{ccc|c} 2 & 1 & 1 & 2 \\ 0 & -3 & 1 & -4 \\ 2 & 6 & -2 & 10 \end{array}\right) \tag{4.1.7}$$

Similarly we multiply the first row by $m_{31} = 1$, subtract it from row 3, and replace the result in row 3:

$$\tilde{\mathcal{E}} := \left(\begin{array}{ccc|c} 2 & 1 & 1 & 2 \\ 0 & -3 & 1 & -4 \\ 0 & 5 & -3 & 8 \end{array}\right). \tag{4.1.8}$$

In this setting we have $\tilde{a}_{22} = -3$ and $\tilde{a}_{32} = 5$. Now we let $m_{32} = \tilde{a}_{32}/\tilde{a}_{22} = -5/3$, then multiplying the second row in $\tilde{\mathcal{E}}$ by m_{32} and subtracting the result from row 3 yields

$$\left(\begin{array}{ccc|c} 2 & 1 & 1 & 2 \\ 0 & -3 & 1 & -4 \\ 0 & 0 & -\frac{4}{3} & \frac{4}{3} \end{array}\right), \tag{4.1.9}$$

where we have obtained the upper triangular matrix

$$U = \left(\begin{array}{ccc} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -\frac{4}{3} \end{array}\right). \tag{4.1.10}$$

The new equivalent equation system is

$$\left\{\begin{array}{rcl} 2x_1 + x_2 + x_3 = & & 2 \\ -3x_2 + x_3 = & & -4 \\ -\frac{4}{3}x_3 = & & \frac{4}{3} \end{array}\right. \tag{4.1.11}$$

with the solution $x_1 = 1$, $x_2 = 1$ and $x_3 = -1$ which, as we can verify, is also the solution of the original equation system (4.1.4).

Remark 4.2. *For an $n \times n$ matrix A, the number of operation in the Gauss elimination procedure, leading to an upper triangular matrix U, is of order $\mathcal{O}(n^3)$. Recall that the number of operations to solve an $n \times n$ upper triangular system is $\mathcal{O}(n^2)$. Therefore the total number of operations to solve an $n \times n$ linear system of equations using the Gauss elimination is of order $\mathcal{O}(n^3)$.*

Definition 4.1. *We define the lower triangular matrices:*

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{pmatrix}, L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{pmatrix} \text{ and } L = \begin{pmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{pmatrix},$$

where m_{21}, m_{31} and m_{32} are the multipliers. The matrices L_1, L_2 and L are unit (ones on the diagonal) lower triangular 3×3-matrices with the property that

$$L = (L_2 L_1)^{-1} = L_1^{-1} L_2^{-1}, \qquad \text{and} \qquad A = LU. \tag{4.1.12}$$

Example 4.2. *Continuing the previous example, we have $m_{21} = 2, m_{31} = 1$ and $m_{32} = -5/3$, consequently*

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{5}{3} & 1 \end{pmatrix} \quad \text{and} \quad L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{5}{3} & 1 \end{pmatrix}.$$

Thus

$$L_1 A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 4 & -1 & 3 \\ 2 & 6 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 5 & -3 \end{pmatrix} = \tilde{A},$$

which corresponds to the first two elementary row operations in the Gaussian elimination. Further

$$L_2 L_1 A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{5}{3} & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 5 & -3 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -\frac{4}{3} \end{pmatrix} = U,$$

which corresponds to the last (third) elementary row operation performed in the previous example.

Note that, the last relation means also

$$A = (L_2L_1)^{-1}U = LU.$$

In the general setting we have the following result:

Proposition 4.1. *The $n \times n$ unit lower triangular matrix L is given by*

$$L = (L_{n-1}L_{n-2}\dots L_1)^{-1},$$

where $L_i,\ i = 1,\dots,n-1$ are the corresponding $n\times n$ row-operation matrices, viz example above. For $n = 3$ we have $(L_2L_1)^{-1} = L$, where

$$L = \begin{pmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{pmatrix},$$

and m_{ij} are the multipliers defined above.

LU factorization of an $n \times n$ matrix A

To generalize the above procedure from the 3×3 case, to an $n \times n$ linear system of equations we use the factorization $A = LU$ of the coefficient matrix A, where L is a unit lower triangular matrix and U is an upper triangular matrix obtained from A by Gaussian elimination.
To solve the system $A\mathbf{x} = \mathbf{b}$, assuming we have an LU factorization, we let $\mathbf{y} = U\mathbf{x}$, and first solve $L\mathbf{y} = \mathbf{b}$ by forward substitution (from the first row to the last) and obtain the vector $\mathbf{y}$, then using $\mathbf{y}$ as the known right hand side finally we solve $U\mathbf{x} = \mathbf{y}$ by backward substitution (from the last row to the first) and get the solution $\mathbf{x}$.

Thus

$$\begin{aligned} A\mathbf{x} = \mathbf{b} &\iff (LU)\mathbf{x} = \mathbf{b} \iff L(U\mathbf{x}) = \mathbf{b} \\ &\iff (L\mathbf{y} = \mathbf{b} \quad \wedge \quad U\mathbf{x} = \mathbf{y}). \end{aligned}$$

Observe that, in the Gauss elimination procedure, $\mathbf{y} = \mathbf{b}$ is solved "by automatic", viz

$$\mathbf{y} = L^{-1}\mathbf{b} = L_2L_1\mathbf{b}.$$

Then, we solve

$$U\mathbf{x} = L_2L_1A\mathbf{x} = L_2L_1\mathbf{b} = L^{-1}\mathbf{b} = \mathbf{y},$$

by backward substitution.

Below we illustrate this procedure through an example.

Example 4.3. *We return to the previous example where we have that*

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{5}{3} & 1 \end{pmatrix} \quad \textit{and} \quad b = \begin{pmatrix} 2 \\ 0 \\ 10 \end{pmatrix}.$$

• $L\mathbf{y} = \mathbf{b}$ *yields the system of equations*

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{5}{3} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 10 \end{pmatrix} \Longleftrightarrow \begin{cases} y_1 = & 2 \\ 2y_1 + y_2 = & 0 \\ y_1 - \frac{5}{3}y_2 + y_3 = & 10. \end{cases}$$

Using forward substitution we get $y_1 = 2,\ y_2 = -4,\ y_3 = 4/3$. *Further with*

$$U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -\frac{4}{3} \end{pmatrix} \quad \textit{and} \quad \mathbf{y} = \begin{pmatrix} 2 \\ -4 \\ \frac{4}{3} \end{pmatrix}.$$

• $U\mathbf{x} = \mathbf{y}$ *yields*

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -\frac{4}{3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ -4 \\ \frac{4}{3} \end{pmatrix} \Longleftrightarrow \begin{cases} 2x_1 + x_2 + x_3 = & 2 \\ -3x_2 + x_3 = & -4 \\ -\frac{4}{3}x_3 = & \frac{4}{3}. \end{cases}$$

Using backward substitution, we get the solution: $x_1 = 1,\ x_2 = 1,\ x_3 = -1$.

Remark 4.3. *In some cases we can predict the solubility of an equation system: Let A be an $n \times n$ matrix. Then we have the following properties:*
• *If $det(A) \neq 0$, then the system of equations $A\mathbf{x} = \mathbf{b}$ have a unique solution, given by $\mathbf{x} = A^{-1}\mathbf{b}$.*
• *A has an LU factorization if $det(\Delta_k) \neq 0, \quad k = 1, \ldots, n-1$, where Δ_k is Sylvester's matrix:*

$$\Delta_k = \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \ldots & \ldots & \ldots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix}.$$

• *If the LU factorization exists and $det(A) \neq 0$, then the LU factorization id unique and $det(A) = U_{11} \cdot \ldots \cdot U_{nn}$.*

In particular for the "symmetric matrices" we have:

Theorem 4.1 (Cholesky's method)**.** *Let A be a symmetric matrix, $(a_{ij} = a_{ji})$, then the following statements are equivalent:*

(i) A is positive definite.

(ii) The eigenvalues of A are positive.

(iii) Sylvester's criterion: $\det(\Delta_k) > 0$ *for* $k = 1, 2, \ldots, n$.

(iv) $A = LL^T$ where L is lower triangular and has positive diagonal elements. (Cholesky factorization)

We do not give a proof of this theorem. The interested reader is referred to literature in linear algebra and matrix theory, e.g. G. Goloub, [13].

4.2 Iterative methods

To speed up the solution procedure, instead of solving $A\mathbf{x} = \mathbf{b}$ directly, for the exact solution $\mathbf{x}$, we consider iterative solution methods based on computing a sequence of approximations $\mathbf{x}^{(k)}, k = 1, 2, \ldots$ such that

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x} \quad \text{or} \quad \lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\| = 0, \quad \text{for some matrix norm.}$$

Example 4.4. *For an $n \times n$ matrix A,L_1 and L_∞ matrix norms are defined by*

$$||A||_1 = \max_j \sum_{i=1}^{n} |a_{ij}|,$$
$$||A||_\infty = \max_i \sum_{i=1}^{n} |a_{ij}|.$$

Thus consider the general $n \times n$ linear system of equations $A\mathbf{x} = \mathbf{b}$, where both the coefficient matrix A and the vector $\mathbf{b}$ have real entries,

$$A\mathbf{x} = \mathbf{b} \Longleftrightarrow \begin{cases} a_{11}x_1 + & a_{12}x_2 & + \ldots & +a_{1n}x_n & = b_1 \\ a_{2,1}x_1 & +a_{22}x_2 & \ldots & +a_{2n}x_n & = b_2 \\ \ldots & \ldots & + \ldots & \ldots & \ldots \\ a_{n1}x_1 & + \ldots & \ldots & +a_{nn}x_n & = b_n. \end{cases} \tag{4.2.1}$$

For the system (4.2.1) we shall introduce the two main iterative methods.

Jacobi iteration: Assume that $a_{ii} \neq 0$, then (4.2.1) can be rewritten as:

$$\begin{cases} x_1 = -\dfrac{1}{a_{11}}[a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n - b_1] \\ x_2 = -\dfrac{1}{a_{22}}[a_{21}x_1 + a_{23}x_2 + \ldots + a_{2n}x_n - b_2] \\ \ldots \\ x_n = -\dfrac{1}{a_{nn}}[a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{n,n-1}x_{n-1} - b_n]. \end{cases}$$

We start with a given initial approximation for the solution, viz

$$\mathbf{x}^{(0)} = (x_1^{(0)} = c_1,\, x_2^{(0)} = c_2,\, \ldots,\, x_n^{(0)} = c_n),$$

and, based on the above system, define a successive iteration procedure, for

$k = 0, 1, 2, \ldots$, as:

$$\begin{cases} x_1^{(k+1)} = -\dfrac{1}{a_{11}}[a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} - b_1] \\ \\ x_2^{(k+1)} = -\dfrac{1}{a_{22}}[a_{21}x_1^{(k)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2] \\ \ldots \qquad\qquad\qquad\qquad \ldots \qquad\qquad\qquad\qquad \ldots \\ x_n^{(k+1)} = -\dfrac{1}{a_{nn}}[a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \ldots + a_{n,n-1}x_{n-1}^{(k)} - b_n], \end{cases}$$

or in compact form in *Jacobi coordinates*, by

$$\begin{cases} \sum_{j=1}^n a_{ij}x_j = b_i \Longleftrightarrow a_{ii}x_i = -\sum_{\substack{j=1 \\ j\neq i}}^n a_{ij}x_j + b_i, \\ a_{ii}x_i^{(k+1)} = -\sum_{\substack{j=1 \\ j\neq i}}^n a_{ij}x_j^{(k)} + b_i. \end{cases} \tag{4.2.2}$$

If n is large and the number of iterations are small $< n$, then the Jacobi method requires less operations than the usual Gauss elimination method. Below we state a convergence criterion (the proof is left as an exercise),

Convergence criterion:
The Jacobi method gives convergence to the exact solution if the matrix A is *strictly diagonally dominant*, i.e.

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad i = 1, 2, \ldots, n. \tag{4.2.3}$$

Problem 4.1. *Show that* $A = \begin{pmatrix} 4 & 2 & 1 \\ 1 & 5 & 1 \\ 0 & 1 & 3 \end{pmatrix}$ *is diagonally dominant.*

Example 4.5. *Solve* $A\mathbf{x} = \mathbf{b}$ *where* $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ *and* $\mathbf{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

A is diagonally dominant and the matrix equation $A\mathbf{x} = \mathbf{b}$ *is equivalent to the linear equation system*

$$\begin{cases} 2x_1 - x_2 = 1 \\ -x_1 + 2x_2 = 1. \end{cases} \tag{4.2.4}$$

We choose zero initial values for x_1 *and* x_2*, i.e.* $x_1^{(0)} = 0$ *and* $x_2^{(0)} = 0$ *and use* (4.2.4) *to build the Jacobi iteration system*

$$\begin{cases} 2x_1^{(k+1)} = x_2^{(k)} + 1 \\ 2x_2^{(k+1)} = x_1^{(k)} + 1, \end{cases} \tag{4.2.5}$$

where k *is the iteration step. Then we have*

$$\begin{cases} 2x_1^{(1)} = x_2^{(0)} + 1 \\ 2x_2^{(1)} = x_1^{(0)} + 1 \end{cases} \quad \textit{with the solution} \quad \begin{cases} x_1^{(1)} = 1/2 \\ x_2^{(1)} = 1/2. \end{cases} \tag{4.2.6}$$

In the next iteration step:

$$\begin{cases} 2x_1^{(2)} = x_2^{(1)} + 1 \\ 2x_2^{(2)} = x_1^{(1)} + 1 \end{cases} \Rightarrow \begin{cases} 2x_1^{(2)} = 1/2 + 1 \\ 2x_2^{(2)} = 1/2 + 1 \end{cases} \Rightarrow \begin{cases} x_1^{(2)} = 3/4 \\ x_2^{(2)} = 3/4. \end{cases} \tag{4.2.7}$$

Continuing we obviously have $\lim_{k\to\infty} x_i^{(k)} = x_i$*,* $i = 1, 2$*, where* $x_1 = x_2 = 1$*. Below we show the first few iterations giving the corresponding* $x_1^{(k)}$ *and* $x_2^{(k)}$ *values*

k	$x_1^{(k)}$	$x_2^{(k)}$
0	0	0
1	1/2	1/2
2	3/4	3/4
3	7/8	7/8

Now if we use the maximum norm: $\|\mathbf{e}_k\|_\infty := \max_{i=1,2} |x_i^{(k)} - x_i|$, *then*

$$\|\mathbf{e}_0\|_\infty = \max(|x_1^{(0)} - x_1|, |x_2^{(0)} - x_2|) = \max\left(\left|0-1\right|, \left|0-1\right|\right) = 1$$
$$\|\mathbf{e}_1\|_\infty = \max(|x_1^{(1)} - x_1|, |x_2^{(1)} - x_2|) = \max\left(\left|\frac{1}{2}-1\right|, \left|\frac{1}{2}-1\right|\right) = \frac{1}{2}$$

$$\|\mathbf{e}_2\|_\infty = \max(|x_1^{(2)} - x_1|, |x_2^{(2)} - x_2|) = \max\left(\left|\frac{3}{4}-1\right|, \left|\frac{3}{4}-1\right|\right) = \frac{1}{4}$$

$$\|\mathbf{e}_3\|_\infty = \max(|x_1^{(3)} - x_1|, x_2^{(3)} - x_2|) = \max\left(\left|\frac{7}{8}-1\right|, \left|\frac{7}{8}-1\right|\right) = \frac{1}{8}.$$

In this way $\|\mathbf{e}_{k+1}\|_\infty = \frac{1}{2}\|\mathbf{e}_k\|_\infty$, *where* $\mathbf{e}_k$ *is the error in step* $k \geq 0$. *Iterating, we see that for the* k*-th Jacobi iteration the convergence rate is* $\left(\frac{1}{2}\right)^k$*:*

$$\|\mathbf{e}_k\|_\infty = \frac{1}{2}\|\mathbf{e}_{k-1}\|_\infty = \left(\frac{1}{2}\right)^2\|\mathbf{e}_{k-2}\|_\infty = \ldots = \left(\frac{1}{2}\right)^k\|\mathbf{e}_0\|_\infty = \left(1/2\right)^k.$$

Gauss-Seidel iteration

We start again with an initial approximation of the solution of the form

$$\mathbf{x} \approx \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right),$$

then using the fact that the first row in the k-th Jacobi iteration gives $x_1^{(k+1)}$ and in the $i+1$-th row $x_1^{(k+1)}, \ldots, x_i^{(k+1)}$ are already computed values given on the left hand sides of the previous (first i) rows. The idea in the Gauss-Seidel procedure is: that, in the very same iteration step, one simultaneously inserts these previously computed values. More specifically the Gauss-Seidel

iteration steps are given by:

$$\begin{cases} x_1^{(k+1)} = \dfrac{-1}{a_{11}}[a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} - b_1] \\ \\ x_2^{(k+1)} = \dfrac{-1}{a_{22}}[a_{21}x_1^{(k+1)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2] \\ \ldots \\ x_{n-1}^{(k+1)} = \dfrac{-1}{a_{n-1,n-1}}[a_{n-1,1}x_1^{(k+1)} + \ldots + a_{n-1,n-2}x_{n-2}^{(k+1)} + a_{n-1,n}x_n^{(k)} - b_{n-1}] \\ \\ x_n^{(k+1)} = \dfrac{-1}{a_{nn}}[a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + \ldots + a_{n,n-1}x_{n-1}^{(k+1)} - b_n], \end{cases}$$

or in a compact way in *Gauss-Seidel coordinates*, as

$$A\mathbf{x} = \mathbf{b} \Longleftrightarrow \sum_{j=1}^{n} a_{ij}x_j = b_i \Longleftrightarrow \sum_{j=1}^{i} a_{ij}x_j + \sum_{j=i+1}^{n} a_{ij}x_j = b_i. \tag{4.2.8}$$

Therefore the iterative form for the Gauss-Seidel method is given by

$$\begin{cases} \sum_{j=1}^{i} a_{ij}x_j^{(k+1)} = -\sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i \Longleftrightarrow \\ a_{ii}x_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i. \end{cases} \tag{4.2.9}$$

Example 4.6. *We consider the same example as above:* $A\mathbf{x} = \mathbf{b}$ *with*

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \textit{and} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Recall the Jacobi iteration system

$$\begin{cases} 2x_1^{(k+1)} = x_2^{(k)} + 1, \\ 2x_2^{(k+1)} = x_1^{(k)} + 1. \end{cases} \tag{4.2.10}$$

The corresponding Gauss-Seidel *iteration system reads as follows:*

$$\begin{cases} 2x_1^{(k+1)} = x_2^{(k)} + 1, \\ 2x_2^{(k+1)} = x_1^{(k+1)} + 1. \end{cases} \tag{4.2.11}$$

We choose the same initial values for x_1 *and* x_2 *as in the Jacobi iterations, i.e.* $x_1^{(0)} = 0$, *and* $x_2^{(0)} = 0$. *Now the first equation in* (4.2.11)*:*

$$2x_1^{(1)} = x_2^{(0)} + 1 \Longrightarrow x_1^{(1)} = \frac{1}{2}.$$

Inserting this value of $x_1^{(1)} = \frac{1}{2}$ *into the second equation in* (4.2.11) *yields*

$$2x_2^{(1)} = x_1^{(1)} + 1 \Longrightarrow 2x_2^{(1)} = \frac{1}{2} + 1 \Longrightarrow x_2^{(1)} = \frac{3}{4}.$$

Below we list the first few iteration steps for this Gauss-Seidel approach:

k	$x_1^{(k)}$	$x_2^{(k)}$
0	0	0
1	1/2	3/4
2	7/8	15/16
3	31/32	63/64

Obviously $\lim\limits_{k\to\infty} x_1^{(k)} = \lim\limits_{k\to\infty} x_2^{(k)} = 1$. *Now with* $\|\mathbf{e}_k\|_\infty = \max\limits_{i=1,2} |x_i^{(k)} - x_i|$, *we get the successive iteration errors:*

$$\|\mathbf{e}_1\|_\infty = \max(|x_1^{(1)} - x_1|, |x_2^{(1)} - x_2|) = \max\left(\left|\frac{1}{2} - 1\right|, \left|\frac{3}{4} - 1\right|\right) = \frac{1}{2}$$

$$\|\mathbf{e}_2\|_\infty = \max\left(\left|\frac{7}{8} - 1\right|, \left|\frac{15}{16} - 1\right|\right) = \frac{1}{8}, \quad \|\mathbf{e}_3\|_\infty = \max\left(\frac{1}{32}, \frac{1}{64}\right) = \frac{1}{32}.$$

Thus for the Gauss-Seidel iteration $\|\mathbf{e}_{k+1}\|_\infty = \frac{1}{4}\|\mathbf{e}_k\|_\infty$, *where* $\mathbf{e}_k$ *is the error for step* k, *and hence we can conclude that the Gauss-Seidel method converges faster than the Jacobi method:*

$$\|\mathbf{e}_k\|_\infty = \frac{1}{4}\|\mathbf{e}_{k-1}\|_\infty = \left(\frac{1}{4}\right)^2 \|\mathbf{e}_{k-2}\|_\infty = \cdots = \left(\frac{1}{4}\right)^k \|\mathbf{e}_0\|_\infty = \left(\frac{1}{4}\right)^k.$$

The successive over-relaxation method (S.O.R.).
The S.O.R. method is a modified version of the Gauss-Seidel iteration. The iteration procedure is given by

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\Big[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big] \qquad (4.2.12)$$

For $\omega > 1$ the method is called an *over-relaxation method* and if $0 < \omega < 1$, it is referred to as an *under-relaxation method*. In the *S.O.R. coordinates* we have

$$a_{ii}x_i^{(k+1)} = a_{ii}x_i^{(k)} - \omega\Big(\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i}^{n} a_{ij}x_j^{(k)} - b_i\Big). \qquad (4.2.13)$$

Remark 4.4. *Note that for* $\omega = 1$, (4.2.13) *gives the Gauss-Seidel method.*

Abstraction of iterative methods
In our procedures we have considered $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} = B\mathbf{x} + \mathbf{c}$ as equivalent linear systems of equations, where B is the iteration matrix and $\mathbf{x}_{k+1} = B\mathbf{x}_k + \mathbf{c}$.

Some advantages of iterative methods over direct methods are:
(i) Iterative methods are faster (depends on B, accuracy is required)
(ii) Iterative methods require less memory (sparsity of A can be preserved).

By a spars matrix we mean a matrix with most zero elements, i.e. a matrix that has very few non-zero elements.

Questions: For a qualitative study of an iteration procedure the following questions are of vital interest;
(Q1) For a given A, what is a good choice for B?
(Q2) When does $\mathbf{x}^{(k)} \to \mathbf{x}$?
(Q3) What is the rate of convergence?
The error at step k is $\mathbf{e}_k = \mathbf{x}^{(k)} - \mathbf{x}$ and that of step $(k+1)$ is $\mathbf{e}_{k+1} = \mathbf{x}^{(k+1)} - \mathbf{x}$. Then we have $\mathbf{e}_{k+1} = \mathbf{x}^{(k+1)} - \mathbf{x} = (B\mathbf{x}^{(k)} + \mathbf{c}) - (B\mathbf{x} + \mathbf{c}) = B(\mathbf{x}^{(k)} - \mathbf{x}) = B\mathbf{e}_k$. Iterating, we have

$$\mathbf{e}_k = B\mathbf{e}_{k-1} = B \cdot B\mathbf{e}_{k-2} = B^3\mathbf{e}_{k-3} = \ldots = B^k\mathbf{e}_{k-k} = B^k\,\mathbf{e}_0.$$

Thus we have shown that $\mathbf{e}_k = B^k \mathbf{e}_0$. Let now

$$L = \begin{pmatrix} 0 & \dots & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n,n-1} & 0 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & a_{n-1,n} \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

and

$$D = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{nn} \end{pmatrix},$$

then $A = L + D + U$. We can rewrite $A\mathbf{x} = \mathbf{b}$ as $(L + D + U)\mathbf{x} = \mathbf{b}$ then $D\mathbf{x} = -(L + U)\mathbf{x} + \mathbf{b}$, and we may reformulate the iterative methods as follows.

Jacobi's method

$$D\mathbf{x}^{(k+1)} = -(L + U)\mathbf{x}^{(k)} + \mathbf{b} \Rightarrow B_J = -D^{-1}(L + U),$$

where B_J is the *Jacobi iteration matrix*.

Example 4.7. *We consider the previous example and write the linear system*

in matrix form as $\mathbf{x} = B_J\mathbf{x} + \mathbf{c}$*, i.e.*

$$\begin{cases} 2x_1 - x_2 = 1 \\ -x_1 + 2x_2 = 1 \end{cases} \Rightarrow \begin{cases} x_1 = \frac{1}{2}x_2 + \frac{1}{2} \\ \\ x_2 = \frac{1}{2}x_1 + \frac{1}{2} \end{cases} \textit{ which in matrix form is}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \textit{ where}$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, B_J = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} \textit{ and } \mathbf{c} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Example 4.8. *Determine the same matrix* B_J *by the formula:* $B_J = -D^{-1}(L + U)$*,*

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, L = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, U = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}, D = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

We can easily see that

$$D^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and thus

$$B_J = -D^{-1}(L + U) = -\frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Gauss-Seidel's method

As in the Jacobi case, we may write $A\mathbf{x} = \mathbf{b}$ as $(L + D + U)\mathbf{x} = \mathbf{b}$ but now we choose $(L + D)\mathbf{x} = -U\mathbf{x} + \mathbf{b}$. Similar to the previous procedure we then have $(L + D)\mathbf{x}^{(k+1)} = -U\mathbf{x}^{(k)} + \mathbf{b}$, and then $B_{GS} = -(L + D)^{-1}U$, where B_{GS} stands for the *Gauss-Seidel iteration matrix.*

Relaxation
Gauss-Seidel corresponds to $(L+D)\mathbf{x} = -U\mathbf{x} + \mathbf{b}$, thus the iteration procedure is:

$$D\mathbf{x}^{(k+1)} = D\mathbf{x}^{(k)} - [L\mathbf{x}^{(k+1)} + (D+U)\mathbf{x}^{(k)} - \mathbf{b}].$$

Now we define the general relaxation procedure

$$\mathbf{x}^{(k+1)} = (1-\omega)\mathbf{x}^{(k)} + \omega D^{-1}\Big(\mathbf{b} - L\mathbf{x}^{(k+1)} - U\mathbf{x}^{(k)}\Big), \tag{4.2.14}$$

where the coefficient of ω, in the second term, is just the right hand side of Gauss-Seidel's method: the term in the bracket in (4.2.12). We may rewrite (4.2.14) as

$$(\omega L + D)\mathbf{x}^{(k+1)} = [(1-\omega)D - \omega U]\mathbf{x}^{(k)} + \omega\mathbf{b},$$

where ω is the relaxation parameter, ($\omega = 1$ gives the Gauss-Seidel iteration). Thus the *relaxation iteration matrix* is:

$$B_\omega = (\omega L + D)^{-1}[(1-\omega)D - \omega U].$$

4.3 Exercises

Problem 4.2. *Illustrate the LU factorization for the matrix*

$$A = \begin{bmatrix} 1 & 3 & 2 \\ -2 & -6 & 1 \\ 2 & 5 & 7 \end{bmatrix}.$$

Problem 4.3. *Solve $A^4 x = b$ for*

$$A = \begin{bmatrix} -1 & 2 \\ 2 & -3 \end{bmatrix} \qquad b = \begin{bmatrix} 144 \\ -233 \end{bmatrix}$$

Problem 4.4. *Find the unique $LD\tilde{U}$ factorization for the matrix*

$$A = \begin{bmatrix} 1 & 1 & -3 \\ 0 & 1 & 1 \\ 3 & -1 & 1 \end{bmatrix}.$$

Hint: First find the LU factorization of A. Then D is a diagonal matrix with its diagonal elements being those of U. $\tilde{U}$ is an upper triangular matrix obtained from U by replacing the diagonal elements of U by ones.

Problem 4.5. *Show that every orthogonal 2×2 matrix is of the form*

$$A_1 = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \qquad or \qquad A_2 = \begin{bmatrix} c & s \\ s & -c \end{bmatrix},$$

where $c^2 + s^2 = 1$

Problem 4.6. *Find the LU factorization for the matrix A*

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 2 & 4 & 1 \\ 0 & 0 & 3 & 5 \end{bmatrix}.$$

Problem 4.7. *Solve the following system of equations*

$$\begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

using the following iteration methods and a starting value of $u^0 = (0, 0)^T$.

(a) Jacobi Method.

(b) Gauss-Seidel Method.

(c) Optimal SOR (you must compute the optimal value of $\omega = \omega_0$ first).

Problem 4.8. *Prove strict diagonal dominance* (4.2.3) *implies convergence of Jacobi and Gauss-Seidel methods.*

Hint: One may use fix-point iterations.

Problem 4.9. *A is an orthogonal matrix if $A^T A = I$ (I being the identity matrix). Show that for any vector $\mathbf{v}$ with $||\mathbf{v}||_2 = 1$, the matrix $I - \mathbf{v}\mathbf{v}^T$ is orthogonal.*

Problem 4.10. *Suppose that $\det(a_{ij}) = 0$, but $a_{ii} \neq 0$ for all i. Prove that the Jacobi's method does not converge.*

Problem 4.11. *Consider the $N \times N$ matrix $A = (a_{ij})$ defined by*

$$a_{ij} = \begin{cases} -1 & |i-j| = 1 \\ 1 & i = j = 1 \\ 1 & i = j = N \\ 2 & 1 < i = j < N \end{cases}$$

a) Show that A is positive semi-definite, i.e. $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbf{R}^N$.
b) Show that $\mathbf{x}^T A\mathbf{x} = 0$ if and only if $\mathbf{x}$ is a constant multiple of the one vector $(1, 1, \ldots, 1)^T$.

Problem 4.12. *We define the L_p-norm ($p = 1, 2, \infty$) of a vector $\mathbf{v}$ as*

$$\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$$
$$\|\mathbf{v}\|_2 = \Big(\sum_{i=1}^{n} |v_i|^2 \Big)^{1/2}$$
$$\|\mathbf{v}\|_\infty = \max\{|v_i|;\ i = 1, \ldots, n\}.$$

Show that

$$\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1.$$

Chapter 5

Two-point boundary value problems

In this chapter we focus on finite element approximation procedure for two-point *boundary value problems* (BVPs) of Dirichlet, Neumann and mixed type. For each case we formulate a corresponding *variational formulation* (VF) and a *minimization problem* (MP) and prove that to solve the boundary value problem is "equivalent" to solve a VF problem which in its turn is equivalent to solve a minimization problem MP, i.e,

$$(BVP) \text{ " } \Longleftrightarrow \text{ " } (VF) \Longleftrightarrow (MP).$$

We also prove *a priori* and *a posteriori* error estimates. The $\Longleftarrow$ in the equivalence " $\Longleftrightarrow$ " is subject to a regularity requirement on the solution up to the order of the underlying PDE.

5.1 A Dirichlet problem

Assume that a horizontal elastic bar which occupies the interval $I := [0, 1]$, is fixed at the end-points. Let $u(x)$ denote the displacement of the bar at a point $x \in I$, $a(x)$ be the *modulus of elasticity*, and $f(x)$ a given *load function*, then one can show that u satisfies the following boundary value problem for Poisson's equation

$$(BVP)_1 \qquad \begin{cases} -\Big(a(x)u'(x)\Big)' = f(x), \quad 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{5.1.1}$$

We assume that $a(x)$ is piecewise continuous in $(0,1)$, bounded for $0 \leq x \leq 1$ and $a(x) > 0$ for $0 \leq x \leq 1$.
Let $v(x)$ and its derivative $v'(x), x \in I$, be square integrable functions, that is: v, $v' \in L_2(0,1)$, and define the L_2-based Sobolev space

$$H_0^1(0,1) = \Big\{ v(x) : \int_0^1 (v(x)^2 + v'(x)^2)dx < \infty, \quad v(0) = v(1) = 0 \Big\}. \quad (5.1.2)$$

The variational formulation (VF). We multiply the equation in $(\mathrm{BVP})_1$ by a so called test function $v(x) \in H_0^1(0,1)$ and integrate over $(0,1)$ to obtain

$$-\int_0^1 (a(x)u'(x))'v(x)dx = \int_0^1 f(x)v(x)dx. \quad (5.1.3)$$

By integration by parts we get

$$-\Big[a(x)u'(x)v(x)\Big]_0^1 + \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx. \quad (5.1.4)$$

Now since $v(0) = v(1) = 0$ we have thus the obtained the *variational formulation* for the problem (5.1.1) as follows: find $u(x) \in H_0^1$ such that

$$(\mathrm{VF})_1 \qquad \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \quad \forall v(x) \in H_0^1. \quad (5.1.5)$$

In other words we have shown that if u satisfies $(\mathrm{BVP})_1$, then u also satisfies the $(\mathrm{VF})_1$ above. We write this as $(\mathrm{BVP})_1 \Longrightarrow (\mathrm{VF})_1$. Now the question is whether the reverse implication is true, i.e. under which conditions can we deduce the implication $(\mathrm{VF})_1 \Longrightarrow (\mathrm{BVP})_1$? It appears that this question has an affirmative answer, *provided that the solution u to $(\mathrm{VF})_1$ is twice differentiable*. Then, modulo this regularity requirement, the two problems are indeed equivalent. We prove this in the following theorem.

Theorem 5.1. *The following two properties are equivalent*

i) u satisfies $(\mathrm{BVP})_1$

ii) u is twice differentiable and satisfies $(\mathrm{VF})_1$.

Proof. We have already shown $(\mathrm{BVP})_1 \Longrightarrow (\mathrm{VF})_1$. It remains to show that $(\mathrm{VF})_1 \Longrightarrow (\mathrm{BVP})_1$. Integrating by parts on the left hand side in (5.1.5), *assuming that u is twice differentiable*, and using $v(0) = v(1) = 0$ we return to the relation (5.1.3):

$$-\int_0^1 (a(x)u'(x))'v(x)dx = \int_0^1 f(x)\ v(x)dx, \qquad \forall v(x) \in H_0^1 \tag{5.1.6}$$

which can be rewritten as

$$\int_0^1 \Big\{ -\Big(a(x)u'(x)\Big)' - f(x)\Big\}v(x)dx = 0, \qquad \forall v(x) \in H_0^1. \tag{5.1.7}$$

We *claim* that (5.1.7) implies that

$$-\Big(a(x)u'(x)\Big)' - f(x) \equiv 0, \qquad \forall x \in (0,1). \tag{5.1.8}$$

Suppose not. Then there exists at least one point $\xi \in (0,1)$, such that

$$-\Big(a(\xi)u'(\xi)\Big)' - f(\xi) \neq 0, \tag{5.1.9}$$

where we may assume, without loss of generality, that

$$-\Big(a(\xi)u'(\xi)\Big)' - f(\xi) > 0 \quad (\text{or } < 0). \tag{5.1.10}$$

Thus, assuming that $f \in C(0,1)$ and $a \in C^1(0,1)$, by continuity, $\exists \delta > 0$, such that in a δ-neighborhood of ξ,

$$g(x) := -\Big(a(x)u'(x)\Big)' - f(x) > 0, \qquad \text{for all} \quad x \in (\xi - \delta, \xi + \delta). \tag{5.1.11}$$

Now if we take the test function $v(x)$ in (5.1.7) as the *hat-function* $v^*(x) > 0$, with $v^*(\xi) = 1$ and the support $I_\delta := (\xi - \delta, \xi + \delta)$, see Figure 5.1. Then $v^*(x) \in H_0^1$ and

$$\int_0^1 \Big\{ -\Big(a(x)u'(x)\Big)' - f(x)\Big\}v^*(x)dx = \int_{I_\delta} \underbrace{g(x)}_{>0}\underbrace{v^*(x)}_{>0}dx > 0.$$

This contradicts (5.1.7). Thus, our claim is true and the proof is complete. □

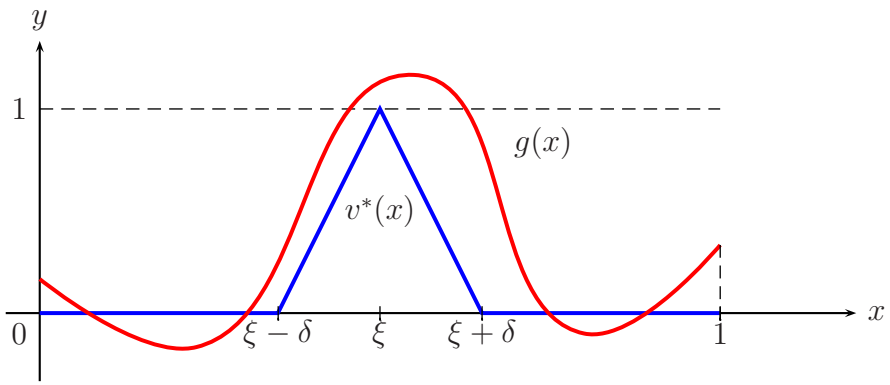


Figure 5.1: The hat function $v^*(x)$ over the interval $(\xi - \delta, \xi + \delta)$.

Corollary 5.1. *(i) If $f(x)$ is continuous and $a(x)$ is continuously differentiable, i.e. $f \in C(0,1)$ and $a \in C^1(0,1)$, then (BVP) and (VF) have the same solution.*
(ii) If $a(x)$ is discontinuous, then (BVP) is not always well-defined but (VF) has still a meaning. Therefore (VF) covers a larger set of data than (BVP).
(iii) More important: in (VF) $u \in C^1(0,1)$,while (BVP) is formulated for u having 2 derivatives, e.g. $u \in C^2(0,1)$.

The minimization problem. For the problem (5.1.1), we may formulate yet another equivalent problem, viz:

Find $u \in H_0^1$ such that $F(u) \leq F(w)$, $\forall w \in H_0^1$, where $F(w)$ is the total potential energy of the displacement $w(x)$, given by

$$\text{(MP)} \qquad F(w) = \underbrace{\frac{1}{2}\int_0^1 a(w')^2 dx}_{\text{Internal (elastic) energy}} - \underbrace{\int_0^1 fw dx}_{\text{Load potential}}. \tag{5.1.12}$$

This means that the solution u minimizes the energy functional $F(w)$.

Below we show that the above minimization problem is equivalent to the variational formulation $(\text{VF})_1$ and hence also to the boundary value problem $(\text{BVP})_1$.

Theorem 5.2. *The following two properties are equivalent*

a) u satisfies the variational formulation $(\text{VF})_1$

b) u is the solution for the minimization problem (MP)

i.e.

$$F(u) \le F(w), \forall w \in H_0^1 \Longleftrightarrow \int_0^1 au'v'dx = \int_0^1 fvdx, \quad \forall v \in H_0^1. \quad (5.1.13)$$

Proof. ($\Longleftarrow$): First we show that the variational formulation $(\mathrm{VF})_1$ implies the minimization problem (MP). To this end, for $w \in H_0^1$ we let $v = w - u$, then $v \in H_0^1$ and

$$\begin{aligned} F(w) = F(u+v) &= \frac{1}{2}\int_0^1 a\Big((u+v)'\Big)^2 dx - \int_0^1 f(u+v)dx = \\ &= \underbrace{\frac{1}{2}\int_0^1 2au'v'dx}_{(i)} + \underbrace{\frac{1}{2}\int_0^1 a(u')^2dx}_{(ii)} + \frac{1}{2}\int_0^1 a(v')^2dx \\ &\quad - \underbrace{\int_0^1 fudx}_{(iii)} - \underbrace{\int_0^1 fvdx.}_{(iv)} \end{aligned}$$

Now using $(\mathrm{VF})_1$ we have $(i) - (iv) = 0$. Further by the definition of the functional F, $(ii) - (iii) = F(u)$. Thus

$$F(w) = F(u) + \frac{1}{2}\int_0^1 a(x)(v'(x))^2dx, \quad (5.1.14)$$

and since $a(x) > 0$ we get $F(w) \ge F(u)$.
($\Longrightarrow$): Next we show that the minimization problem (MP) implies the variational formulation $(\mathrm{VF})_1$. To this end, assume that $F(u) \le F(w)\ \forall w \in H_0^1$, and for an arbitrary function $v \in H_0^1$, set $g(\varepsilon, v) = F(u+\varepsilon v)$, then by (MP), g (as a function of ε) has a *minimum* at $\varepsilon = 0$. In other words $\frac{\partial}{\partial\varepsilon}g(\varepsilon, v)\Big|_{\varepsilon=0} = 0$. We have that

$$\begin{aligned} g(\varepsilon, v) = F(u+\varepsilon v) &= \frac{1}{2}\int_0^1 a\Big((u+\varepsilon v)'\Big)^2 dx - \int_0^1 f(u+\varepsilon v)dx = \\ &= \frac{1}{2}\int_0^1 \{a(u')^2 + a\varepsilon^2(v')^2 + 2a\varepsilon u'v'\}dx - \int_0^1 fudx - \varepsilon\int_0^1 fvdx. \end{aligned}$$

Now we compute the derivative $g'_\varepsilon(\varepsilon, v)$,

$$g'_\varepsilon(\varepsilon, v) = \frac{1}{2}\int_0^1 \{2a\varepsilon(v')^2 + 2au'v'\}dx - \int_0^1 fvdx, \tag{5.1.15}$$

where $g'_\varepsilon\Big|_{(\varepsilon=0)} = 0$, yields

$$\int_0^1 au'v'dx - \int_0^1 fvdx = 0, \tag{5.1.16}$$

which is the variational formulation $(\mathrm{VF})_1$. Hence, we conclude that $F(u) \leq F(w)$, $\forall w \in H_0^1 \Longrightarrow (\mathrm{VF})_1$, and the proof is complete. □

We summarize the two theorems in short as

Corollary 5.2.

$$(BVP)_1\text{"} \Longleftrightarrow \text{"}(VF)_1 \Longleftrightarrow (MP).$$

Recall that " $\Longleftrightarrow$ " is a conditional equivalence, requiring, e.g. u to be twice differentiable, for the reverse implication.

5.2 A mixed Boundary Value Problem

Obviously changing the boundary conditions would require changes in the variational formulation. This can be seen, e.g. in deriving the variational formulation corresponding to the following mixed boundary value problem: find u such that

$$(\mathrm{BVP})_2 \qquad \begin{cases} -\Big(a(x)u'(x)\Big)' = f(x), \quad 0 < x < 1 \\ u(0) = 0, \qquad a(1)u'(1) = g_1. \end{cases} \tag{5.2.1}$$

As usual, we multiply the equation by a suitable test function $v(x)$, and integrate over the interval $(0, 1)$. Note that this time the test function satisfies only $v(0) = 0$. This is due to the fact that now $u(1)$ is not given, and to get an approximate value of u at $x = 1$, we need to supply a test function (a half-hat-function) at $x = 1$. So we have

$$-\int_0^1 (a(x)u'(x))'v(x)dx = \int_0^1 f(x)v(x)dx, \tag{5.2.2}$$

and by integration by parts we have

$$-\Big[a(x)u'(x)v(x)\Big]_0^1 + \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx. \qquad (5.2.3)$$

Now using the boundary data $a(1)u'(1) = g_1$ and $v(0) = 0$ we get

$$\int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx + g_1v(1), \quad \forall v \in \tilde{H}_0^1, \qquad (5.2.4)$$

where

$$\tilde{H}_0^1 = \{v(x) : \int_0^1 (v(x)^2 + v'(x)^2)dx < \infty, \text{ such that } v(0) = 0\}. \qquad (5.2.5)$$

Hence, (5.2.4) yields the variational formulation: find $u \in \tilde{H}_0^1$ such that

$$(\mathrm{VF})_2 \qquad \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx + g_1v(1), \quad \forall v \in \tilde{H}_0^1.$$

Now, as in the *Dirichlet* case, we want to show that

Theorem 5.3. $(\mathrm{BVP})_2 \Longleftrightarrow (\mathrm{VF})_2$, *in the sense that the two problems have the same solution.*

Proof. That $(\mathrm{BVP})_2 \Longrightarrow (\mathrm{VF})_2$, is already shown by (5.2.2)-(5.2.4). To prove $(\mathrm{VF})_2 \Longrightarrow (\mathrm{BVP})_2$: that a solution of the variational problem $(\mathrm{VF})_2$ is also a solution of the two-point boundary value problem $(\mathrm{BVP})_2$, we have to prove the following claims:

(i) the solution satisfies the differential equation

(ii) the solution satisfies the boundary conditions

We start with $(VF)_2$ and perform a reversed order integration by parts to get

$$\int_0^1 a(x)u'(x)v'(x)dx = [a(x)u'(x)v(x)]_0^1 - \int_0^1 (a(x)u'(x))'v(x)\ dx. \qquad (5.2.6)$$

Since $v(0) = 0$, we get

$$\int_0^1 a(x)u'(x)v'(x)dx = a(1)u'(1)v(1) - \int_0^1 (a(x)u'(x))'v(x)dx \qquad (5.2.7)$$

Thus the variational formulation $(\mathrm{VF})_2$ can be rewritten as

$$-\int_0^1 (a(x)u'(x))'v(x)dx + a(1)u'(1)v(1) = \int_0^1 f(x)v(x)dx + g_1 v(1). \quad (5.2.8)$$

The equation (5.2.8) is valid for every $v(x) \in \tilde{H}_0^1(0,1)$, including the special class of test functions $v(x)$ with $v(0) = v(1) = 0$ as in the Dirichlet problem: $-(au')' = f,\ u(0) = u(1) = 0$. This is simply because $H_0^1(0,1) \subset \tilde{H}_0^1(0,1)$. Consequently choosing $v(1) = 0$, (5.2.8) is reduced to

$$-\int_0^1 (a(x)u'(x))'v(x)dx = \int_0^1 f(x)v(x)dx, \qquad \forall v(x) \in H_0^1 \qquad (5.2.9)$$

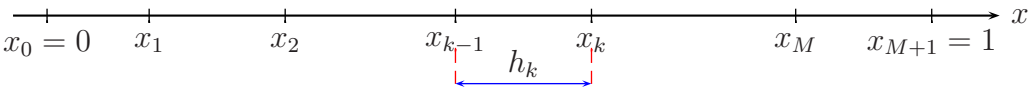
Now, as in the Dirichlet case, the variational formulation (5.2.9) gives the differential equation in (5.2.1) and hence the claim (i) is true. Inserting $(au')' = f$ into (5.2.8) we get $g_1 v(1) = a(1)u'(1)v(1)$. Choosing $v(1) \neq 0$, e.g. $v(1) = 1$, gives the boundary condition $g_1 = a(1)u'(1)$. The boundary condition $u(0) = 0$ follows directly from the definition of $\tilde{H}_0^1(0,1)$, and the proof is complete. □

Remark 5.1. *i) The Dirichlet boundary condition is called an essential boundary condition and is "strongly imposed" in the trial/test function space: Enforced explicitly to the trial and test functions in (VF).*

ii) Neumann and Robin boundary conditions are called natural boundary conditions. These boundary conditions are automatically satisfied, within the corresponding variational formulations, and are therefore weakly imposed.

5.3 The finite element method (FEM)

We now formulate the finite element procedure for boundary value problems. To this end we let $\mathcal{T}_h = \{0 = x_0 < x_1 < \ldots < x_M < x_{M+1} = 1\}$ be a partition of the interval $I = [0, 1]$ into subintervals $I_k = [x_{k-1}, x_k]$ and set $h_k = x_k - x_{k-1}$. Define the piecewise *constant* function $h(x) := x_k - x_{k-1} = h_k$



for $x \in I_k$. Let $\mathcal{C}(I, P_1(I_k))$ denote the set of all continuous piecewise linear functions on $\mathcal{T}_h$ (continuous in the whole interval I, linear on each subinterval I_k), and define

$$V_h^{(0)} = \{v : v \in \mathcal{C}(I, P_1(I_k)), \quad v(0) = v(1) = 0\}. \tag{5.3.1}$$

Note that $V_h^{(0)}$ is a finite dimensional ($dim(V_h^{(0)} = M$) subspace of

$$H_0^1 = \Big\{v(x) : \int_0^1 (v(x)^2 + v'(x)^2)dx < \infty, \quad \text{and} \quad v(0) = v(1) = 0\Big\}. \tag{5.3.2}$$

Continuous Galerkin of degree 1, cG(1). A finite element formulation for our Dirichlet boundary value problem $(\text{BVP})_1$ is given by: find $u_h \in V_h^{(0)}$ such that the following discrete variational formulation holds true

$$\text{(FEM)} \qquad \int_0^1 a(x)u_h'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \quad \forall v \in V_h^{(0)}. \tag{5.3.3}$$

Note that the finite element method (FEM) is a finite dimensional version of the variational formulation $(\text{VF})_1$. Here the test functions are in a finite dimensional subspace $V_h^{(0)}$, of H_0^1, spanned by the hat-functions, $\varphi_j(x)$, $j = 1, \ldots, M$. Now the purpose is to *estimate the error* arising in approximating the solution for $(BVP)_1$ (a similar procedure is applied for $(BVP)_2$) by functions in $V_h^{(0)}$. To this end we need to introduce some measuring environment for the error. We recall the definition of L_p-norms:

L_p-norm	$\|v\|_{L_p} = \Big(\int_0^1 \|v(x)\|^p dx \Big)^{1/p}, \quad 1 \le p < \infty$
L_∞-norm	$\|v\|_{L_\infty} = \sup_{x\in[0,1]} \|v(x)\|,$
Weighted L_2-norm	$\|v\|_a = \Big(\int_0^1 a(x)\|v(x)\|^2 dx \Big)^{1/2}, \quad a(x) > 0$
Energy-norm	$\|v\|_E = \Big(\int_0^1 a(x)\|v'(x)\|^2 dx \Big)^{1/2},$
Note that	$\|v\|_E = \|v'\|_a.$

$\|v\|_E$ describes the *internal elastic energy* of the elastic bar modeled by our Dirichlet boundary value problem $(BVP)_1$.

5.4 Error estimates in the energy norm

We shall study two types of error estimates:

i) An *a priori error estimate*; where a certain norm of the error is estimated by some norm of the *exact solution* $u(x)$. In such estimates the error analysis gives information about the size of the error, depending on the (unknown) exact solution u, before any computational steps.

ii) An *a posteriori error estimate*; where the error is estimated by some norm of the *residual* of the approximate solution. Recall that the residual is the difference between the left and right hand side in the equation when the exact solution $u(x)$ is replaced by the approximate solution $u_h(x)$. Hence a posteriori error estimates give quantitative information about the size of the error after that the approximate solution $u_h(x)$ has been computed.

   Below, first we shall prove a qualitative theorem which shows that the finite element solution is the best approximate solution to the Dirichlet problem in the energy norm.

Theorem 5.4. *Let $u(x)$ be the solution to the Dirichlet boundary value problem*

$$\textit{(BVP)} \qquad \begin{cases} -\Big(a(x)u'(x)\Big)' = f(x), & 0 < x < 1 \\ u(0) = 0 \quad u(1) = 0, \end{cases} \tag{5.4.1}$$

and $u_h(x)$ its finite element approximation given by (5.3.3)*. Then we have*

$$\|u - u_h\|_E \leq \|u - v\|_E, \qquad \forall v(x) \in V_h^{(0)}. \tag{5.4.2}$$

This means that the finite element solution $u_h \in V_h^{(0)}$ is the best approximation of the solution u, in the energy norm, by functions in $V_h^{(0)}$.

Proof. Recall the variational formulation associated with the problem (5.4.1):

$$\text{(VF)} \quad \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \qquad \forall v \in H_0^1. \tag{5.4.3}$$

We take an arbitrary $v \in V_h^{(0)}$, then by the definition of the energy norm

$$\begin{aligned}
\|u - u_h\|_E^2 &= \int_0^1 a(x)(u'(x) - u_h'(x))^2\, dx \\
&= \int_0^1 a(x)(u'(x) - u_h'(x))(u'(x) - v'(x) + v'(x) - u_h'(x))dx \\
&= \int_0^1 a(x)(u'(x) - u_h'(x))(u'(x) - v'(x))dx \\
&\quad + \int_0^1 a(x)(u'(x) - u_h'(x))(v'(x) - u_h'(x))\, dx
\end{aligned} \tag{5.4.4}$$

By Galerkin orthogonality, the last integral is identically zero. Here is an argument: since $v - u_h \in V_h^{(0)} \subset H_0^1$, by the variational formulation (5.4.3) we can write

$$\int_0^1 a(x)u'(x)(v'(x) - u_h'(x))dx = \int_0^1 f(x)(v(x) - u_h(x))\, dx, \tag{5.4.5}$$

and its finite element counterpart, see (5.3.3),

$$\int_0^1 a(x)u_h'(x)(v'(x) - u_h'(x))dx = \int_0^1 f(x)(v(x) - u_h(x))\, dx. \tag{5.4.6}$$

Subtracting these two relations, the last line of the estimate (5.4.4) above

will vanish, so that we end up with

$$\begin{aligned}\|u-u_h\|_E^2 &= \int_0^1 a(x)(u'(x)-u_h'(x))(u'(x)-v'(x))dx \\ &= \int_0^1 a(x)^{\frac{1}{2}}(u'(x)-u_h'(x))a(x)^{\frac{1}{2}}(u'(x)-v'(x))dx \\ &\leq \Big(\int_0^1 a(x)(u'(x)-u_h'(x))^2dx\Big)^{\frac{1}{2}}\Big(\int_0^1 a(x)(u'(x)-v'(x))^2dx\Big)^{\frac{1}{2}} \\ &= \|u-u_h\|_E\cdot\|u-v\|_E,\end{aligned} \tag{5.4.7}$$

where, in the last estimate, we used Cauchy-Schwarz inequality. Thus

$$\|u-u_h\|_E \leq \|u-v\|_E, \qquad \forall v\in V_h^{(0)}, \tag{5.4.8}$$

and the proof is complete. □

The next step is to show that there exists a function $v(x)\in V_h^{(0)}$ such that $\|u-v\|_E$ is not *too large*. The function that we have in mind is $\pi_h u(x)$: the *piecewise linear interpolant of* $u(x)$, introduced in Chapter 3. Recall the interpolation error estimates in L_p-norms:

Theorem 5.5. *Let $0=x_0<x_1<x_2<\ldots<x_M<x_{M+1}=1$ be a partition of $[0,1]$ and $h(x):=(x_{j+1}-x_j)$, $x\in(x_j,x_{j+1})$, $j=0,1,\ldots,M$, be the mesh function. Further, assume that $\pi_h v(x)$ is the piecewise linear interpolant of $v(x)$. Then, for each estimate below, there is an interpolation constant c_i such that*

$$\|\pi_h v-v\|_{L_p}\leq c_i\|h^2v''\|_{L_p} \quad 1\leq p\leq\infty \tag{5.4.9}$$

$$\|(\pi_h v)'-v'\|_{L_p}\leq c_i\|hv''\|_{L_p} \tag{5.4.10}$$

$$\|\pi_h v-v\|_{L_p}\leq c_i\|hv'\|_{L_p}. \tag{5.4.11}$$

Theorem 5.6. *[An a priori error estimate] Let u and u_h be the solutions of the Dirichlet problem (BVP) and the finite element problem (FEM), respectively. Then there exists an interpolation constant C_i, depending only on $a(x)$, such that*

$$\|u-u_h\|_E\leq C_i\|hu''\|_a. \tag{5.4.12}$$

Proof. By the general estimate (5.4.2) in Theorem 5.4, we have that

$$\|u - u_h\|_E \leq \|u - v\|_E, \qquad \forall v \in V_h^{(0)}. \tag{5.4.13}$$

Now since $\pi_h u(x) \in V_h^{(0)}$, we may take $v = \pi_h u(x)$ in (5.4.13) and use, e.g. the second estimate in the interpolation theorem above to get

$$\begin{aligned} \|u - u_h\|_E &\leq \|u - \pi_h u\|_E = \|u' - (\pi_h u)'\|_a \\ &\leq C_i \|hu''\|_a = C_i \Big(\int_0^1 a(x) h^2(x) u''(x)^2 \, dx \Big)^{1/2}, \end{aligned} \tag{5.4.14}$$

which is the desired result and the proof is complete. □

Remark 5.2. *Note that the interpolation theorem is not in the weighted norm. The $a(x)$ dependence of the interpolation constant C_i can be shown as follows:*

$$\begin{aligned} \|u' - (\pi_h u)'\|_a &= \Big(\int_0^1 a(x)(u'(x) - (\pi_h u)'(x))^2 \, dx \Big)^{1/2} \\ &\leq \Big(\max_{x\in[0,1]} a(x)^{1/2} \Big) \cdot \|u' - (\pi_h u)'\|_{L_2} \leq c_i \Big(\max_{x\in[0,1]} a(x)^{1/2} \Big) \|hu''\|_{L_2} \\ &= c_i \Big(\max_{x\in[0,1]} a(x)^{1/2} \Big) \Big(\int_0^1 h(x)^2 u''(x)^2 \, dx \Big)^{1/2} \\ &\leq c_i \frac{(\max_{x\in[0,1]} a(x)^{1/2})}{(\min_{x\in[0,1]} a(x)^{1/2})} \cdot \Big(\int_0^1 a(x) h(x)^2 u''(x)^2 \, dx \Big)^{1/2}. \end{aligned}$$

Thus

$$C_i = c_i \frac{(\max_{x\in[0,1]} a(x)^{1/2})}{(\min_{x\in[0,1]} a(x)^{1/2})}, \tag{5.4.15}$$

where c_i is the interpolation constant in the second estimate in Theorem 5.5.

Remark 5.3. *If the objective is to divide (0,1) into a fixed, finite, number of subintervals, then one can use the result of Theorem 5.6: to obtain an optimal (best possible) partition of (0,1); in the sense that: whenever $a(x)u''(x)^2$ gets large we compensate by making $h(x)$ smaller. This, however, "requires that the exact solution $u(x)$ is known" [1]. Now we shall study a posteriori error analysis, where instead of the unknown solution $u(x)$, "we use the known, computed approximate solution $u_h(x)$".*

[1]Note that when a is a given constant then, $-u''(x) = (1/a)f(x)$ is known.

Theorem 5.7 (An a posteriori error estimate)**.** *There is an interpolation constant c_i depending only on $a(x)$ such that the error in the finite element approximation of the Dirichlet boundary value problem* (5.4.1)*, satisfies*

$$\|e(x)\|_E \leq c_i \Big(\int_0^1 \frac{1}{a(x)} h^2(x) R^2(u_h(x)) dx \Big)^{1/2}, \tag{5.4.16}$$

where $e(x) = u(x) - u_h(x)$ and $R(u_h(x)) = f + (a(x)u_h'(x))'$ is the residual. Note that $e \in H_0^1$.

Proof. By the definition of the *energy norm* we have

$$\begin{aligned} \|e(x)\|_E^2 &= \int_0^1 a(x)(e'(x))^2 dx = \int_0^1 a(x)(u'(x) - u_h'(x))e'(x)dx \\ &= \int_0^1 a(x)u'(x)e'(x)dx - \int_0^1 a(x)u_h'(x)e'(x)dx \end{aligned} \tag{5.4.17}$$

Since $e \in H_0^1$ the variational formulation (VF) gives that

$$\int_0^1 a(x)u'(x)e'(x)dx = \int_0^1 f(x)e(x)dx. \tag{5.4.18}$$

Hence, we can write

$$\|e(x)\|_E^2 = \int_0^1 f(x)e(x)dx - \int_0^1 a(x)u_h'(x)e'(x)dx. \tag{5.4.19}$$

Adding and subtracting the interpolant $\pi_h e(x)$ and its derivative $(\pi_h e)'(x)$ to e and e' in the integrands above yields

$$\begin{aligned} \|e(x)\|_E^2 = &\int_0^1 f(x)(e(x) - \pi_h e(x))dx + \underbrace{\int_0^1 f(x)\pi_h e(x)dx}_{(i)} \\ &- \int_0^1 a(x)u_h'(x)(e'(x) - (\pi_h e)'(x))dx - \underbrace{\int_0^1 a(x)u_h'(x)(\pi_h e)'(x)dx}_{(ii)}. \end{aligned}$$

Since $u_h(x)$ is the solution of the (FEM) given by (5.3.3) and $\pi_h e(x) \in V_h^{(0)}$ we have that $-(ii) + (i) = 0$. Hence

$$\begin{aligned} \|e(x)\|_E^2 &= \int_0^1 f(x)(e(x) - \pi_h e(x))dx - \int_0^1 a(x)u_h'(x)(e'(x) - (\pi_h e)'(x))dx \\ &= \int_0^1 f(x)(e(x) - \pi_h e(x))dx - \sum_{k=1}^{M+1} \int_{x_{k-1}}^{x_k} a(x)u_h'(x)(e'(x) - (\pi_h e)'(x))dx. \end{aligned}$$

To continue we integrate by parts in the integrals in the summation above

$$
\begin{aligned}
&-\int_{x_{k-1}}^{x_k} a(x)u_h'(x)(e'(x)-(\pi_h e)'(x))dx \\
&= -\Big[a(x)u_h'(x)(e(x)-\pi_h e(x))\Big]_{x_{k-1}}^{x_k} + \int_{x_{k-1}}^{x_k} (a(x)u_h'(x))'(e(x)-\pi_h e(x))\,dx.
\end{aligned}
$$

Now, using $e(x_k) = \pi_h e(x_k)$, $k = 0, 1 \ldots, M+1$, where the x_k:s are the interpolation nodes, the boundary terms vanish and thus we end up with

$$
-\int_{x_{k-1}}^{x_k} a(x)u_h'(x)(e'(x)-(\pi_h e)'(x))dx = \int_{x_{k-1}}^{x_k} (a(x)u_h'(x))'(e(x)-\pi_h e(x))dx.
$$

Thus, summing over k, we have

$$
-\int_0^1 a(x)u_h'(x)(e'(x)-(\pi_h e(x))'dx = \int_0^1 (a(x)u_h'(x))'(e(x)-\pi_h e(x))dx,
$$

where $(a(x)u_h'(x))'$ should be interpreted locally on each subinterval $[x_{k-1}, x_k]$. (Since $u_h'(x)$ in general is discontinuous, $u_h''(x)$ does not exist globally on $[0,1]$.) Therefore

$$
\begin{aligned}
\|e(x)\|_E^2 &= \int_0^1 f(x)(e(x)-\pi_h e(x))dx + \int_0^1 (a(x)u_h'(x))'(e(x)-\pi_h e(x))dx \\
&= \int_0^1 \{f(x)+(a(x)u_h'(x))'\}(e(x)-\pi_h e(x))dx.
\end{aligned}
$$

Now let $R(u_h(x)) = f(x) + (a(x)u_h'(x))'$, i.e. $R(u_h(x))$ is the *residual error*, which is a well-defined function except in the set $\{x_k\}$, $k = 1, \ldots, M$; where $(a(x_k)u_h'(x_k))'$ is not defined. Then, using Cauchy-Schwarz' inequality we get the following estimate

$$
\begin{aligned}
\|e(x)\|_E^2 &= \int_0^1 R(u_h(x))(e(x)-\pi_h e(x))dx = \\
&= \int_0^1 \frac{1}{\sqrt{a(x)}} h(x) R(u_h(x)) \cdot \sqrt{a(x)} \Big(\frac{e(x)-\pi_h e(x)}{h(x)}\Big)\,dx \\
&\le \Big(\int_0^1 \frac{1}{a(x)} h^2(x) R^2(u_h(x))dx\Big)^{1/2} \Big(\int_0^1 a(x)\Big(\frac{e(x)-\pi_h e(x)}{h(x)}\Big)^2 dx\Big)^{1/2}.
\end{aligned}
$$

Further, by the definition of the weighted L_2-norm we have,

$$\Big\|\frac{e(x)-\pi_h e(x)}{h(x)}\Big\|_a = \Big(\int_0^1 a(x)\Big(\frac{e(x)-\pi_h e(x)}{h(x)}\Big)^2 dx\Big)^{1/2}. \tag{5.4.20}$$

To estimate (5.4.20) we can use the third interpolation estimate for $e(x)$ in each subinterval and get

$$\Big\|\frac{e(x)-\pi_h e(x)}{h(x)}\Big\|_a \le C_i\|e'(x)\|_a = C_i\|e(x)\|_E, \tag{5.4.21}$$

where C_i as before depends on $a(x)$. Thus

$$\|e(x)\|_E^2 \le \Big(\int_0^1 \frac{1}{a(x)}h^2(x)R^2(u_h(x))dx\Big)^{1/2}\cdot C_i\|e(x)\|_E, \tag{5.4.22}$$

and the proof is complete. □

Remark 5.4. *The detailed derivation of* (5.4.21) *is as follows:*

$$\begin{aligned}
\Big\|\frac{e(x)-\pi_h e(x)}{h(x)}\Big\|_a &= \Big(\sum_{j=1}^{M+1}\int_{I_j} a(x)\Big(\frac{e(x)-\pi_h e(x)}{h_j}\Big)^2 dx\Big)^{1/2}\\
&\le \max_{x\in]0,1]} a(x)^{1/2}\Big(\sum_{j=1}^{M+1}\frac{1}{h_j^2}\int_{I_j} a(x)(e(x)-\pi_h e(x))\Big)^2 dx)^{1/2}\\
&\le \max_{x\in]0,1]} a(x)^{1/2}\Big(\sum_{j=1}^{M+1}\frac{1}{h_j^2}\cdot c_i^2\cdot\int_{I_j}(h_j e'(x)\Big)^2 dx)^{1/2}\\
&\le c_i\frac{\max_{x\in]0,1]} a(x)^{1/2}}{\min_{x\in]0,1]} a(x)^{1/2}}\|e'\|_a.
\end{aligned}$$

Adaptivity

Below we briefly outline the adaptivity procedure based on the a posteriori error estimate which uses the *approximate* solution and which can be used for mesh-refinements. Loosely speaking, the estimate (5.4.16) predicts local mesh refinement, i.e. indicates the regions (subintervals) which should be subdivided further. More specifically the idea is as follows: assume that one seeks an error less than a given error tolerance TOL > 0:

$$\|e(x)\|_E \le \text{TOL}. \tag{5.4.23}$$

Then, one may use the following steps as a mesh refinement strategy:

(i) Make an initial partition of the interval

(ii) Compute the corresponding FEM solution $u_h(x)$ and residual $R(u_h(x))$.

(iii) If $\|e(x)\|_E > \text{TOL}$, refine the mesh in the places where $\frac{1}{a(x)}R^2(u_h(x))$ is large and perform the steps (ii) and (iii) again.

5.5 FEM for convection–diffusion–absorption boundary value problems

We now return to the Galerkin approximation of a solution to boundary value problems and give a framework for the cG(1) finite element procedure leading to a linear system of equations of the form $A\xi = \mathbf{b}$. More specifically, we shall extend the approach in Chapter 2, for the stationary heat equation, to cases involving *absorption* and/or *convection* terms. We also consider non-homogeneous Dirichlet boundary conditions. We illustrate this procedure through the following two examples.

Example 5.1. *Determine the coefficient matrix and load vector for the cG(1) finite element approximation of the boundary value problem*

$$-u''(x) + 4u(x) = 0, \quad 0 < x < 1; \qquad u(0) = \alpha \neq 0, \quad u(1) = \beta \neq 0,$$

on a uniform partition $\mathcal{T}_h$ of the interval $[0, 1]$ into $n + 1$ subintervals of length $h = 1/(n + 1)$.
Solution: The problem is to construct an approximate solution u_h in a finite dimensional space spanned by the piecewise linear basis functions (hat-functions) $\varphi_j(x)$, $j = 0, 1, \ldots, n + 1$ on the partition $\mathcal{T}_h$. This results in a discrete problem represented by a linear system of equations $A\xi = \mathbf{b}$, for the unknown $\xi = \{c_j\}_{j=1}^n$, ($c_0 = \alpha$ and $c_{n+1} = \beta$ will be given by the boundary conditions.)
The continuous solution is assumed to be in the Hilbert space

$$H^1 = \left\{ w : \int_0^1 \left(w(x)^2 + w'(x)^2 \right) dx < \infty \right\}.$$

Since both $u(0) = \alpha$ *och* $u(1) = \beta$ *are given, we need to take the trial functions in*

$$V := \{w : w \in H^1, \quad w(0) = \alpha,\ w(1) = \beta\},$$

and the test functions in

$$V^0 := H_0^1 = \{w : w \in H^1, \quad w(0) = w(1) = 0\}.$$

We multiply the PDE by a test function $v \in V^0$ *and integrate over* $(0, 1)$. *Integrating by parts we get*

$$-u'(1)v(1) + u'(0)v(0) + \int_0^1 u'v'\,dx + 4\int_0^1 uv\,dx = 0 \quad \Longleftrightarrow$$

$$(VF): \quad \textit{Find} \quad u \in V \quad \textit{so that} \quad \int_0^1 u'v'\,dx + 4\int_0^1 uv\,dx = 0, \qquad \forall v \in V^0.$$

The partition $\mathcal{T}_h$, *of* $[0, 1]$ *into* $n + 1$ *uniform subintervals* $I_1 = [0, h]$, $I_2 = [h, 2h]$, *...,* *and* $I_{n+1} = [nh, (n + 1)h]$, *is also described by the nodes* $x_0 = 0, x_1 = h, \ldots, x_n = nh$ *and* $x_{n+1} = (n + 1)h = 1$. *The corresponding discrete function spaces are*

$$V_h := \{w_h : w_h \textit{ is piecewise linear, continuous on } \mathcal{T}_h,\ w_h(0) = \alpha,\ w_h(1) = \beta\},$$

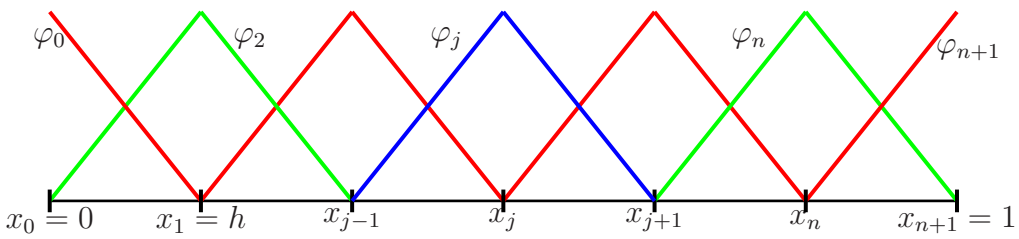
and

$$V_h^0 := \{v_h : v_h \textit{ is piecewise linear and continuous on } \mathcal{T}_h,\ \ v_h(0) = v_h(1) = 0\}.$$

Note that here, the basis functions needed to represent functions in V_h *are the hat-functions* $\varphi_j, j = 0, \ldots, n+1$ *(including the two* half-hat-functions φ_0 *and* φ_{n+1}*), whereas the basis functions describing* V_h^0 *are* φ_i*:s for* $i = 1, \ldots, n$, *i.e. all* full-hat-functions *but not* φ_0 *and* φ_{n+1}. *This is due to the fact that the values* $u(0) = \alpha$ *och* $u(1) = \beta$ *are given and therefore we do not need to determine those two nodal values approximately. See the figure below for all* $\varphi_j, j = 0, \ldots, n + 1$.

Now the finite element formulation (the discrete variational formulation) is: find $u_h \in V_h$ *such that*

$$(FEM) \qquad \int_0^1 u_h'v'\,dx + 4\int_0^1 u_h v\,dx = 0, \quad \forall v \in V_h^0.$$



We have that $u_h(x) = c_0\varphi_0(x) + \sum_{j=1}^{n} c_j\varphi_j(x) + c_{n+1}\varphi_{j+1}(x)$, *where* $c_0 = \alpha$, $c_{n+1} = \beta$ *and*

$$\varphi_0(x) = \frac{1}{h}\begin{cases} h - x & 0 \le x \le h \\ 0, & else \end{cases}, \quad \varphi_j(x) = \frac{1}{h}\begin{cases} x - x_{j-1}, & x_{j-1} \le x \le x_j \\ x_{j+1} - x & x_j \le x \le x_{j+1} \\ 0 & x \notin [x_{j-1}, x_{j+1}]. \end{cases}$$

and

$$\varphi_{n+1}(x) = \frac{1}{h}\begin{cases} x - x_n & nh \le x \le (n+1)h \\ 0, & else. \end{cases}.$$

Inserting u_h *into (FEM), and choosing* $v = \varphi_i(x)$, $i = 1, \ldots, n$ *we get*

$$\begin{aligned} &\sum_{j=1}^{n}\Big(\int_0^1 \varphi_j'(x)\varphi_i'(x)\,dx + 4\int_0^1 \varphi_j(x)\varphi_i(x)\,dx\Big)c_j \\ &\quad = -\Big(\int_0^1 \varphi_0'(x)\varphi_i'(x)\,dx + 4\int_0^1 \varphi_0(x)\varphi_i(x)\,dx\Big)c_0 \\ &\quad - \Big(\int_0^1 \varphi_{n+1}'(x)\varphi_i'(x)\,dx + 4\int_0^1 \varphi_{n+1}(x)\varphi_i(x)\,dx\Big)c_{n+1}. \end{aligned}$$

In matrix form this corresponds to $A\xi = \mathbf{b}$ *with* $A = S + 4M$, *where* $S = A_{unif}$

is the, previously computed, stiffness matrix:

$$S = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & -1 & 2 & -1 \\ 0 & \dots & \dots & \dots & -1 & 2 \end{bmatrix}. \tag{5.5.1}$$

M is the mass-matrix:

$$M = \begin{bmatrix} \int_0^1 \varphi_1\varphi_1 & \int_0^1 \varphi_2\varphi_1 & \dots & \int_0^1 \varphi_n\varphi_1 \\ \int_0^1 \varphi_1\varphi_2 & \int_0^1 \varphi_2\varphi_2 & \dots & \int_0^1 \varphi_n\varphi_2 \\ \dots & \dots & \dots & \dots \\ \int_0^1 \varphi_1\varphi_n & \int_0^1 \varphi_2\varphi_n & \dots & \int_0^1 \varphi_n\varphi_n \end{bmatrix}. \tag{5.5.2}$$

Remark 5.5. *Note the index locations in the matrix M:*

$$m_{ij} = \int_0^1 \varphi_j(x)\varphi_i(x)\,dx.$$

This, however, does not make any difference in the current example, since M is a symmetric matrix.

To compute the entries of M, we follow the same procedure as in Chapter 2, and notice that, as S, also M is symmetric and its elements m_{ij} are given by

$$m_{ij} = m_{ji} = \begin{cases} \int_0^1 \varphi_i\varphi_j\,dx = 0, & \forall i,j \quad \textit{with} \quad |i-j| > 1 \\ \int_0^1 \varphi_j^2(x)\,dx, & \textit{for} \quad i = j \\ \int_0^1 \varphi_j(x)\varphi_{j+1}(x), & \textit{for} \quad i = j+1. \end{cases} \tag{5.5.3}$$

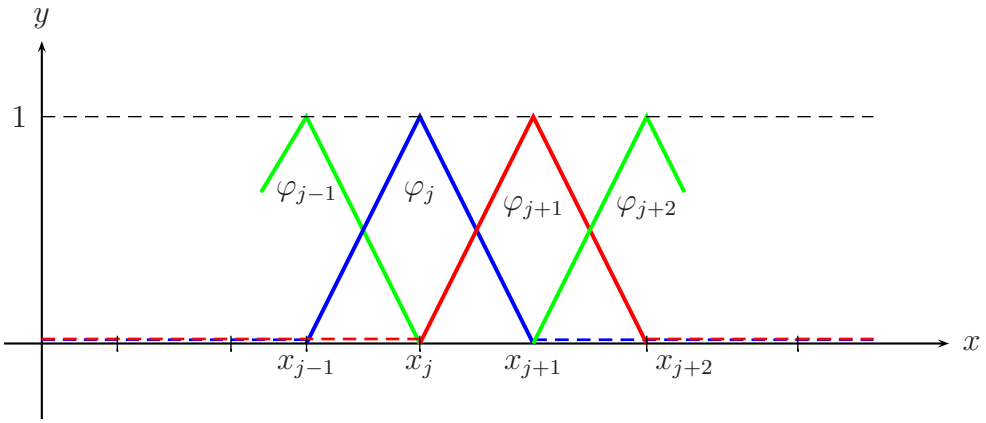


Figure 5.2: φ_j and φ_{j+1}.

The diagonal elements are

$$\begin{aligned} m_{jj} &= \int_0^1 \varphi_j(x)^2\,dx = \frac{1}{h^2}\Big(\int_{x_{j-1}}^{x_j}(x-x_{j-1})^2\,dx + \int_{x_j}^{x_{j+1}}(x_{j+1}-x)^2\Big) \\ &= \frac{1}{h^2}\Big[\frac{(x-x_{j-1})^3}{3}\Big]_{x_{j-1}}^{x_j} - \frac{1}{h^2}\Big[\frac{(x_{j+1}-x)^3}{3}\Big]_{x_j}^{x_{j+1}} \\ &= \frac{1}{h^2}\cdot\frac{h^3}{3} + \frac{1}{h^2}\cdot\frac{h^3}{3} = \frac{2}{3}h, \quad j=1,\ldots,n, \end{aligned} \tag{5.5.4}$$

and the two super- and sub-diagonals can be computed as

$$\begin{aligned} m_{j,j+1} = m_{j+1,j} &= \int_0^1 \varphi_j\varphi_{j+1}\,dx = \frac{1}{h^2}\int_{x_j}^{x_{j+1}}(x_{j+1}-x)(x-x_j) = [PI] \\ &= \frac{1}{h^2}\Big[(x_{j+1}-x)\frac{(x-x_j)^2}{2}\Big]_{x_j}^{x_{j+1}} - \frac{1}{h^2}\int_{x_j}^{x_{j+1}} -\frac{(x-x_j)^2}{2}\,dx \\ &= \frac{1}{h^2}\Big[\frac{(x-x_j)^3}{6}\Big]_{x_j}^{x_{j+1}} = \frac{1}{6}h, \quad j=1,\ldots,n-1. \end{aligned}$$

Thus the mass matrix in this case is

$$M = h\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 & \dots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & \dots & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \dots & \dots & \dots & \frac{1}{6} & \frac{2}{3} \end{bmatrix} = \frac{h}{6}\begin{bmatrix} 4 & 1 & 0 & 0 & \dots & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 & 4 & 1 \\ 0 & \dots & \dots & \dots & 1 & 4 \end{bmatrix}.$$

Hence, for $i, j = 1, \dots, n$, the coefficient matrix $A = S + 4M$ is given by

$$[A]_{ij} = \int_0^1 \varphi_i'\varphi_j'\,dx + 4\int_0^1 \varphi_i\varphi_j(x)\,dx = \begin{cases} \frac{2}{h} + \frac{8h}{3}, & i = j, \\ -\frac{1}{h} + \frac{2h}{3}, & |i-j| = 1, \\ 0 & \textit{else.} \end{cases}$$

Finally, with $c_0 = \alpha$ och $c_{n+1} = \beta$, we get the load vector viz,

$$\begin{aligned} b_1 &= -(-\frac{1}{h} + \frac{2h}{3})c_0 = \alpha(\frac{1}{h} - \frac{2h}{3}), \\ b_2 &= \dots = b_{n-1} = 0, \\ b_n &= -(-\frac{1}{h} + \frac{2h}{3})c_{n+1} = \beta(\frac{1}{h} - \frac{2h}{3}). \end{aligned}$$

Now, for each particular choice of h (i.e. n), α and β we may solve $A\xi = \mathbf{b}$ to obtain the nodal values of the approximate solution u_h at the inner nodes x_j, $j = 1, \dots, n$. That is: $\xi = (c_1, \dots, c_n)^T := (u_h(x_1), \dots, u_h(x_n))^T$. Connecting the points $(x_j, u_h(x_j))$, $j = 0, \dots, n+1$ by straight lines we obtain the desired piecewise linear approximation for the solution of our boundary value problem.

Remark 5.6. *An easier way of to compute the above integrals $m_{j,j+1}$ (as well as m_{jj}) is through Simpson's rule, which is exact for polynomials of degree ≤ 2. Since $\varphi_j(x)\varphi_{j+1}(x) = 0$ at $x = x_j$ and $x = x_{j+1}$, we need to evaluate only the midterm of the Simpson's formula, i.e.*

$$\int_0^1 \varphi_j\varphi_{j+1}\,dx = 4\frac{h}{6}\varphi_j\Big(\frac{x_j + x_{j+1}}{2}\Big)\cdot\varphi_{j+1}\Big(\frac{x_j + x_{j+1}}{2}\Big)\cdot = 4\cdot\frac{h}{6}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{h}{6}.$$

Example 5.2. *Below we consider a convection-diffusion problem:*

$$-\varepsilon u''(x) + pu'(x) = r, \quad 0 < x < 1; \qquad u(0) = 0, \quad u'(1) = \beta \neq 0,$$

where ε and p are positive real numbers and $r \in \mathbf{R}$. Here $-\varepsilon u''$ is the diffusion term, pu' corresponds to convection, and r is a given (here for simplicity a constant) source ($r > 0$) or sink ($r < 0$). We would like to answer the same question as in the previous example. This time with $c_0 = u(0) = 0$. Then, the test function at $x = 0$; φ_0 will not be necessary. But since $u(1)$ is not given, we shall need the test function at $x = 1$: φ_{n+1}. The function space for the continuous solution: the trial function space, and the test function space are both the same:

$$V := \left\{ w : \int_0^1 \left(w(x)^2 + w'(x)^2 \right) dx < \infty, \text{ and } w(0) = 0 \right\}.$$

We multiply the PDE by a test function $v \in V$ and integrate over $(0, 1)$. Then, integration by parts yields

$$-\varepsilon u'(1)v(1) + \varepsilon u'(0)v(0) + \varepsilon \int_0^1 u'v'dx + p \int_0^1 u'vdx = r \int_0^1 v\, dx.$$

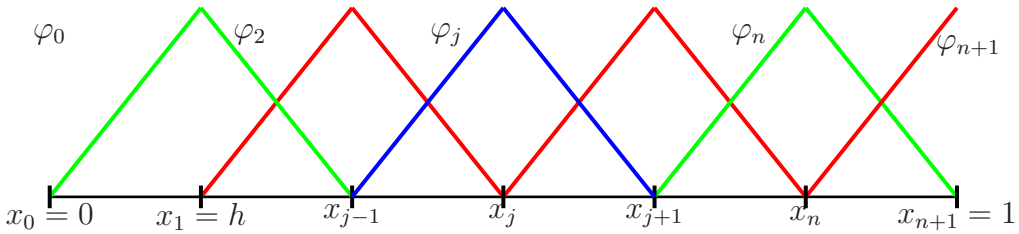
Hence, we end up with the variational formulation: find $u \in V$ such that

$$\text{(VF)} \qquad \varepsilon \int_0^1 u'v'\, dx + p \int_0^1 u'v\, dx = r \int_0^1 v\, dx + \varepsilon\beta v(1), \qquad \forall v \in V.$$

The corresponding discrete test and trial function space is

$$V_h^0 := \{w_h : w_h \text{ is piecewise linear and continuous on } \mathcal{T}_h, \text{ and } w_h(0) = 0\}.$$

Thus, the basis functions for V_h^0 are the hat-functions $\varphi_j, j = 1, \ldots, n+1$ (including the half-hat-function φ_{n+1}), and hence $\dim(V_h^0) = n + 1$.



Now the finite element formulation (the discrete variational formulation) reads as follows: find $u_h \in V_h^0$ *such that*

$$(FEM) \qquad \varepsilon \int_0^1 u_h' v'\,dx + p\int_0^1 u_h' v\,dx = r\int_0^1 v\,dx + \varepsilon\beta v(1), \quad \forall v \in V_h^0.$$

Inserting the ansatz $u_h(x) = \sum_{j=1}^{n+1} \xi_j \varphi_j(x)$ *into (FEM), and choosing* $v = \varphi_i(x)$, $i = 1, \ldots, n+1$, *we get*

$$\sum_{j=1}^{n+1}\Big(\varepsilon\int_0^1 \varphi_j'(x)\varphi_i'(x)\,dx + p\int_0^1 \varphi_j'(x)\varphi_i(x)\,dx\Big)\xi_j = r\int_0^1 \varphi_i(x)\,dx + \varepsilon\beta\varphi_i(1), \quad i = 1, \ldots, n+1.$$

In matrix form this corresponds to the linear system of equations $A\xi = \mathbf{b}$ *with* $A = \varepsilon\,\tilde{S} + p\,C$, *where* $\tilde{S}$ *is computed as* A_{unif} *and is the* $(n+1)\times(n+1)$*-stiffness matrix with its last diagonal element* $\tilde{s}_{n+1,n+1} = \int_0^1 \varphi_{n+1}'(x)\varphi_{n+1}'(x)\,dx = 1/h$, *and* C *is the convection matrix with the elements*

$$c_{ij} = \int_0^1 \varphi_j'(x)\varphi_i(x)\,dx.$$

Hence we have, evidently,

$$\tilde{S} = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & 0 & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & -1 & 2 & -1 \\ 0 & \ldots & \ldots & \ldots & -1 & 1 \end{bmatrix}.$$

To compute the entries for C, *we note that like,* S, M *and* $\tilde{S}$, *also* C *is a*

tridiagonal matrix. But C is anti-symmetric. Its entries are

$$\begin{cases} c_{ij} = 0, & \text{for} \quad |i-j| > 1 \\ c_{ii} = \int_0^1 \varphi_i(x)\varphi_i'(x)\,dx = 0, & \text{for} \quad i = 1, \ldots, n \\ c_{n+1,n+1} = \int_0^1 \varphi_{n+1}(x)\varphi_{n+1}'(x)\,dx = 1/2, & \\ c_{i,i+1} = \int_0^1 \varphi_i(x)\varphi_{i+1}'(x)\,dx = 1/2, & \text{for} \quad i = 1, \ldots, n \\ c_{i+1,i} = \int_0^1 \varphi_{i+1}(x)\varphi_i'(x)\,dx = -1/2, & \text{for} \quad i = 1, \ldots, n. \end{cases} \tag{5.5.5}$$

Finally, we have the entries b_i of the load vector $\mathbf{b}$ as

$$b_1 = \ldots = b_n = rh, \qquad b_{n+1} = rh/2 + \varepsilon\beta.$$

Thus,

$$C = \frac{1}{2}\begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ -1 & 0 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & -1 & 0 & 1 \\ 0 & \ldots & \ldots & \ldots & -1 & 1 \end{bmatrix}, \quad \mathbf{b} = rh\begin{bmatrix} 1 \\ 1 \\ 1 \\ . \\ 1 \\ 1/2 \end{bmatrix} + \varepsilon\beta\begin{bmatrix} 0 \\ 0 \\ 0 \\ . \\ 0 \\ 1 \end{bmatrix}.$$

Remark 5.7. *In the convection dominated case $\frac{\varepsilon}{p} << 1$ this standard FEM will not work. Spurious oscillations in the approximate solution will appear. The standard FEM has to be modified in this case.*

5.6 Exercises

Problem 5.1. *Consider the two-point boundary value problem*

$$-u'' = f, \quad 0 < x < 1; \qquad u(0) = u(1) = 0. \tag{5.6.1}$$

Let $V = \{v : \|v\| + \|v'\| < \infty, \quad v(0) = v(1) = 0\}$.

a. Use V to derive a variational formulation of (5.6.1).

b. Discuss why V is valid as a vector space of test functions.

c. Classify whether the following functions are admissible test functions or not:

$$\sin \pi x, \qquad x^2, \qquad x \ln x, \qquad e^x - 1, \qquad x(1-x).$$

Problem 5.2. *Assume that $u(0) = u(1) = 0$, and that u satisfies*

$$\int_0^1 u'v'\,dx = \int_0^1 fv\,dx,$$

for all $v \in V = \{v : \|v\| + \|v'\| < \infty, \quad v(0) = v(1) = 0\}$.

a. Show that u minimizes the functional

$$F(v) = \frac{1}{2}\int_0^1 (v')^2\,dx - \int_0^1 fv\,dx. \tag{5.6.2}$$

Hint: $F(v) = F(u + w) = F(u) + \ldots \geq F(u)$.

b. Prove that the above minimization problem is equivalent to

$$-u'' = f, \quad 0 < x < 1; \qquad u(0) = u(1) = 0.$$

Problem 5.3. *Consider the two-point boundary value problem*

$$-u'' = 1, \quad 0 < x < 1; \qquad u(0) = u(1) = 0. \tag{5.6.3}$$

Let $\mathcal{T}_h : x_j = \frac{j}{4}$, $j = 0, 1, \ldots, 4$, denote a partition of the interval $0 < x < 1$ into four subintervals of equal length $h = 1/4$ and let V_h be the corresponding space of continuous piecewise linear functions vanishing at $x = 0$ and $x = 1$.

a. Compute a finite element approximation $U \in V_h$ to (5.6.3).

b. Prove that $U \in V_h$ is unique.

Problem 5.4. *Consider once again the two-point boundary value problem*

$$-u'' = f, \quad 0 < x < 1; \qquad u(0) = u(1) = 0.$$

a. Prove that the finite element approximation $U \in V_h$ to u satisfies

$$\|(u - U)'\| \leq \|(u - v)'\|,$$

for all $v \in V_h$.

b. Use this result to deduce that

$$\|(u - \pi_h u)'\| \leq C\|hu''\|, \tag{5.6.4}$$

where C *is a constant and* $\pi_h u$ *a piecewise linear interpolant to* u.

Problem 5.5. *Consider the two-point boundary value problem*

$$\begin{aligned} -(au')' &= f, & 0 < x < 1, \\ u(0) &= 0, & a(1)u'(1) = g_1, \end{aligned} \tag{5.6.5}$$

where $a > 0$ *is a positive function and* g_1 *is a constant.*

a. Derive the variational formulation of (5.6.5).

b. Discuss how the boundary conditions are implemented.

Problem 5.6. *Consider the two-point boundary value problem*

$$-u'' = 0, \quad 0 < x < 1; \qquad u(0) = 0, \quad u'(1) = 7. \tag{5.6.6}$$

Divide the interval $0 \leq x \leq 1$ *into two subintervals of length* $h = \frac{1}{2}$ *and let* V_h *be the corresponding space of continuous piecewise linear functions vanishing at* $x = 0$.

a. Formulate a finite element method for (5.6.6).

b. Calculate by hand the finite element approximation $U \in V_h$ *to* (5.6.6).

Study how the boundary condition at $x = 1$ *is approximated.*

Problem 5.7. *Consider the two-point boundary value problem*

$$-u'' = 0, \quad 0 < x < 1; \qquad u'(0) = 5, \quad u(1) = 0. \tag{5.6.7}$$

Let $\mathcal{T}_h : x_j = jh$, $j = 0, 1, \ldots, N$, $h = 1/N$ *be a uniform partition of the interval* $0 < x < 1$ *into* N *subintervals and let* V_h *be the corresponding space of continuous piecewise linear functions.*

a. Use V_h *to formulate a finite element method for* (5.6.7).

b. Compute the finite element approximation $U \in V_h$ *assuming* $N = 3$.

Problem 5.8. *Consider the problem of finding a solution approximation to*

$$-u'' = 1, \quad 0 < x < 1; \qquad u'(0) = u'(1) = 0. \tag{5.6.8}$$

Let $\mathcal{T}_h$ be a partition of the interval $0 < x < 1$ into two subintervals of equal length $h = \frac{1}{2}$ and let V_h be the corresponding space of continuous piecewise linear functions.

a. Find the exact solution to (5.6.8) *by integrating twice.*

b.Compute a finite element approximation $U \in V_h$ to u if possible.

Problem 5.9. *Consider the two-point boundary value problem*

$$-((1+x)u')' = 0, \quad 0 < x < 1; \qquad u(0) = 0, \quad u'(1) = 1. \tag{5.6.9}$$

Divide the interval $0 < x < 1$ into 3 subintervals of equal length $h = \frac{1}{3}$ and let V_h be the corresponding space of continuous piecewise linear functions vanishing at $x = 0$.

a. Use V_h to formulate a finite element method for (5.6.9).

b. Verify that the stiffness matrix $\mathbf{A}$ *and the load vector* $\mathbf{b}$ *are given by*

$$\mathbf{A} = \frac{1}{2}\begin{bmatrix} 16 & -9 & 0 \\ -9 & 20 & -11 \\ 0 & -11 & 11 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

c. Show that $\mathbf{A}$ *is symmetric tridiagonal, and positive definite.*

d. Derive a simple way to compute the energy norm $\|U\|_E^2$, defined by

$$\|U\|_E^2 = \int_0^1 (1+x)U'(x)^2\,dx,$$

where $U \in V_h$ is the finite element solution approximation.

Problem 5.10. *Consider the two-point boundary value problem*

$$-u'' = 0, \quad 0 < x < 1; \qquad u(0) = 0, \quad u'(1) = k(u(1) - 1). \tag{5.6.10}$$

Let $\mathcal{T}_h : 0 = x_0 < x_1 < x_2 < x_3 = 1$, where $x_1 = \frac{1}{3}$ and $x_2 = \frac{2}{3}$ be a partition of the interval $0 \le x \le 1$ and let V_h be the corresponding space of continuous piecewise linear functions, which vanish at $x = 0$.

a. Compute a solution approximation $U \in V_h$ to (5.6.10) *assuming* $k = 1$.

b. Discuss how the parameter k influence the boundary condition at $x = 1$.

Problem 5.11. *Consider the finite element method applied to*

$$-u'' = 0, \quad 0 < x < 1; \qquad u(0) = \alpha, \quad u'(1) = \beta,$$

where α and β are given constants. Assume that the interval $0 \le x \le 1$ is divided into three subintervals of equal length $h = 1/3$ and that $\{\varphi_j\}_0^3$ is a nodal basis of V_h, the corresponding space of continuous piecewise linear functions.

a. Verify that the ansatz

$$U(x) = \alpha\varphi_0(x) + \xi_1\varphi_1(x) + \xi_2\varphi_2(x) + \xi_3\varphi_3(x),$$

yields the following system of equations

$$\frac{1}{h}\begin{bmatrix} -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} \alpha \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \beta \end{bmatrix}. \tag{5.6.11}$$

b. If $\alpha = 2$ and $\beta = 3$ show that (5.6.11) *can be reduced to*

$$\frac{1}{h}\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \begin{bmatrix} -2h^{-1} \\ 0 \\ 3 \end{bmatrix}.$$

c. Solve the above system of equations to find $U(x)$.

Problem 5.12. *Compute a finite element solution approximation to*

$$-u'' + u = 1; \qquad 0 \le x \le 1, \qquad u(0) = u(1) = 0, \tag{5.6.12}$$

using the continuous piecewise linear ansatz $U = \xi_1\varphi_1(x) + \xi_2\varphi_2(x)$ where

$$\varphi_1(x) = \begin{cases} 3x, & 0 < x < \frac{1}{3} \\ 2 - 3x, & \frac{1}{3} < x < \frac{2}{3}, \\ 0, & \frac{2}{3} < x < 1 \end{cases} \qquad \varphi_2(x) = \begin{cases} 0, & 0 < x < \frac{1}{3} \\ 3x - 1, & \frac{1}{3} < x < \frac{2}{3}. \\ 3 - 3x, & \frac{2}{3} < x < 1 \end{cases}$$

Problem 5.13. *Consider the following eigenvalue problem*

$$-au'' + bu = 0; \qquad 0 \le x \le 1, \qquad u(0) = u'(1) = 0, \tag{5.6.13}$$

where $a, b > 0$ are constants. Let $\mathcal{T}_h : 0 = x_0 < x_1 < \ldots < x_N = 1$, be a non-uniform partition of the interval $0 \le x \le 1$ into N intervals of length $h_i = x_i - x_{i-1},\ i = 1, 2, \ldots, N$ and let V_h be the corresponding space of continuous piecewise linear functions. Compute the stiffness and mass matrices.

Problem 5.14. *Show that the FEM with the mesh size h for the problem:*

$$\begin{cases} -u'' = 1 & 0 < x < 1 \\ u(0) = 1 & u'(1) = 0, \end{cases} \tag{5.6.14}$$

with

$$U(x) = 7\varphi_0(x) + U_1\varphi_1(x) + \ldots + U_m\varphi_m(x). \tag{5.6.15}$$

leads to the linear system of equations: $\tilde{A} \cdot \tilde{U} = \tilde{b}$, where

$$\underset{m \times (m+1)}{\tilde{A} = \frac{1}{h}\begin{bmatrix} -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \ldots \end{bmatrix}}, \quad \underset{(m+1) \times 1}{\tilde{U} = \begin{bmatrix} 7 \\ U_1 \\ \ldots \\ U_m \end{bmatrix}}, \quad \underset{m \times 1}{\tilde{b} = \begin{bmatrix} h \\ \ldots \\ h \\ h/2 \end{bmatrix}},$$

. which is reduced to $AU = b$, with

$$A = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}, \quad U = \begin{bmatrix} U_1 \\ U_2 \\ \ldots \\ U_m \end{bmatrix}, \quad b = \begin{bmatrix} h + \frac{7}{h} \\ h \\ \ldots \\ h \\ h/2 \end{bmatrix}.$$

Problem 5.15. *Prove an a priori and an a posteriori error estimate for a finite element method (for example cG(1)) for the problem*

$$-u'' + \alpha u = f, \quad \text{in } I = (0,1), \qquad u(0) = u(1) = 0,$$

where the coefficient $\alpha = \alpha(x)$ is a bounded positive function on I, ($0 \leq \alpha(x) \leq K,\ x \in I$).

Problem 5.16. *a) Formulate a cG(1) method for the problem*

$$\begin{cases} (a(x)u'(x))' = 0, & 0 < x < 1, \\ a(0)u'(0) = u_0, & u(1) = 0. \end{cases}$$

and give an a posteriori error estimate.

b) Let $u_0 = 3$ and compute the approximate solution in a) for a uniform partition of $I = [0,1]$ into 4 intervals and

$$a(x) = \begin{cases} 1/4, & x < 1/2, \\ 1/2, & x > 1/2. \end{cases}$$

c) Show that, with these special choices, the computed solution is equal to the exact one, i.e. the error is equal to 0.

Problem 5.17. *Prove an a priori error estimate for the finite element method for the problem*

$$-u''(x) + u'(x) = f(x), \quad 0 < x < 1, \qquad u(0) = u(1) = 0.$$

Problem 5.18. *(a) Prove an a priori error estimate for the cG(1) approximation of the boundary value problem*

$$-u'' + cu' + u = f \quad \text{in } I = (0,1), \qquad u(0) = u(1) = 0,$$

where $c \geq 0$ is constant.

(b) For which value of c is the a priori error estimate optimal?

Problem 5.19. *Let U be the piecewise linear finite element approximation for*

$$-u''(x) + 2xu'(x) + 2u(x) = f(x), \quad x \in (0,1), \qquad u(0) = u(1) = 0,$$

in a partition $\mathcal{T}_h$ of the interval $[0,1]$. Set $e = u - U$ and derive a priori and a posteriori error estimates in the energy-norm:

$$||e||_E^2 = ||e'||^2 + ||e||^2, \quad \textit{where} \quad ||w||^2 = \int_0^1 w(x)^2\,dx.$$

Chapter 6

Scalar Initial Value Problems

This chapter is devoted to finite element methods for time discretizations. Here, we shall consider problems depending on the time variable, only. The approximation techniques developed in this chapter, combined with those of the previous chapter for boundary value problems, can be used for the numerical study of initial boundary value problems; such as, e.g. the heat and wave equations, by finite element methods.

As a model problem we shall consider the classical example of population dynamics described by the following ordinary differential equation (ODE)

$$\begin{array}{ll} (\mathrm{DE}) & \left\{ \begin{array}{rl} \dot{u}(t) + a(t)u(t) = f(t), & 0 < t \leq T, \\ u(0) = u_0, & \end{array} \right. \\ (\mathrm{IV}) & \end{array} \tag{6.0.1}$$

where $f(t)$ is the source term and $\dot{u}(t) = \dfrac{du}{dt}$. The coefficient $a(t)$ is a bounded function. If $a(t) \geq 0$ the problem $(6.0.1)$ is called $parabolic$, while $a(t) \geq \alpha > 0$ yields a $dissipative\ problem$, in the sense that, with increasing t, perturbations of solutions to $(6.0.1)$, e.g. introduced by numerical discretization, will decay. In general, in numerical approximations for $(6.0.1)$, the error accumulates when advancing in time, i.e. the error of previous time steps adds up to the error of the present time step. The different types of error accumulation/perturbation growth are referred to as stability properties of the initial value problem.

6.1 Solution formula and stability

Theorem 6.1. *The solution of the problem* (6.0.1) *is given by*

$$u(t) = u_0 \cdot e^{-A(t)} + \int_0^t e^{-(A(t)-A(s))} f(s) ds, \tag{6.1.1}$$

where $A(t) = \int_0^t a(s)ds$ *and* $e^{A(t)}$ *is the integrating factor.*

Proof. Multiplying the (DE) by the integrating factor $e^{A(t)}$ we have

$$\dot{u}(t)e^{A(t)} + \dot{A}(t)e^{A(t)}u(t) = e^{A(t)}f(t), \tag{6.1.2}$$

where we used that $a(t) = \dot{A}(t)$. Equation (6.1.2) can be rewritten as

$$\frac{d}{dt}\Big(u(t)e^{A(t)}\Big) = e^{A(t)}f(t).$$

We denote the variable by s and integrate from 0 to t to get

$$\int_0^t \frac{d}{ds}\Big(u(s)e^{A(s)}\Big)ds = \int_0^t e^{A(s)}f(s)ds,$$

i.e.

$$u(t)e^{A(t)} - u(0)e^{A(0)} = \int_0^t e^{A(s)}f(s)ds.$$

Now since $A(0) = 0$ and $u(0) = u_0$ we get the desired result

$$u(t) = u_0 e^{-A(t)} + \int_0^t e^{-(A(t)-A(s))} f(s) ds. \tag{6.1.3}$$

□

This representation for $u(t)$ is also referred to as the *Variation of constants formula*.

Theorem 6.2 (Stability estimates)**.** *Using the solution formula, we can derive the following stability estimates:*

(i) If $a(t) \geq \alpha > 0$, *then* $|u(t)| \leq e^{-\alpha t}|u_0| + \dfrac{1}{\alpha}(1 - e^{-\alpha t}) \max\limits_{0\leq s\leq t} |f(s)|$

(ii) If $a(t) \geq 0$ (i.e. $\alpha = 0$; the parabolic case), then

$$|u(t)| \leq |u_0| + \int_0^t |f(s)|ds \quad \text{or} \quad |u(t)| \leq |u_0| + \|f\|_{L_1(0,t)} \tag{6.1.4}$$

Proof. (i) For $a(t) \geq \alpha > 0$ we have that $A(t) = \int_0^t a(s)ds$ is an increasing function of t, $A(t) \geq \alpha t$ and

$$A(t) - A(s) = \int_0^t a(r)\,dr - \int_0^s a(r)\,dr = \int_s^t a(r)\,dr \geq \alpha(t-s). \tag{6.1.5}$$

Thus $e^{-A(t)} \leq e^{-\alpha t}$ and $e^{-(A(t)-A(s))} \leq e^{-\alpha(t-s)}$. Hence, using (6.1.3) we get

$$|u(t)| \leq |u_0|e^{-\alpha t} + \int_0^t e^{-\alpha(t-s)}|f(s)|ds, \tag{6.1.6}$$

which after integration yields

$$|u(t)| \leq e^{-\alpha t}|u_0| + \max_{0\leq s\leq t}|f(s)|\Big[\frac{1}{\alpha}e^{-\alpha(t-s)}\Big]_{s=0}^{s=t}, \quad \text{i.e.}$$
$$|u(t)| \leq e^{-\alpha t}|u_0| + \frac{1}{\alpha}(1-e^{-\alpha t})\max_{0\leq s\leq t}|f(s)|.$$

(ii) Let $\alpha = 0$ in (6.1.6) (which is true also in this case), then $|u(t)| \leq |u_0| + \int_0^t |f(s)|ds$, and the proof is complete. □

Remark 6.1. *(i) expresses that the effect of the initial data u_0 decays exponentially with time, and that the effect of the source term f on the right hand side does not depend on the length of the time interval, only on the maximum value of f, and on the value of α. Regarding (ii): in this case the influence of u_0 remains bounded in time, and the integral of f indicates an accumulation in time.*

6.2 Galerkin finite element methods for IVP

Recall that we refer to the set of functions where we seek the approximate solution as the *trial space* and the space of functions used for the orthogonality condition (multipliers), as the *test space*.

The polynomial approximation procedure introduced in Chapter 2, along (2.1.21)-(2.1.29), for the initial value problem (2.1.1), and consequently also for (6.0.1), being over the whole time interval $(0, T)$ is known as the *global Galerkin method*. In this section, first we shall introduce two versions of the global Galerkin method and then extend them to partitions of the interval $(0, T)$ using piecewise polynomial test and trial functions. However, in most practical applications, if suffices to consider the following two simple, low degree polynomial, approximation cases that are studied in detail in this chapter:
The continuous Galerkin method of degree 1; cG(1). In this case the trial functions are piecewise linear and continuous while the test functions are piecewise constant and discontinuous, i.e. *unlike the cG(1) for BVP*, here the trial and test functions belong to polynomial spaces of different degree.

The discontinuous Galerkin method of degree 0; dG(0). Here both the trial and test functions are piecewise constant and discontinuous.

Below we introduce the global versions of the above two cases formulated for the whole time interval $(0, T)$ and then derive the local versions formulated for a partition of $(0, T)$.

6.2.1 The continuous Galerkin method

Recall the global Galerkin method of degree q for the initial value problem (6.0.1): find $U \in \mathcal{P}^q(0, T)$, with $U(0) = u_0$ such that

$$\int_0^T (\dot{U} + aU)v dt = \int_0^T f v \, dt, \quad \forall v \in \mathcal{P}^q(0, T), \text{ with } \quad v(0) = 0, \qquad (6.2.1)$$

where $v \in span\{t, t^2, \ldots, t^q\}$.

We now formulate a variation of the above global method: Find $U \in \mathcal{P}^q(0, T)$ with $U(0) = u_0$ such that

$$\int_0^T (\dot{U} + aU)v dt = \int_0^T f v dt, \quad \forall v \in \mathcal{P}^{q-1}(0, T), \qquad (6.2.2)$$

where now the test functions $v \in span\{1, t, t^2, \ldots, t^{q-1}\}$. Hence, the difference between these two methods is in the choice of their test function spaces. We

shall focus on (6.2.2), due to the fact that this method yields a more accurate approximation of degree q than the original method (6.2.1).

Before generalizing (6.2.2) to piecewise polynomial approximation, which is the $cG(q)$ method, we consider an example.

Example 6.1. *Consider* (6.2.2) *with* $q = 1$*, then we may choose* $v \equiv 1$*, thus*

$$\int_0^T (\dot{U} + aU)v dt = \int_0^T (\dot{U} + aU)dt = U(T) - U(0) + \int_0^T aU(t)dt \quad (6.2.3)$$

But $U(t)$ *being a linear function is given by*

$$U(t) = U(0)\frac{T-t}{T} + U(T)\frac{t}{T}. \quad (6.2.4)$$

Inserting into (6.2.3) *we get*

$$U(T) - U(0) + \int_0^T a\Big(U(0)\frac{T-t}{T} + U(T)\frac{t}{T}\Big)\, dt = \int_0^T f\, dt, \quad (6.2.5)$$

which gives an equation for the unknown quantity $U(T)$*. Consequently, using* (6.2.4) *with a given* $U(0)$*, we get the linear approximation* $U(t)$ *for all* $t \in [0, T]$*. We shall now generalize this example to piecewise linear approximation and demonstrate the iteration procedure for the cG(1) scheme.*

The cG(1) Algorithm

For a partition $\mathcal{T}_k$ *of the interval* $[0, T]$ *into subintervals* $I_k = (t_{k-1}, t_k]$*, we perform the following steps:*

(1) Given $U(0) = U_0 = u_0$ *and a source term* f*, apply (6.2.5) to the first subinterval* $(0, t_1]$ *and compute* $U(t_1)$*. Then, using* (6.2.4) *one gets* $U(t), \forall t \in [0, t_1]$*.*

(2) Assume that $U(t)$ *is computed for* t *in all the successive subintervals* $(t_{k-1}, t_k]$ *for* $k = 1, 2, \ldots, n-1$*, where* $U_{k-1} := U(t_{k-1})$ *and* f *are considered as the data and the unknown* $U_k := U(t_k)$ *has been computed using*

$$U_k - U_{k-1} + \int_{t_{k-1}}^{t_k} a\Big(\frac{t_k - t}{t_k - t_{k-1}}U_{k-1} + \frac{t - t_{k-1}}{t_k - t_{k-1}}U_k\Big)dt = \int_{t_{k-1}}^{t_k} f dt.$$

(3) Compute $U(t)$ for $t \in (t_{n-1}, t_n]$, $n = 1, \ldots, N$. This is step 2 with k replaced by n. Now, since U_{n-1} is assumed to be known we can calculate U_n and then $U(t)$, for all $t \in (t_{n-1}, t_n]$.

The cG(q) method
The global continuous Galerkin method of degree q formulated for test and trial functions defined on a partition $\mathcal{T}_k,\ 0 = t_0 < t_1 < \ldots < t_N = T$ of the interval $(0, T)$, is referred to as the $cG(q)$ method. Whence the method reads as follows: find $U(t) \in V_k^{(q)}$, such that $U(0) = u_0$, and

$$\int_0^{t_N} (\dot{U} + aU)w dt = \int_0^{t_N} f w dt, \qquad \forall w \in W_k^{(q-1)}, \tag{6.2.6}$$

where

$$V_k^{(q)} = \{v : v \ \text{ continuous, piecewise polynomials of degree } q \text{ on } \mathcal{T}_k\},$$

$$W_k^{(q-1)} = \{w : w \text{ discontinuous, piecewise polynomials of degree } q-1 \ \text{ on } \mathcal{T}_k\}.$$

So, the difference between the global continuous Galerkin method and cG(q) is that now we have *piecewise polynomials* on a partition rather than global polynomials in the whole interval $(0, T)$.

6.2.2 The discontinuous Galerkin method

We start presenting the *global discontinuous Galerkin method of degree q*: find $U(t) \in \mathcal{P}^q(0, T)$ such that

$$\int_0^T (\dot{U} + aU)v dt + \beta(U(0) - u(0))v(0) = \int_0^T f v dt, \quad \forall v \in \mathcal{P}^q(0, T), \tag{6.2.7}$$

where β is a coefficient that weights the relative importance of the residual error for the initial value $U(0) - u(0)$, against the conventional residual $R(U) := \dot{U} + aU - f$ of the differential equation. This approach gives up the requirement that $U(t)$ satisfies the initial condition. Instead, the initial condition is imposed in a variational sence by the term $(U(0) - u(0))v(0)$. As in the cG(q) case, the above strategy can be localized (formulated for the subintervals in a partition $\mathcal{T}_k$) to derive the *discontinuous Galerkin method of degree q*: the $dG(q)$ scheme below. To this end, we recall the following

standard notation for the right/left limits: $v_n^{\pm} = \lim_{s\to 0^+} v(t_n \pm s)$ and the corresponding *jump* term $[v_n] = v_n^+ - v_n^-$ at the time level $t = t_n$. Then, the $dG(q)$ method (with $\beta \equiv 1$) for (6.0.1) reads as follows: for $n = 1, \ldots, N$; find $U(t) \in \mathcal{P}^q(t_{n-1}, t_n)$ such that

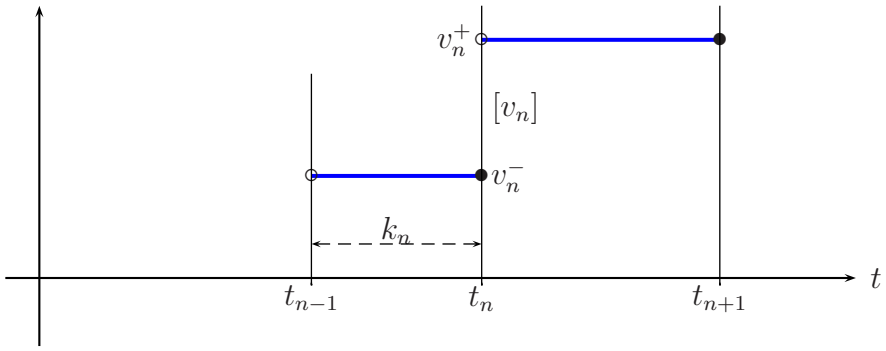


Figure 6.1: The jump $[v_n]$ and the right and left limits $v_n^{\pm}$

$$\int_{t_{n-1}}^{t_n} (\dot{U} + aU)v dt + U_{n-1}^+ v_{n-1}^+ = \int_{t_{n-1}}^{t_n} fv dt + U_{n-1}^- v_{n-1}^+, \quad \forall v \in \mathcal{P}^q(t_{n-1}, t_n). \tag{6.2.8}$$

Example 6.2 (dG(0))**.** *Let $q = 0$, then v is constant generated by the single basis function: $v \equiv 1$. Further, we have $U(t) = U_n = U_{n-1}^+ = U_n^-$ on $I_n = (t_{n-1}, t_n]$, and $\dot{U} \equiv 0$. Thus, for $q = 0$ (6.2.8) yields the following dG(0) formulation: for $n = 1, \ldots, N$; find piecewise constants U_n such that*

$$\int_{t_{n-1}}^{t_n} aU_n dt + U_n = \int_{t_{n-1}}^{t_n} f dt + U_{n-1}. \tag{6.2.9}$$

Summing over n in (6.2.8), we get the following general dG(q) formulation: Find $U(t) \in W_k^{(q)}$, with $U_0^- = u_0$ such that

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (\dot{U} + aU)w dt + \sum_{n=1}^{N} [U_{n-1}]w_{n-1}^+ = \int_0^{t_N} fw dt, \quad \forall w \in W_k^{(q)}. \tag{6.2.10}$$

Remark 6.2. *To compare the $cG(1)$ and $dG(0)$ methods, one can show that (see Johnson et al) $cG(1)$ converges faster than $dG(0)$, whereas $dG(0)$ has better stability properties than $cG(1)$: More specifically, in the parabolic case when $a > 0$ is constant and ($f \equiv 0$) we can easily verify that (see Exercises at the end of this chapter) the $dG(0)$ solution U_n corresponds to the Backward Euler scheme*

$$U_n = \Big(\frac{1}{1+ak}\Big)^n u_0,$$

and the $cG(1)$ solution $\tilde{U}_n$ is given by the Crank-Nicolson scheme:

$$\tilde{U}_n = \Big(\frac{1-\frac{1}{2}ak}{1+\frac{1}{2}ak}\Big)^n u_0,$$

where k is the constant time step.

6.3 A posteriori error estimates

In this section we derive a posteriori error estimates for the most considered Galerkin approaches for time discretization, namely the cG(1) and dG(0) methods. These estimates can be extended to the general $cG(q)$ and $dG(q)$ methods. The details are, however, "much too" involved. Some lower degree polynomial case (e.g. $cG(2)$) can be found in the exercise section.

6.3.1 A posteriori error estimate for cG(1)

Recall the initial value problem (6.0.1) of finding the function u such that

$$\dot{u}(t) + a(t)u(t) = f(t), \quad \forall t \in (0,T], \qquad u(0) = u_0. \tag{6.3.1}$$

A general *variational form* reads as: find u such that

$$\int_0^T (\dot{u} + au)v dt = \int_0^T fv dt, \qquad \text{for all test functions } v.$$

Integrating by parts we get the equivalent equation

$$u(T)v(T) - u(0)v(0) + \int_0^T u(t)(-\dot{v}(t) + av(t))dt = \int_0^T fv dt. \tag{6.3.2}$$

If we now choose the test function $v(t)$ to be the solution of *the dual problem*:

$$-\dot{v}(t) + av(t) = 0, \quad \text{for } t \in (0,T), \tag{6.3.3}$$

then (6.3.2) is simplified to

$$u(T)v(T) = u(0)v(0) + \int_0^T fv dt, \quad \forall v(t). \tag{6.3.4}$$

In other words, choosing v to be the solution of the dual problem (6.3.3) we may get the final value $u(T)$ of the solution directly coupled to the initial value $u(0)$ and the data f. This type of representation will be crucial in the a posteriori error analysis in the proof of the next theorem. Let us state *the dual problem* explicitly:

The dual problem
The dual problem for (6.3.1) is formulated as follows: find $\varphi(t)$ such that

$$\begin{cases} -\dot{\varphi}(t) + a(t)\varphi(t) = 0, & t_N > t \geq 0, \\ \varphi(t_N) = e_N, & e_N = u_N - U_N = u(t_N) - U(t_N). \end{cases} \tag{6.3.5}$$

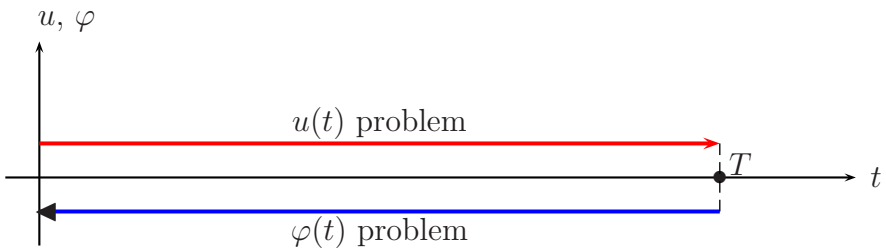


Figure 6.2: Time directions in forward and dual solutions

Note that (6.3.5) runs *backward in time* starting at time $t_N = T$.

Theorem 6.3 (A posteriori error estimate for cG(1))**.** *For $N = 1, 2, \ldots$, the cG(1) solution $U(t)$ satisfies*

$$|e_N| \leq S(t_N) \cdot \max_{t \in [0,t_N]} |k\, r(U)|, \tag{6.3.6}$$

where $k(t)$ is the time step function: $k(t) = k_n = |I_n|$ for $t \in I_n = (t_{n-1}, t_n]$, and $r(U) = \dot{U} + aU - f$ is the residual error for U. Further $S(t_N)$, specified below, is the stability factor satisfying the quantitative bound

$$S(t_N) := \frac{\int_0^{t_N} |\dot{\varphi}| dt}{|e_N|} \leq \begin{cases} e^{\lambda t_N}, & \text{if} \quad |a(t)| \leq \lambda, \quad \text{for} \quad 0 \leq t \leq t_n, \\ 1, & \text{if} \quad a(t) \geq 0, \qquad \text{for} \quad 0 \leq t \leq t_n. \end{cases} \tag{6.3.7}$$

Proof. Let $e(t) = u(t) - U(t)$. Using the dual problem $-\dot{\varphi}(t) + a(t)\varphi(t) = 0$ we may write

$$e_N^2 = e_N^2 + 0 = e_N^2 + \int_0^{t_N} e(-\dot{\varphi} + a\varphi)\, dt, \tag{6.3.8}$$

and by integration by parts we get

$$\begin{aligned} \int_0^{t_N} e(-\dot{\varphi} + a\varphi) dt &= [-e(t)\varphi(t)]_{t=0}^{t_N} + \int_0^{t_N} \dot{e}\varphi\, dt + \int_0^{t_N} ea\varphi\, dt \\ &= -e(t_N)\varphi(t_N) + \int_0^{t_N} (\dot{e} + ae)\varphi\, dt = -e_N^2 + \int_0^{t_N} (\dot{e} + ae)\varphi\, dt, \end{aligned}$$

where we used that $e(0) = 0$ and $\varphi(t_N) = e_N$ when evaluating the boundary terms. Note that

$$\begin{aligned} \dot{e}(t) + a(t)e(t) &= \dot{u}(t) - \dot{U}(t) + a(t)u(t) - a(t)U(t) \\ &= f(t) - \dot{U}(t) - a(t)U(t) := -r(U), \end{aligned} \tag{6.3.9}$$

where we used that $f(t) = \dot{u}(t) + a(t)u(t)$ and the residual $r(U) := \dot{U} + aU - f$. Consequently, we get the following *error representation formula*:

$$e_N^2 = -\int_0^{t_N} r(U(t))\varphi(t) dt. \tag{6.3.10}$$

To continue we use the L_2-projection $\pi_k\varphi = \frac{1}{k_n}\int_{I_n} \varphi(s) ds$ of φ onto the space of piecewise constant polynomials $W_k^{(0)}$ and write

$$e_N^2 = -\int_0^{t_N} r(U)(\varphi(t) - \pi_k\varphi(t)) dt - \int_0^{t_N} r(U)\pi_k\varphi(t) dt. \tag{6.3.11}$$

Now, by the cG(1) method (6.2.6),

$$\int_0^{t_N} (\dot{U} + aU)v\,dt = \int_0^{t_N} fv\,dt, \qquad \forall v \in W_k^{(0)}, \tag{6.3.12}$$

where $W_k^{(0)}$ is the space of discontinuous, piecewise constant test functions. This gives the *orthogonality* relation, $r(U) \perp v,\ \forall v \in W_k^{(0)}$, and implies that

$$\int_0^{t_N} r(U)\pi_k\varphi(t)dt = 0, \tag{6.3.13}$$

since $\pi_k\varphi \in W_k^{(0)}$. Thus, the final error representation formula becomes

$$e_N^2 = -\int_0^{t_N} r(U)(\varphi(t) - \pi_k\varphi(t))dt. \tag{6.3.14}$$

Next, we shall need the L_2-projection error estimate (proved as the interpolation error estimate) for the function φ in the interval I_n, with $|I_n| = k_n$:

$$\int_0^{t_N} |\varphi - \pi_k\varphi|dt \le \sum_{n=1}^{N} k_n \int_{I_n} |\dot{\varphi}|dt. \tag{6.3.15}$$

To show (6.3.15) let $\bar{\varphi} = \frac{1}{k_n}\int_{I_n} \varphi(s)ds$ be the mean value of φ over I_n, then

$$\begin{aligned}\int_{I_n} |\varphi - \pi_k\varphi|dt &= \int_{I_n} |\varphi - \bar{\varphi}|dt \\ &= \int_{I_n} |\varphi(t) - \varphi(\xi)|dt = \int_{I_n} |\int_{\xi}^{t} \dot{\varphi}(s)\,ds|dt \\ &\le \int_{I_n} |\int_{I_n} \dot{\varphi}(s)\,ds|dt = k_n \cdot \int_{I_n} |\dot{\varphi}|dt,\end{aligned}$$

where we have used the mean value theorem for integrals, with $\xi \in I_n$.

Summing over n, yields the global estimate

$$\int_0^{t_N} |\varphi - \pi_k\varphi|dt = \sum_{n=1}^{N} \int_{I_n} |\varphi - \pi_k\varphi|dt \le \sum_{n=1}^{N} k_n \int_{I_n} |\dot{\varphi}|dt. \tag{6.3.16}$$

Now let $|v|_J := \max_{t\in J} |v(t)|$, then using (6.3.16) and the final form of the error representation formula (6.3.14) we have that

$$|e_N|^2 \le \sum_{n=1}^{N} \Big(|r(U)|_{I_n} k_n \int_{I_n} |\dot{\varphi}|dt\Big) \le \max_{1\le n\le N}(k_n|r(U)|_{I_n}) \int_0^{t_N} |\dot{\varphi}|dt.$$

Since, by the definition of $S(t_N)$, $\int_0^{t_N} |\dot{\varphi}|dt = |e_N| \cdot S(t_N)$, we finally get

$$|e_N|^2 \leq |e_N| S(t_N) \max_{t\in[0,t_N]} (|kr(U)|). \tag{6.3.17}$$

This completes the proof of the first assertion of the theorem.

To prove the second assertion, we transform the dual problem (6.3.5) to a forward problem, using the change of variables, viz $s = t_N - t,\ (t = t_N - s)$. Thus for $\psi(s) := \varphi(t_N - s)$, using the chain rule we get

$$\frac{d\psi}{ds} = \frac{d\varphi}{dt} \cdot \frac{dt}{ds} = -\dot{\varphi}(t_N - s). \tag{6.3.18}$$

The dual problem (6.3.5) is now reformulated as: find $\varphi(t)$ such that

$$-\dot{\varphi}(t_N - s) + a(t_N - s)\varphi(t_N - s) = 0. \tag{6.3.19}$$

The corresponding forward problem for ψ reads as follows: find $\psi(s)$ such that

$$\begin{cases} \dfrac{d\psi(s)}{ds} + a(t_N - s)\psi(s) = 0, & 0 < s \leq t_N \\ \psi(0) = \varphi(t_N) = e_N, & e_N = u_N - U_N = u(t_N) - U(t_N). \end{cases}$$

By the solution formula (6.1.1) this problem has the solution $\psi(s)$ given by

$$\begin{aligned} \psi(s) &= \psi(0) \cdot e^{-\int_0^s a(t_N-u)\,du} = [t_N - u = v,\ -du = dv] \\ &= e_N \cdot e^{-\int_{t_N}^{t_N-s} a(v)\,dv} = e_N \cdot e^{A(t_N-s)-A(t_N)}. \end{aligned}$$

Inserting back into the relation $\varphi(t) = \psi(s)$, with $t = t_N - s$, we end up with

$$\varphi(t) = e_N e^{A(t)-A(t_N)}, \quad \text{and} \quad \dot{\varphi}(t) = e_N a(t) e^{A(t)-A(t_N)}. \tag{6.3.20}$$

To estimate $S(t_N)$ we note that for $a(t) \geq 0$, the second relation in (6.3.20) yields

$$\begin{aligned} \int_0^{t_N} |\dot{\varphi}(t)|dt &= |e_N| \int_0^{t_N} a(t) e^{A(t)-A(t_N)} dt = |e_N| [e^{A(t)-A(t_N)}]_{t=0}^{t_N} \\ &= |e_N| \Big(1 - e^{A(0)-A(t_N)}\Big) = |e_N| \Big(1 - e^{-A(t_N)}\Big) \leq |e_N|, \end{aligned}$$

which gives $S(t_N) = \dfrac{\int_0^{t_N} |\dot\varphi(t)|dt}{|e_N|} \le 1$.

As for the case $|a(t)| \le \lambda$, we use again (6.3.20): $\dot\varphi(t) = a(t)e_N e^{A(t)-A(t_N)}$ and write

$$|\dot\varphi(t)| \le \lambda|e_N|e^{A(t)-A(t_N)} = \lambda|e_N|e^{\int_{t_N}^t a(s)ds} \le \lambda|e_N|e^{\lambda(t_N-t)}. \tag{6.3.21}$$

Integrating over $(0, t_N)$ we get

$$\int_0^{t_N} |\dot\varphi(t)|dt \le |e_N| \int_0^{t_N} \lambda e^{\lambda(t_N-t)}dt = |e_N|\Big[-e^{\lambda(t_N-t)}\Big]_0^{t_N} = |e_N|(-1+e^{\lambda t_N}),$$

which gives that $S(t_N) \le (-1+e^{\lambda t_N}) \le e^{\lambda t_N}$, and completes the proof of the second assertion. □

The convergence rate in (6.3.6) appears to be of order $\mathcal{O}(k)$, but is actually of order $\mathcal{O}(k^2)$ since $|r(U)| \le Ck$ if f is a smooth function. To see this convergence of order $\mathcal{O}(k^2)$, we use some further orthogonality observations in the following theorem.

Theorem 6.4 (Convergence order $\mathcal{O}(k^2)$). *For $N = 1, 2, \ldots$, and with S_N as in the previous theorem, the error for the cG(1) solution $U(t)$ satisfies*

$$|e_N| \le S(t_N) \max_{t\in[0,t_N]} \Big|k^2 \frac{d}{dt}(aU - f)\Big|. \tag{6.3.22}$$

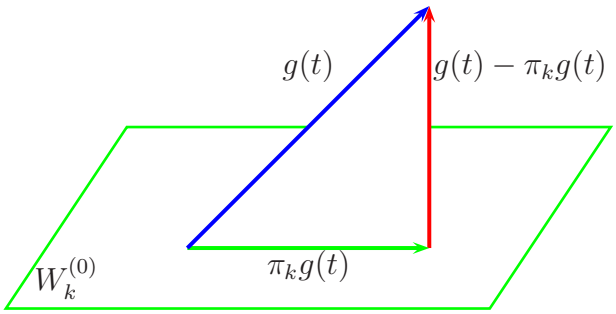


Figure 6.3: Orthogonality: $(g(t) - \pi_k g(t)) \perp$ (constants) $\forall g(t)$.

Proof. Let $\pi_k g$ be the L_2-projection of g onto the space $W_k^{(0)}$ of piecewise constant polynomials. Then, by the orthogonality relation $\Big(g(t) - \pi_k g(t)\Big) \perp$ (constants) $\forall g(t)$, since $\dot{U}(t)$ is constant on each subinterval I_n we have that $\int_0^{t_N} \dot{U}(\varphi - \pi_k\varphi)dt = 0$. Thus using the error representation formula (6.3.14), we may write

$$\begin{aligned} e_N^2 &= -\int_0^{t_N} r(U)\Big(\varphi(t) - \pi_k\varphi(t)\Big)dt = \int_0^{t_N} (f - aU - \dot{U})(\varphi - \pi_k\varphi)dt \\ &= \int_0^{t_N} (f - aU)(\varphi - \pi_k\varphi)dt - \int_0^{t_N} \dot{U}(\varphi - \pi_k\varphi)dt \\ &= -\int_0^{t_N} (aU - f)(\varphi - \pi_k\varphi)dt. \end{aligned}$$

Similarly, using the fact that $\pi_k(aU - f)$ is constant on each subinterval I_n, we have

$$\int_0^{t_N} \pi_k(aU - f)(\varphi - \pi_k\varphi)dt = 0. \tag{6.3.23}$$

Consequently we can write

$$e_N^2 = -\int_0^{t_N} \Big((aU - f) - \pi_k(aU - f)\Big)(\varphi - \pi_k\varphi)dt. \tag{6.3.24}$$

(ompare (6.3.24) with (6.3.14)). To proceed recall that for $t \in I_n$:

$$g(t) - \pi_k g(t) = g(t) - \bar{g}(t) = g(t) - g(\xi) = \int_\xi^t \dot{g}(s)\, ds,$$

where $\bar{g}$ is the mean value of g over I_n, $\xi \in I_n$ and we have used the mean value theorem for the integrals. This implies

$$|g(t) - \pi_k g(t)| \le \int_{I_n} |\dot{g}(t)|\, dt \le \max_{t\in I_n} |\dot{g}(t)| \cdot k_n, \tag{6.3.25}$$

Now we apply (6.3.6) with $r(U)$ replaced by $(aU - f) - \pi_k(aU - f)$, (6.3.25) with $g := aU - f$, and the L_2-projection error estimate to get

$$|e_N| \le S(t_N)\Big|k|(aU - f) - \pi_k(aU - f)|\Big|_{[0,t_N]} \le S(t_N)\Big|k^2 \frac{d}{dt}(aU - f)\Big|_{[0,t_N]},$$

which is the desired result. □

Remark 6.3. *Note that the estimate* (6.3.22) *is less practical than the estimate* (6.3.6) *in the sense that: it requires the cumbersome procedure of computing the time derivative of the residual.*

6.3.2 A posteriori error estimate for dG(0)

We shall now derive an a posteriori error estimate for the dG(0) method. In this case, the residual contains jump discontinuities.

Theorem 6.5. *For $N = 1, 2, \ldots$, the dG(0) solution $U(t)$ satisfies*

$$|u(t_N) - U_N| \leq S(t_N)|kR(U)|_{[0,t_N]}, \quad U_N = U(t_N) \tag{6.3.26}$$

where

$$R(U) = |f - aU| + \frac{|U_n - U_{n-1}|}{k_n} \quad \textit{for} \quad t_{n-1} < t \leq t_n. \tag{6.3.27}$$

Proof. The proof is similar to that of the $cG(1)$ approach of Theorem 6.3, but we have to take the discontinuities of $U(t)$ at the time nodes t_j into account. Below we shall first demonstrate how to rearrange the contribution from these jumps, and in this way derive a corresponding error representation formula. To this end, recalling the dual problem $-\dot{\varphi}(t) + a(t)\varphi(t) = 0$, with $\varphi(t_N) = e_N$, we may write, using integration by parts on each subinterval

$$\begin{aligned} e_N^2 &= e_N^2 + \sum_{n=1}^N \int_{t_{n-1}}^{t_n} e(-\dot{\varphi}(t) + a(t)\varphi(t))dt \\ &= e_N^2 + \sum_{n=1}^N \Big(\int_{t_{n-1}}^{t_n} (\dot{e} + ae)\varphi(t)dt - [e\varphi]_{t_{n-1}}^{t_n} \Big) \\ &= e_N^2 + \sum_{n=1}^N \int_{t_{n-1}}^{t_n} (f - aU)\varphi dt - \sum_{n=1}^N [e\varphi]_{t_{n-1}}^{t_n}, \end{aligned} \tag{6.3.28}$$

where in the last step we have used $\dot{e} + ae = \dot{u} - \dot{U} + au - aU = f - aU$ since $\dot{U} = 0$ on each subinterval (where U is constant). For a given function g we use the notation $g(t_n^-) = g_n^-$ and $g(t_{n-1}^+) = g_{n-1}^+$, and rewrite the last sum as

$$\begin{aligned} \sum_{n=1}^N [e\varphi]_{t_{n-1}}^{t_n} &= \sum_{n=1}^N \Big(e(t_n^-)\varphi(t_n^-) - e(t_{n-1}^+)\varphi(t_{n-1}^+) \Big) \\ &= \sum_{n=1}^N (e_n^-\varphi_n^- - e_{n-1}^+\varphi_{n-1}^+) = (e_1^-\varphi_1^- - e_0^+\varphi_0^+) + (e_2^-\varphi_2^- - e_1^+\varphi_1^+) \\ &\quad + \ldots + (e_{N-1}^-\varphi_{N-1}^- - e_{N-2}^+\varphi_{N-2}^+) + (e_N^-\varphi_N^- - e_{N-1}^+\varphi_{N-1}^+). \end{aligned}$$

We use the identity $\varphi_i^- = (\varphi_i^- - \varphi_i^+ + \varphi_i^+),\ i = 1, \ldots N-1$, and rewrite the contribution from the jumps as

$$\begin{aligned}-\sum_{n=1}^{N}[e\varphi]_{t_{n-1}}^{t_n} = -e_N^-\varphi_N^- + e_0^+\varphi_0^+ - e_1^-(\varphi_1^- - \varphi_1^+ + \varphi_1^+) + e_1^+\varphi_1^+ \\ - e_2^-(\varphi_2^- - \varphi_2^+ + \varphi_2^+) + e_2^+\varphi_2^+ - \ldots \\ - e_{N-1}^-(\varphi_{N-1}^- - \varphi_{N-1}^+ + \varphi_{N-1}^+) + e_{N-1}^+\varphi_{N-1}^+.\end{aligned}$$

Here the general i-th term is written as

$$\begin{aligned}- e_i^-(\varphi_i^- - \varphi_i^+ + \varphi_i^+) + e_i^+\varphi_i^+ = -e_i^-\varphi_i^- + e_i^-\varphi_i^+ - e_i^-\varphi_i^+ + e_i^+\varphi_i^+ \\ = e_i^-(\varphi_i^+ - \varphi_i^-) + \varphi_i^+(e_i^+ - e_i^-) = e_i^-[\varphi_i] + \varphi_i^+[e_i],\end{aligned}$$

with $[g] = g^+ - g^-$ representing the jump. Hence we have

$$-\sum_{n=1}^{N}[e\varphi]_{t_{n-1}}^{t_n} = -e_N^2 + e_0^+\varphi_0^+ + \sum_{n=1}^{N-1}[e_n]\varphi_n^+ + \sum_{n=1}^{N-1} e_n^-[\varphi_n]. \tag{6.3.29}$$

Inserting into (6.3.28) and using the fact that φ and u are smooth, i.e. $[\varphi_n] = [u_n] = 0$ and $[e_n] = [-U_n]$, we get the following error representation formula

$$\begin{aligned}e_N^2 &= e_N^2 + \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(f - aU)\varphi dt - e_N^2 + e_0^+\varphi_0^+ + \sum_{n=1}^{N-1}[e_n]\varphi_n^+ + \sum_{n=1}^{N-1}[\varphi_n]e_n^- \\ &= e_0^+\varphi_0^+ + \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(f - aU)\varphi dt - \sum_{n=1}^{N-1}[U_n]\varphi_n^+ \\ &= \sum_{n=1}^{N}\Big(\int_{t_{n-1}}^{t_n}(f - aU)\varphi dt - [U_{n-1}]\varphi_{n-1}^+\Big) \\ &= \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(f - aU)(\varphi - \pi_k\varphi)\,dt - [U_{n-1}](\varphi - \pi_k\varphi)_{n-1}^+,\end{aligned}$$

where we have used $e_0^+ = u_0^+ - U_0^+ = u_0 - U_0^+ = U_0^- - U_0^+ = -[U_0]$, and in the last step, once again, we use definition (6.2.10) of dG(q) with $q = 0$ and the L_2 projection $w = \pi_k\varphi \in W_k^{(0)}$ (space of piecewise constenats) of φ. The remaining part is not substantially different from the proof of Theorem 6.3, and is therefore omitted. □

6.3.3 Adaptivity for dG(0)

To guarantee a desired *local* bound for the error of the dG(0) approximation $U(t)$, such as

$$|e_n| = |u(t_n) - U(t_n)| \leq TOL, \tag{6.3.30}$$

where TOL is a given tolerance, we seek to determine the time step $k_n = t_n - t_{n-1}$ so that, using (6.3.26), locally, we have

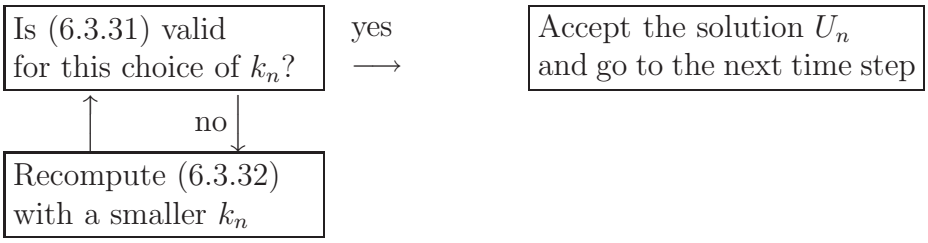
$$S(t_N) \max_{t \in I_n} |k_n R(U)| \leq TOL, \qquad n = 1, 2, \ldots, N. \tag{6.3.31}$$

Adaptivite algorithm

(i) Compute U_n from U_{n-1} using a predicted step k_n, and the relation

$$\int_{t_{n-1}}^{t_n} aU_n dt + U_n = \int_{t_{n-1}}^{t_n} f dt + U_{n-1}. \tag{6.3.32}$$

(ii) Compute $|kR(U)|_{I_n} := \max_{t \in I_n} |k_n R(U)|$, where $R(U) = |f - aU| + \frac{|U_n - U_{n-1}|}{k_n}$ in I_n, and follow chart:



6.4 A priori error analysis

In this section we shall derive a priori error estimates for the $dG(0)$ method. First we consider the case of general $a(t)$ (but we simplify, the analysis by assuming that a is a constant) and then we consider the parabolic case $a(t) \geq 0$, where we perform a refined analysis. The corresponding a priori error estimate for the $cG(1)$ method is obtained similarly. The weaker stability properties of $cG(1)$, however, do not allow a refined analysis in the parabolic case.

6.4.1 A priori error estimates for the dG(0) method

The dG(0) method for $\dot{u} + au = f$, with a constant a, is formulated as: find $U = U(t),\ t \in I_n$, such that

$$\int_{t_{n-1}}^{t_n} \dot{U}dt + a\int_{t_{n-1}}^{t_n} Udt = \int_{I_n} fdt, \quad n = 1, 2, \ldots \tag{6.4.1}$$

Note that $U(t) = U_n$ is constant for $t \in I_n$. Let $U_n = U(t_n),\ U_{n-1} = U(t_{n-1})$ and $k_n = t_n - t_{n-1}$, then

$$\int_{t_{n-1}}^{t_n} \dot{U}dt + a\int_{t_{n-1}}^{t_n} Udt = U_n - U_{n-1} + ak_nU_n.$$

Hence with a given initial data $u(0) = u_0$, the equation (6.4.1) is written as

$$U_n - U_{n-1} + ak_nU_n = \int_{I_n} fdt. \quad n = 1, 2, \ldots \qquad U_0 = u_0. \tag{6.4.2}$$

For the *exact* solution $u(t)$ of $\dot{u} + au = f$, the same procedure yields

$$u(t_n) - u(t_{n-1}) + ak_nu(t_n) = \int_{I_n} fdt + ak_nu(t_n) - a\int_{t_{n-1}}^{t_n} u(t)dt, \tag{6.4.3}$$

where we have moved the term $a\int_{t_{n-1}}^{t_n} u(t)dt$ to the right hand side and added $ak_nu(t_n)$ to both sides. Thus from (6.4.2) and (6.4.3) we have that, denoting $u_n = u(t_n)$,

$$\begin{cases} (1 + k_na)U_n = U_{n-1} + \displaystyle\int_{I_n} fdt, \\ (1 + k_na)u_n = u_{n-1} + \displaystyle\int_{I_n} fdt + ak_nu_n - a\int_{t_{n-1}}^{t_n} u(t)dt. \end{cases} \tag{6.4.4}$$

Now let $e_n = u_n - U_n$, then $e_{n-1} = u_{n-1} - U_{n-1}$, and subtracting the first equation in (6.4.4) from the second one yields

$$e_n = (1+k_n a)^{-1}(e_{n-1}+\rho_n), \qquad where \quad \rho_n := ak_n u_n - a\int_{t_{n-1}}^{t_n} u(t)dt. \tag{6.4.5}$$

Thus, to estimate e_n we use an iteration procedure combined with an estimate for ρ_n.

Lemma 6.1. *We have that*

$$|\rho_n| \le \frac{1}{2}|a|k_n^2 \max_{I_n}|\dot{u}(t)| \tag{6.4.6}$$

Proof. Evidently,

$$\begin{aligned} |\rho_n| &= \Big|ak_n u_n - a\int_{t_{n-1}}^{t_n} u(t)\,dt\Big| = |a|k_n\Big|u_n - \frac{1}{k_n}\int_{I_n} u(t)\,dt\Big| \\ &= |a|\Big|\int_{t_{n-1}}^{t_n}(u_n - u(t))\,dt\Big| = |a|\Big|\int_{t_{n-1}}^{t_n}\int_t^{t_n}\dot{u}(s)\,ds\,dt\Big| \\ &\le |a|\max_{I_n}|\dot{u}(t)|\cdot\int_{t_{n-1}}^{t_n}(t_n - t)\,dt = \frac{1}{2}|a|k_n^2\cdot\max_{I_n}|\dot{u}(t)|. \end{aligned} \tag{6.4.7}$$

□

To simplify the estimate for e_n we split, and gather, the proof of technical details in the following lemma:

Lemma 6.2. *For $k_n|a| \le 1/2$, and $n \ge 1$ we have that*

(i) $(1 - k_n|a|)^{-1} \le e^{2k_n|a|}$,

(ii) $|e_N| \le \frac{1}{2}\sum_{n=1}^{N}(e^{2|a|\tau_n}|a|k_n)\Big(\max_{1\le n\le N} k_n|\dot{u}|_{I_n}\Big), \qquad \tau_n = t_N - t_{n-1}.$

(iii) $\sum_{n=1}^{N} e^{2|a|\tau_n}|a|k_n \le e\int_0^{t_N}|a|e^{2|a|\tau}d\tau.$

We postpone the proof of this lemma and first show that using $(i)-(iii)$ we can prove the following a priori error estimate

Theorem 6.6. *If $k_n|a| \leq \frac{1}{2}, n \geq 1$ then the error for the $dG(0)$ approximation U satisfies*

$$|u(t_N) - U(t_N)| = |e_N| \leq \frac{e}{4}\Big(e^{2|a|t_N} - 1\Big) \max_{1\leq n\leq N} k_n|\dot{u}(t)|_{I_n}. \tag{6.4.8}$$

Proof. Using estimates (ii) and (iii) of the above lemma we have that

$$\begin{aligned}|e_N| &\leq \frac{1}{2}\sum_{n=1}^{N}(e^{2|a|\tau_n}|a|k_n) \max_{1\leq n\leq N} k_n|\dot{u}|_{I_n} \leq \frac{1}{2}\Big(e\int_0^{t_N} |a|e^{2|a|\tau}d\tau\Big) \max_{1\leq n\leq N} k_n|\dot{u}|_{I_n} \\ &= \frac{1}{2}e\Big[\frac{e^{2|a|\tau}}{2}\Big]_0^{t_N} \max_{1\leq n\leq N} k_n|\dot{u}(t)|_{I_n} = \frac{e}{4}\Big(e^{2|a|t_N} - 1\Big) \max_{1\leq n\leq N} k_n|\dot{u}(t)|_{I_n},\end{aligned}$$

and the proof is complete. □

Note that the stability factor $\frac{e}{4}\Big(e^{2|a|t_N}-1\Big)$ grows exponentially depending on $|a|$ and t_N, hence this result may not be satisfactory at all. We deal with this question below in the parabolic case $a(t) \geq 0$.

Remark 6.4. *The a priori error estimate for the cG(1) method is similar to that of Theorem 6.6, with $k_n|\dot{u}|$ replaced by $k_n^2|\ddot{u}|$.*

We now return to the proof of our technical results $(i) - (iii)$:

Proof of Lemma 6.3. (i) For $0 < x := k_n|a| \leq 1/2$, we have that $1/2 \leq 1-x < 1$ implies $0 \leq 1-2x < 1$. We may multiply both sides of $\dfrac{1}{1-x} < e^{2x}$ (which is the result we want to prove) by $1 - x \geq 1/2 > 0$ to obtain the equivalent relation

$$f(x) := (1-x)e^{2x} - 1 > 0. \tag{6.4.9}$$

Note that since $f(0) = 0$ and $f'(x) = (1-2x)e^{2x} \geq 0$ (equality for $x = \frac{1}{2}$), thus (6.4.9), and hence the relation (i) is valid.

To prove (ii) we recall that $e_n = (1+k_na)^{-1}(e_{n-1}+\rho_n)$. To deal with the factor $(1+k_na)^{-1}$ we use $k_n|a| \leq \dfrac{1}{2}$, then by (i) $(1+k_na)^{-1} \leq (1-k_n|a|)^{-1} \leq e^{2k_n|a|}$, $n \geq 1$. Thus

$$|e_N| \leq \frac{1}{1-k_N|a|}|e_{N-1}| + \frac{1}{1-k_N|a|}|\rho_N| \leq |e_{N-1}|e^{2k_N|a|} + |\rho_N|e^{2k_N|a|}. \tag{6.4.10}$$

Relabeling, N to $N-1$,

$$|e_{N-1}| \le |e_{N-2}|e^{2k_{N-1}|a|} + |\rho_{N-1}|e^{2k_{N-1}|a|} = e^{2k_{N-1}|a|}\Big(|e_{N-2}| + |\rho_{N-1}|\Big),$$

and inserting into (6.4.10) yields

$$|e_N| \le e^{2k_N|a|}e^{2k_{N-1}|a|}\Big(|e_{N-2}| + |\rho_{N-1}|\Big) + |\rho_N|e^{2k_N|a|}. \tag{6.4.11}$$

Similarly we have $|e_{N-2}| \le e^{2k_{N-2}|a|}\Big(|e_{N-3}| + |\rho_{N-2}|\Big)$. Now iterating (6.4.11) and using the fact that $e_0 = 0$ we get,

$$\begin{aligned}|e_N| \le & e^{2k_N|a|}e^{2k_{N-1}|a|}e^{2k_{N-2}|a|}|e_{N-3}| + e^{2k_N|a|}e^{2k_{N-1}|a|}e^{2k_{N-2}|a|}|\rho_{N-2}| \\ & + e^{2k_N|a|}e^{2k_{N-1}|a|}|\rho_{N-1}| + |\rho_N|e^{2k_N|a|} \le \cdots \le \\ \le & e^{2|a|\sum_{n=1}^{N}k_n}|e_0| + \sum_{n=1}^{N} e^{2|a|\sum_{m=n}^{N}k_m}|\rho_n| = \sum_{n=1}^{N} e^{2|a|\sum_{m=n}^{N}k_m}|\rho_n|.\end{aligned}$$

Recalling (6.4.6): $|\rho_n| \le \frac{1}{2}|a|k_n^2 \max_{I_n}|\dot{u}(t)|$, we have the error estimate

$$|e_N| \le \Big(\sum_{n=1}^{N} e^{2|a|\sum_{m=n}^{N}k_m}\Big)\frac{1}{2}|a|k_n^2 \max_{I_n}|\dot{u}(t)|. \tag{6.4.12}$$

Note that

$$\sum_{m=n}^{N} k_m = (t_n - t_{n-1}) + (t_{n+1} - t_n) + (t_{n+2} - t_{n+1}) + \ldots + (t_N - t_{N-1}) = t_N - t_{n-1}.$$

Hence, since $\tau_n = t_N - t_{n-1}$, we have shown assertion (ii), i.e.

$$|e_N| \le \sum_{n=1}^{N} e^{2|a|(t_N - t_{n-1})}\frac{1}{2}|a|k_n^2 \max_{I_n}|\dot{u}(t)| \le \frac{1}{2}\sum_{n=1}^{N}(e^{2|a|\tau_n}|a|k_n)\Big(\max_{1\le n\le N} k_n|\dot{u}|_{I_n}\Big).$$

(iii) To prove this part we note that

$$\tau_n = t_N - t_{n-1} = (t_N - t_n) + (t_n - t_{n-1}) = \tau_{n+1} + k_n, \tag{6.4.13}$$

and since $|a|k_n \leq 1/2$ we have $2|a|\tau_n = 2|a|\tau_{n+1} + 2|a|k_n \leq 2|a|\tau_{n+1} + 1$. Further for $\tau_{n+1} \leq \tau \leq \tau_n$, we can write

$$\begin{aligned} e^{2|a|\tau_n}\, k_n &= \int_{\tau_{n+1}}^{\tau_n} e^{2|a|\tau_n} d\tau \leq \int_{\tau_{n+1}}^{\tau_n} e^{(2|a|\tau_{n+1}+1)} d\tau \\ &= \int_{\tau_{n+1}}^{\tau_n} e^1 \cdot e^{2|a|\tau_{n+1}} d\tau \leq e \int_{\tau_{n+1}}^{\tau_n} e^{2|a|\tau} d\tau. \end{aligned} \tag{6.4.14}$$

Multiplying (6.4.14) by $|a|$ and summing over n we get

$$\begin{aligned} \sum_{n=1}^{N} e^{2|a|\tau_n}|a|k_n &\leq e\Big(\sum_{n=1}^{N} \int_{\tau_{n+1}}^{\tau_n} e^{2|a|\tau} d\tau\Big)|a| \\ &= e \int_{\tau_{N+1}}^{\tau_1} e^{2|a|\tau}|a| d\tau = e \int_0^{t_N} |a| e^{2|a|\tau} d\tau, \end{aligned} \tag{6.4.15}$$

which is the desired result and the proof of (iii) is complete.

□

6.5 The parabolic case $(a(t) \geq 0)$

The interest in parabolic case is due to the fact that it does not accumulate the error. This can be seen from the fundamental solution with the presence of the deceasing multiplicative factor $e^{-A(t)}$. Below we state and proof some basic estimate for this case:

Theorem 6.7. *Consider the $dG(0)$ approximation U, of $\dot{u} + au = f$, with $a(t) \geq 0$. Assume that $k_j|a|_{I_j} \leq \frac{1}{2}$; $\forall j$, then we have the error estimates*

$$|u(t_N) - U_N| \leq \begin{cases} 3e^{2\lambda t_N} \max\limits_{0\leq t\leq t_N} |k\dot{u}|, & \text{for } |a(t)| \leq \lambda, \\ 3 \max\limits_{0\leq t\leq t_N} |k\dot{u}|, & \text{for } a(t) \geq 0. \end{cases} \tag{6.5.1}$$

Sketch of the proof. Let $e = u - U = (u - \pi_k u) + (\pi_k u - U) := \tilde{e} + \bar{e}$, where $\tilde{e}$ is the projection error with $\pi_k u$ being the L_2-projection of u into $W_k^{(0)}$. Since the L_2-projection error is of the form (6.5.1), hence it suffices to estimate the

discrete error $\bar{e}$. To this end we shall use the following *discrete dual problem (DDP)*:
Find $\Phi \in W_k^{(0)}$, such that for $n = N, N-1, \ldots, 1$,

$$(DDP) \qquad \begin{cases} \displaystyle\int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi)v dt - [\Phi_n]v_n = 0, \quad \forall v \in W_k^{(0)} \\ \Phi_N^+ = \Phi_{N+1} = (\pi_k u - U)_N :\equiv \bar{e}_N. \end{cases} \tag{6.5.2}$$

Let now $v = \bar{e}$, then

$$|\bar{e}_N|^2 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi)\bar{e} dt - \sum_{n=1}^{N-1} [\Phi_n]\bar{e}_n + \Phi_N \bar{e}_N. \tag{6.5.3}$$

We use $\bar{e} = (\pi_k u - U) = (\pi_k u - u + u - U)$ and rewrite (6.5.3) as

$$\begin{aligned} |\bar{e}_N|^2 = &\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} [-\dot{\Phi} + a(t)\Phi](\pi_k u - u + u - U)\, dt \\ &- \sum_{n=1}^{N-1} [\Phi_n](\pi_k u - u + u - U)_n + \Phi_N(\pi_k u - u + u - U)_N. \end{aligned}$$

By Galerkin orthogonality, the total contribution from $u - U$ terms vanishes, and we end up with the error terms involving only $\pi_k u - u$. Hence, due to the fact that $\dot{\Phi} = 0$ on each subinterval we have the following error representation formula

$$\begin{aligned} |\bar{e}_N|^2 &= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi)(\pi_k u - u) dt - \sum_{n=1}^{N-1} [\Phi_n](\pi_k u - u)_n + \Phi_N(\pi_k u - u)_N \\ &= -\int_0^{t_N} (a(t)\Phi)(u - \pi_k u)\, dt + \sum_{n=1}^{N-1} [\Phi_n](u - \pi_k u)_n - \Phi_N(u - \pi_k u)_N. \end{aligned}$$

□

To continue we shall need the following results:

Lemma 6.3. *If* $|a(t)| \leq \lambda, \forall t \in (0, t_N)$ *and* $k_j |a|_{I_j} \leq \frac{1}{2}$, *for* $j = 1, \ldots, N$, *then the solution of the discrete dual problem satisfies*

(i) $\quad |\Phi_n| \leq e^{2\lambda(t_N - t_{n-1})} |\bar{e}_N|,$

(ii) $$\sum_{n=1}^{N-1} |[\Phi_n]| \le e^{2\lambda t_N} |\bar{e}_N|,$$

(iii) $$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} |a(t)\Phi_n| dt \le e^{2\lambda t_N} |\bar{e}_N|,$$

(iv) $$Max\Big(|\Phi_n|, \sum_{n=1}^{N-1} |[\Phi_n]|, \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} a(t)|\Phi_n|\, dt\Big) \le |\bar{e}_N|, \quad a(t) \ge 0.$$

Proof. We show the last estimate (iv), (the proofs of (i)-(iii) are similar to that of the stability factor in the previous theorem). Consider the discrete, local dual problem with $v \equiv 1$:

$$\begin{cases} \displaystyle\int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi) dt - [\Phi_n] = 0, \\ \Phi_{N+1} = (\pi_k u - U)_N :\equiv \bar{e}_N. \end{cases} \tag{6.5.4}$$

For dG(0) and for a constant Φ, this becomes

$$\begin{cases} -\Phi_{n+1} + \Phi_n + \Phi_n \displaystyle\int_{t_{n-1}}^{t_n} a(t)\, dt = 0, \quad n = N, N-1, \ldots, 1 \\ \Phi_{N+1} = \bar{e}_N, \qquad \Phi_n = \Phi|_{I_n}. \end{cases} \tag{6.5.5}$$

By iterating we get

$$\Phi_n = \prod_{j=n}^{N} \Big(1 + \int_{I_j} a(t) dt\Big)^{-1} \Phi_{N+1}. \tag{6.5.6}$$

For $a(t) \ge 0$ we have $\Big(1 + \int_{I_j} a(t) dt\Big)^{-1} \le 1$, thus (6.5.6) implies that

$$|\Phi_n| \le \Phi_{N+1} = |\bar{e}_N|. \tag{6.5.7}$$

Further we have using (6.5.6) that

$$\Phi_{n-1} = \prod_{j=n-1}^{N} \Big(1 + \int_{I_j} a(t) dt\Big)^{-1} \Phi_{N+1} = \Big(1 + \int_{I_{n-1}} a(t) dt\Big)^{-1} \Phi_n \le \Phi_n$$

which implies that

$$[\Phi_n] = \Phi_n^+ - \Phi_n^- = \Phi_{n+1} - \Phi_n \geq 0. \tag{6.5.8}$$

Now, since by (6.5.6),$\Phi_1 \leq \Phi_{N+1}$, thus

$$\begin{aligned} \sum_{n=1}^{N} |[\Phi_n]| &= \Phi_{N+1} - \Phi_N + \Phi_N - \Phi_{N-1} + \ldots + \Phi_2 - \Phi_1 \\ &= \Phi_{N+1} - \Phi_1 \leq \Phi_{N+1} = |\bar{e}_N|. \end{aligned} \tag{6.5.9}$$

Finally in the discrete equation:

$$\int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi)v dt - [\Phi_n]v_n = 0, \qquad \forall v \in W_k^{(0)}, \tag{6.5.10}$$

we have $v \equiv 1$ and $\dot{\Phi} \equiv 0$ for the dG(0). Hence (6.5.10) can be rewritten as

$$\int_{t_{n-1}}^{t_n} a(t)\Phi_n dt = [\Phi_n]. \tag{6.5.11}$$

Summing over n, this gives that

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} a(t)\Phi_n dt = \sum_{n=1}^{N} [\Phi_n] \leq |\bar{e}_N|. \tag{6.5.12}$$

Combining (6.5.7), (6.5.9), and (6.5.12) completes the proof of (iv). □

Below we state some useful results (their proofs are standard, however, lengthy and therefore are omitted).

Quadrature rule. Assume that a is constant. Then the error representation formula in the proof of the Theorem 6.5 for $dG(0)$, combined with the quadrature rule for f reads as follows:

$$\begin{aligned} e_N^2 = \sum_{n=1}^{N} \Big(&\int_{t_{n-1}}^{t_n} (f - aU)(\varphi - \pi_k\varphi) dt - [U_{n-1}](\varphi - \pi_k\varphi)_{n-1}^+ \\ &+ \int_{t_{n-1}}^{t_n} f\pi_k\varphi dt - (\overline{f\pi_k\varphi})_n k_n \Big), \end{aligned} \tag{6.5.13}$$

where the last two terms represent the quadrature error. Here, e.g. for the endpoint-rule $\bar{g}_n = g(t_n)$, whereas for the midpoint-rule $\bar{g}_n := g(t_{(n-1/2)})$. We also recall the *weak stability factor* $\tilde{S}(t_N) := \int_0^{t_N} |\varphi| dt/|e_N|$, where φ is the solution of the dual problem

$$-\dot{\varphi} + a\varphi = 0, \quad \text{for } t_N > t \geq 0 \qquad \varphi(t_N) = e_N.$$

Recall that $\pi_k\varphi$ is piecewise constant with the *mass-preserving* property

$$\int_{I_n} |\pi_k\varphi(t)| dt = \int_{I_n} |\varphi(t)| dt.$$

It is easy to prove the following relations between the two stability factors:

$$\tilde{S}(t_N) \leq t_N(1 + S(t_N)),$$

and if $a > 0$ is sufficiently small, then we have indeed $\tilde{S}(t_N) >> S(t_N)$.

Finally we state (without proof, see Johnson et al for further details) the following dG(0) estimate

Theorem 6.8 (The modified a posteriori estimate for dG(0))**.** *For each $N = 1, 2, \ldots$, the dG(0) approximation $U(t)$ computed using a quadrature rule on the source term f satisfies*

$$|u(t_n) - U_n| \leq S(t_n)|kR(U)|_{(0,t_N)} + \tilde{S}(t_N)C_j|k^j f^{(j)}|_{(0,t_N)}, \tag{6.5.14}$$

where

$$R(U) = \frac{|U_n - U_{n-1}|}{k_n} + |f - aU|, \quad \text{on } I_n \tag{6.5.15}$$

and $j = 1$ for the rectangle rule, $j = 2$ for the midpoint rule, $C_1 = 1$, $C_2 = 1/2$, $f^{(1)} := \dot{f} := df/dt$ and $f^{(2)} := \ddot{f} := d^2f/dt^2$.

6.5.1 Some examples of error estimates

In this part we shall derive some simple versions of the error estimates above

Example 6.3. *Let U be the cG(1) approximation of the solution u for the initial value problem*

$$\dot{u} + u = f, \quad t > 0, \quad u(0) = u_0. \tag{6.5.16}$$

Then we have that

$$|(u-U)(T)| \leq \max_{[0,T]} |k(f-\dot{U}-U)|, \tag{6.5.17}$$

where k is the time step.

Proof. Evidently the error $e = u - U$ satisfies the Galerkin orthogonality

$$\int_0^T (\dot{e}+e)v dt = 0, \qquad \text{for all piecewise constants } v(t). \tag{6.5.18}$$

Let φ be the solution of the dual equation

$$-\dot{\varphi}+\varphi = 0, \quad t < T, \quad \varphi(T) = e(T), \tag{6.5.19}$$

then $\varphi(t) = e(T)e^{t-T}$. Further,

$$|e(T)|^2 = e(T)\cdot e(T) + \int_0^T e(-\dot{\varphi}+\varphi)dt = e(T)\cdot e(T) - \int_0^T e\dot{\varphi}\, dt + \int_0^T e\varphi dt.$$

Integration by parts yields

$$\int_0^T e\dot{\varphi}dt = [e\cdot\varphi]_{t=0}^T - \int_0^T \dot{e}\varphi dt = e(T)\varphi(T) - e(0)\varphi(0) - \int_0^T \dot{e}\varphi dt.$$

Hence, using $\varphi(T) = e(T)$, and $e(0) = 0$, we have

$$\begin{aligned}|e(T)|^2 &= e(T)\cdot e(T) - e(T)\cdot e(T) + \int_0^T \dot{e}\varphi\, dt + \int_0^T e\varphi\, dt = \int_0^T (\dot{e}+e)\varphi\, dt \\ &= \int_0^T (\dot{e}+e)(\varphi - v)dt = \int_0^T \Big(\dot{u}+u-\dot{U}-U\Big)(\varphi - v)dt.\end{aligned}$$

We have that $r(U) := \dot{U} + U - f$, is the residual and

$$|e(T)|^2 = -\int_0^T r(U)\cdot(\varphi - v)dt \leq \max_{[0,T]} |k\cdot r(U)| \int_0^T \frac{1}{k}|\varphi - v|dt. \tag{6.5.20}$$

Recall that

$$\int_I k^{-1}|\varphi - v|dx \leq \int_I |\dot{\varphi}|dx. \tag{6.5.21}$$

Further $-\dot{\varphi} + \varphi = 0$ implies $\dot{\varphi} = \varphi$, and $\varphi(t) = e(T)e^{t-T}$. Thus

$$\begin{aligned} |e(T)|^2 &\leq \max_{[0,T]} |k\, r(U)| \int_0^T |\dot{\varphi}| dt = \max_{[0,T]} |k\, r(U)| \int_0^T |\varphi(t)|\, dt \\ &\leq \max_{[0,T]} |kr(U)| e(T)| \int_0^T e^{t-T} dt, \end{aligned} \tag{6.5.22}$$

and since

$$\int_0^T e^{t-T} dt = [e^{t-T}]_0^T = e^0 - e^{-T} = 1 - e^{-T} \leq 1, \quad T > 0.$$

Finally, we end up with the desired result:

$$|e(T)| \leq \max_{[0,T]} |k\, r(U)|.$$

and the proof is complete. □

Problem 6.1. *Generalize the above example to the problem $\dot{u} + au = f$, with $a =$ positive constant. Is the statement of this example valid for $\dot{u} - u = f$?*

Problem 6.2. *Study the dG(0)-case for $\dot{u} + au = f, \quad a > 0$*

Example 6.4. *Let $\dot{u} + u = f, t > 0$. Show for the cG(1)-approximation $U(t)$ that*

$$|(u - U)(T)| \leq \max_{[0,T]} |k^2 \ddot{u}| T. \tag{6.5.23}$$

Sketch of the proof via the dual equation. Let φ be the solution for the dual problem

$$\dot{\varphi} + \varphi = 0, \ t < T, \quad \varphi(T) = e(T),$$

and define $\rho := u - \hat{u}$ and $\Theta := \hat{u} - U$, where $\hat{u}$ is the piecewise linear interpolant of u. Then we may compute the error at time T, as

$$\begin{aligned} |e(T)|^2 = |\Theta(T)|^2 &= \Theta(T)\varphi(T) + \int_0^T \bar{\Theta}(-\dot{\Phi} + \Phi)\, dt = \int_0^T (\dot{\Theta} + \Theta)\bar{\Phi} dt \\ &= -\int_0^T (\dot{\rho} + \rho)\bar{\Phi}\, dt = -\int_0^T \rho \cdot \bar{\Phi}\, dt \leq \max_{[0,T]} |k^2 \ddot{u}| \int_0^T |\bar{\Phi}|\, dt \\ &\leq \max_{[0,T]} |k^2 \ddot{u}| T |e(T)|. \end{aligned}$$

Here Φ is cG(1)-approximation of φ such that $\int_0^T v(-\dot{\Phi} + \Phi)\, dt = 0$ for all piecewise constant $v(t)$, and $\bar{w} =$ is the piecewise constant mean value. The details are left to the reader. □

6.6 Exercises

Problem 6.3. *(a) Derive the stiffness matrix and load vector in piecewise polynomial (of degree q) approximation for the following ODE in population dynamics,*

$$\begin{cases} \dot{u}(t) = \lambda u(t), & \text{for } 0 < t \leq 1, \\ u(0) = u_0. \end{cases}$$

(b) Let $\lambda = 1$ and $u_0 = 1$ and determine the approximate solution $U(t)$, for $q = 1$ and $q = 2$.

Problem 6.4. *Consider the initial value problem*

$$\dot{u}(t) + a(t)u(t) = f(t), \quad 0 < t \leq T, \qquad u(0) = u_0.$$

Show that if $a(t) = 1$, $f(t) = 2\sin(t)$, then we have
$u(t) = \sin(t) - \cos(t) = \sqrt{2}\sin(t - \pi/2)$.

Problem 6.5. *Compute the solution for*

$$\dot{u}(t) + a(t)u(t) = t^2, \quad 0 < t \leq T, \qquad u(0) = 1,$$

corresponding to
(a) $a(t) = 4$, *(b)* $a(t) = -t$.

Problem 6.6. *Compute the cG(1) approximation for the differential equations in the above problem. In each case, determine the condition on the step size that guarantees that U exists.*

Problem 6.7. *Without using fundamental theorem, prove that if $a(t) \geq 0$ then, a continuously differentiable solution of* (6.0.1) *is unique.*

Problem 6.8. *Consider the initial value problem*

$$\dot{u}(t) + a(t)u(t) = f(t), \quad 0 < t \leq T, \qquad u(0) = u_0.$$

Show that for $a(t) > 0$, and for $N = 1, 2, \ldots$, the piecewise linear approximate solution U for this problem satisfies the a posteriori error estimate

$$|u(t_N) - U_N| \leq \max_{[0,t_N]} |k(\dot{U} + aU - f)|, \quad k = k_n, \text{ for } t_{n-1} < t \leq t_n.$$

Problem 6.9. *Consider the initial value problem:*

$$\dot{u}(t) + au(t) = 0, \quad t > 0, \quad u(0) = 1.$$

a) Let $a = 40$, and the time step $k = 0.1$. Draw the graph of $U_n := U(nk)$, $k = 1, 2, \ldots,$ approximating u using (i) explicit Euler, (ii) implicit Euler, and (iii) Crank-Nicholson methods.

b) Consider the case $a = i$, $(i^2 = -1)$, having the complex solution $u(t) = e^{-it}$ with $|u(t)| = 1$ for all t. Show that this property is preserved in Crank-Nicholson approximation, (i.e. $|U_n| = 1$), but NOT in any of the Euler approximations.

Problem 6.10. *Consider the initial value problem*

$$\dot{u}(t) + au(t) = 0, \quad t > 0, \quad u(0) = u_0, \quad (a = \text{ constant}).$$

Assume a constant time step k and verify the iterative formulas for $dG(0)$ and $cG(1)$ approximations U and $\tilde{U}$, respectively: i.e.

$$U_n = \Big(\frac{1}{1+ak}\Big)^n u_0, \qquad \tilde{U}_n = \Big(\frac{1-ak/2}{1+ak/2}\Big)^n u_0.$$

Problem 6.11. *Assume that*

$$\int_{I_j} f(s)\, ds = 0, \qquad \text{for } \ j = 1, 2, \ldots,$$

where $I_j = (t_{j-1}, t_j)$, $t_j = jk$ with k being a positive constant. Prove that if $a(t) \geq 0$, then the solution for (6.0.1) satisfies

$$|u(t)| \leq e^{-A(t)}|u_0| + \max_{0\leq s\leq t} |kf(s)|.$$

Problem 6.12. *Let U be the $cG(1)$ approximation of u satisfying the initial value problem*

$$\dot{u} + au = f, \quad t > 0, \qquad u(0) = u_0.$$

Let k be the time step and show that for $a = 1$,

$$|(u - U)(T)| \leq \min\Big(||k(f - \dot{U} - U)||_{L^\infty[0,T]}, T||k^2\ddot{u}||_{L^\infty[0,T]}\Big).$$

Problem 6.13. *Consider the scalar boundary value problem*

$$\dot{u}(t) + a(t)u(t) = f(t), \quad t > 0, \qquad u(0) = u_0.$$

(a) Show that for $a(t) \geq a_0 > 0$, we have the stability estimate

$$|u(t)| \leq e^{-a_0 t}\left(|u_0| + \int_0^t e^{a_0 s}|f(s)|\,ds\right)$$

(b) Formulate the $cG(1)$ method for this problem, and show that the condition $\frac{1}{2}a_0 k > -1$, where k is the time step, guarantees that the method is operational, i.e. no zero division occurs.

(c) Assume that $a(t) \geq 0$, $f(t) \equiv 0$, and estimate the quantity $\frac{\int_0^T |\dot{u}|\,dt}{|u_0|}$.

Problem 6.14. *Consider the initial value problem ($u = u(x,t)$)*

$$\dot{u} + Au = f, \quad t > 0; \qquad u(t=0) = u_0.$$

Show that if there is a constant $\alpha > 0$ such that

$$(Av, v) \geq \alpha||v||^2, \qquad \forall v,$$

then the solution u of the initial value problem satisfies the stability estimate

$$||u(t)||^2 + \alpha\int_0^t ||u(s)||^2\,ds \leq ||u_0||^2 + \frac{1}{\alpha}\int_0^t ||f(s)||^2\,ds.$$

Problem 6.15. *Formulate a continuous Galerkin method using piecewise polynomials based on the original global Galerkin method.*

Problem 6.16. *Formulate the dG(1) method for the differential equations specified in Problem 6.5.*

Problem 6.17. *Write out the a priori error estimates for the equations specified in Problem 6.5.*

Problem 6.18. *Use the a priori error bound to show that the residual of the dG(0) approximation satisfies $\mathcal{R}(U) = \mathcal{O}(1)$.*

Problem 6.19. *Prove the following stability estimate for the dG(0) method described by* (6.2.10)*,*

$$|U_N|^2 + \sum_{n=0}^{N-1} |[U-n]|^2 \leq |u_0|^2.$$

Chapter 7

Initial Boundary Value Problems in 1d

A large class of phenomena in nature, science and technology, such as seasonal periods, heat distribution, wave propagation, etc, are varying both in space and time. To describe these phenomena in a physical domain requires the knowledge of their initial status, as well as information on the boundary of the domain, or asymptotic behavior in the case of unbounded domains. Problems that model such properties are called initial boundary value problems. In this chapter we shall study the two most important equations of this type: namely, the heat equation and the wave equation in one space dimension. We also address (briefly) the one-space dimensional time-dependent convection-diffusion problem.

7.1 Heat equation in 1d

In this section we focus on some basic L_2-stability and finite element error estimates for the one-space dimensional heat equation. In Chapter 1 we derived the one-dimensional stationary heat equation. A general discussion on the heat equation can be found in our *Lecture Notes in Fourier Analysis*. Here, to illustrate, we consider an example of an initial boundary value problem

(IBVP) for the one-dimensional heat equation, viz

$$\begin{cases} \dot{u} - u'' = f(x,t), & 0 < x < 1, \quad t > 0, \\ u(x,0) = u_0(x), & 0 < x < 1, \\ u(0,t) = u_x(1,t) = 0, & t > 0, \end{cases} \tag{7.1.1}$$

where we have used the following notation

$$\dot{u} := u_t = \frac{\partial u}{\partial t}, \quad u' := u_x = \frac{\partial u}{\partial x}, \quad u'' := u_{xx} = \frac{\partial^2 u}{\partial x^2}.$$

Note that the partial differential equation in (7.1.1) contains three derivatives, two derivatives in space and one derivative in time. This yields three degrees of freedom (in the sense that, to recover the solution u from the equation (7.1.1), it is necessary to integrate twice in space and once in time, where each integration introduces an arbitrary "constant".) Therefore, to determine a unique solution to the heat equation, it is necessary to supply three data: here two boundary conditions corresponding to the two spatial derivatives in u'' and an initial condition corresponding to the time derivative in $\dot{u}$.

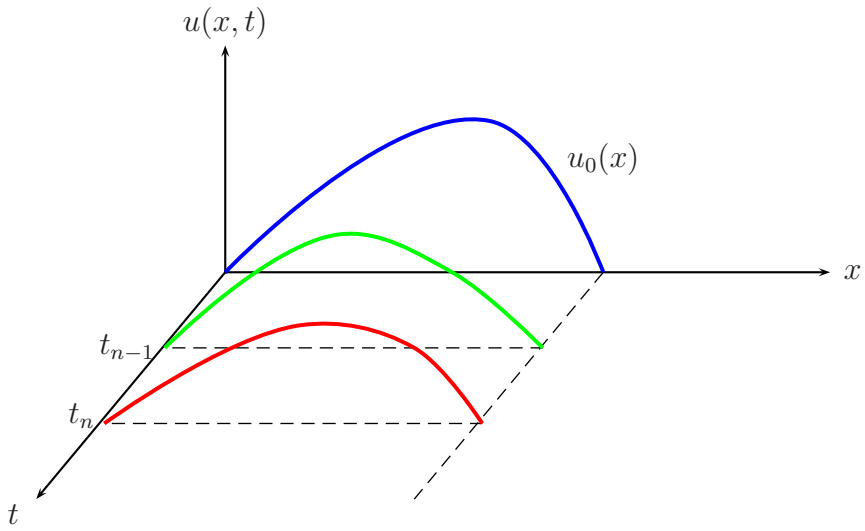


Figure 7.1: A decreasing temperature profile with data $u(0,t) = u(1,t) = 0$.

Below we give an example, where these concepts are described in physical terms:

Example 7.1. *Describe the physical meaning of the functions and parameters in the problem* (7.1.1)*, when* $f = 20 - u$*.*

Answer: The problem is an example of heat conduction where

$u(x,t)$,	means the temperature at the point x and time t.
$u(x,0) = u_0(x)$,	is the initial temperature at time $t = 0$.
$u(0,t) = 0$,	means fixed temperature at the boundary point $x = 0$.
$u'(1,t) = 0$,	means isolated boundary at the boundary point $x = 1$ (where no heat flux occurs).
$f = 20 - u$,	is the heat source, in this case a control system to force $u \to 20$.

Remark 7.1. *Observe that it is possible to generalize* (7.1.1) *to a* u *dependent source term* f*, e.g. as in the above example where* $f = 20 - u$*.*

7.1.1 Stability estimates

We shall derive a general stability estimate for the mixed (Dirichlet at one end point and Neumann in the other) initial boundary value problem above, prove a one-dimensional version of the *Poincare inequality* and finally derive also some stability estimates in the homogeneous ($f \equiv 0$) case. These are tools that we shall need in our finite element analysis.

Theorem 7.1. *The IBVP* (7.1.1) *satisfies the stability estimates*

$$||u(\cdot,t)|| \leq ||u_0|| + \int_0^t ||f(\cdot,s)||\, ds, \tag{7.1.2}$$

$$||u_x(\cdot,t)||^2 \leq ||u_0'||^2 + \int_0^t ||f(\cdot,s)||^2\, ds, \tag{7.1.3}$$

where u_0 *and* u_0' *are assumed to be* $L_2(I)$ *functions with* $I = (0,1)$*, further we have assumed that* $f \in L_1\Big([0,t], L_2(I)\Big) \cap L_2\Big([0,t], L_2(I)\Big)$*. Note further*

that, here $||\cdot||$ *is the time dependent* L_2 *norm:*

$$||w(\cdot,s)|| := ||w(\cdot,s||_{L_2(0,1)} = \Big(\int_\Omega ||w(x,s)||^2\,dx\Big)^{1/2}.$$

Proof. Multiplying the equation in (7.1.1) by u and integrating over $(0,1)$ yields

$$\int_0^1 \dot{u}u\,dx - \int_0^1 u''u\,dx = \int_0^1 fu\,dx. \tag{7.1.4}$$

By integration by parts in the second integral, we have

$$\frac{1}{2}\frac{d}{dt}\int_0^1 u^2\,dx + \int_0^1 (u')^2\,dx - u'(1,t)u(1,t) + u'(0,t)u(0,t) = \int_0^1 fu\,dx.$$

Then, using the boundary conditions and Cauchy-Schwarz' inequality we end up with

$$||u||\frac{d}{dt}||u|| + ||u'||^2 = \int_0^1 fu\,dx \le ||f||\,||u||. \tag{7.1.5}$$

Consequently $||u||\frac{d}{dt}||u|| \le ||f||\,||u||$, and thus

$$\frac{d}{dt}||u|| \le ||f||. \tag{7.1.6}$$

Integrating over time we get

$$||u(\cdot,t)|| - ||u(\cdot,0)|| \le \int_0^t ||f(\cdot,s)||\,ds, \tag{7.1.7}$$

which yields the first assertion (7.1.2) of the theorem. To prove (7.1.3) we multiply the differential equation by $\dot{u}$, integrate over $(0,1)$, and use integration by parts so that we have on the left hand side

$$\int_0^1 (\dot{u})^2\,dx - \int_0^1 u''\dot{u}\,dx = ||\dot{u}||^2 + \int_0^1 u'\dot{u}'\,dx - u'(1,t)\dot{u}(1,t) + u'(0,t)\dot{u}(0,t).$$

Hence, using the boundary data, where $u(0,t) = 0 \Longrightarrow \dot{u}(0,t) = 0$, and Cauchy-Schwarz' inequality

$$||\dot{u}||^2 + \frac{1}{2}\frac{d}{dt}||u'||^2 = \int_0^1 f\dot{u}\,dx \le ||f||\,||\dot{u}|| \le \frac{1}{2}\Big(||f||^2 + ||\dot{u}||^2\Big), \tag{7.1.8}$$

or

$$\frac{1}{2}||\dot{u}||^2 + \frac{1}{2}\frac{d}{dt}||u'||^2 \leq \frac{1}{2}||f||^2. \tag{7.1.9}$$

Therefore, evidently,

$$\frac{d}{dt}||u'||^2 \leq ||f||^2. \tag{7.1.10}$$

Finally, integrating over $(0,t)$ we get the second assertion of the theorem:

$$||u'(\cdot,t)||^2 - ||u'(\cdot,0)||^2 \leq \int_0^t ||f(\cdot,s)||^2\,ds, \tag{7.1.11}$$

and the proof is complete. □

To proceed we give a proof of a one-dimensional version of the *Poincare inequality*: one of the most important inequalities in PDE and analysis.

Theorem 7.2 (The Poincare inequality in $1-d$)**.** *Assume that u and u' are square integrable functions on an interval $[0,L]$. There exists a constant C_L, independent of u, but dependent on L, such that if $u(0)=0$, then*

$$\int_0^L u(x)^2\,dx \leq C_L \int_0^L u'(x)^2\,dx, \quad \textit{i.e.} \quad ||u|| \leq \sqrt{C_L}||u'||. \tag{7.1.12}$$

Proof. For $x \in [0,L]$ we may write

$$\begin{aligned} u(x) &= \int_0^x u'(y)\,dy \leq \int_0^x |u'(y)|\,dy = \int_0^x |u'(y)|\cdot 1\,dy \\ &\leq \Big(\int_0^L |u'(y)|^2\,dy\Big)^{1/2}\cdot\Big(\int_0^L 1^2 dy\Big)^{1/2} = \sqrt{L}\Big(\int_0^L |u'(y)|^2\,dy\Big)^{1/2}, \end{aligned}$$

where in the last step we used the Cauchy-Schwarz inequality. Thus

$$\int_0^L u(x)^2\,dx \leq \int_0^L L\Big(\int_0^L |u'(y)|^2\,dy\Big)\,dx = L^2\int_0^L |u'(y)|^2\,dy, \tag{7.1.13}$$

and hence

$$||u|| \leq L||u'||. \tag{7.1.14}$$

□

Remark 7.2. *The constant $C_L = L^2$ indicates that the Poincare inequality is valid for arbitrary bounded intervals, but not for unbounded intervals. If $u(0) \neq 0$ and, for simplicity $L = 1$, then by a similar argument as above we get the following version of the one-dimensional Poincare inequality:*

$$||u||^2_{L_2(0,1)} \leq 2\Big(u(0)^2 + ||u'||^2_{L_2(0,1)}\Big). \tag{7.1.15}$$

Theorem 7.3 (Stability of the homogeneous heat equation)**.** *The initial boundary value problem for the heat equation*

$$\begin{cases} \dot{u} - u'' = 0, & 0 < x < 1, \quad t > 0 \\ u(0,t) = u_x(1,t) = 0, & \qquad\quad t > 0 \\ u(x,0) = u_0(x), & 0 < x < 1, \end{cases} \tag{7.1.16}$$

satisfies the following stability estimates

$$a) \quad \frac{d}{dt}||u||^2 + 2||u'||^2 = 0, \qquad b) \quad ||u(\cdot,t)|| \leq e^{-t}||u_0||.$$

Proof. a) Multiply the equation by u and integrate over $x \in (0,1)$, to get

$$0 = \int_0^1 (\dot{u} - u'')u\,dx = \int_0^1 \dot{u}u\,dx + \int_0^1 (u')^2\,dx - u'(1,t)u(1,t) + u'(0,t)u(0,t),$$

where we used integration by parts. Using the boundary data we then have

$$\frac{1}{2}\frac{d}{dt}\int_0^1 u^2\,dx + \int_0^1 (u')^2\,dx = \frac{1}{2}\frac{d}{dt}||u||^2 + ||u'||^2 = 0.$$

This gives the proof of a). As for the proof of b), using a) and the Poincare inequality, with $L = 1$, i.e., $||u|| \leq ||u'||$ we get

$$\frac{d}{dt}||u||^2 + 2||u||^2 \leq 0. \tag{7.1.17}$$

Multiplying both sides of (7.1.17) by the integrating factor e^{2t} yields

$$\frac{d}{dt}\Big(||u||^2 e^{2t}\Big) = \Big(\frac{d}{dt}||u||^2 + 2||u||^2\Big)e^{2t} \leq 0. \tag{7.1.18}$$

We replace t by s and integrate in s over $(0,t)$, to obtain

$$\int_0^t \frac{d}{ds}\Big(||u||^2 e^{2s}\Big)\,ds = ||u(\cdot,t)||^2 e^{2t} - ||u(\cdot,0)||^2 \le 0. \tag{7.1.19}$$

This yields

$$||u(\cdot,t)||^2 \le e^{-2t}||u_0||^2 \Longrightarrow ||u(\cdot,t)|| \le e^{-t}||u_0||, \tag{7.1.20}$$

and completes the proof. □

Below, for the sake of generality, we shall denote the spatial domain by Ω and the boundary of Ω by $\partial\Omega$. Thus, the corresponding formulation for the homogeneous heat equation reads as: Given $u_0(x)$ find $u(x,t)$ satisfying

$$\begin{cases} \dot{u} - u'' = 0, & \text{for } x \in \Omega, & t > 0, \\ u(x,t) = 0, & \text{for } x \in \partial\Omega_1, & t > 0, \\ u_x(x,t) = 0, & \text{for } x \in \partial\Omega_2, & t > 0, \\ u(x,0) = u_0(x), & \text{for } x \in \Omega, & \end{cases} \tag{7.1.21}$$

where $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$.

Theorem 7.4 (An energy estimate)**.** *For any small positive constant ε, the solution of the homogeneous heat equation* (7.1.21) *satisfies the estimate*

$$\int_\varepsilon^t \|\dot{u}\|(s)ds \le \frac{1}{2}\sqrt{\ln\frac{t}{\varepsilon}}\|u_0\|. \tag{7.1.22}$$

Proof. We multiply the differential equation: $\dot{u} - u'' = 0$, by $-tu''$, and integrate over Ω to obtain

$$-t\int_\Omega \dot{u}u''\,dx + t\int_\Omega (-u'')^2\,dx = 0. \tag{7.1.23}$$

Integrating by parts and using the boundary data (note that $u = 0$ on $\partial\Omega_1$ implies that $\dot{u} = 0$ on $\partial\Omega_1$) we get

$$\int_\Omega \dot{u}u''\,dx = -\int_\Omega \dot{u}'\cdot u'\,dx = -\frac{1}{2}\frac{d}{dt}\|u'\|^2, \tag{7.1.24}$$

so that (11.1.12) may be rewritten as

$$t\frac{1}{2}\frac{d}{dt}\|u'\|^2 + t\|u''\|^2 = 0. \tag{7.1.25}$$

Using the identity $t\frac{d}{dt}\|u'\|^2 = \frac{d}{dt}(t\|u'\|^2) - \|u'\|^2$ we end up with

$$\frac{d}{dt}(t\|u'\|^2) + 2t\|u''\|^2 = \|u'\|^2. \tag{7.1.26}$$

Now relabeling t as s and integrating in s over $(0,t)$ we get

$$\int_0^t \frac{d}{ds}(s\|u'\|^2(s))\,ds + 2\int_0^t s\|u''\|^2(s)\,ds = \int_0^t \|u'\|^2(s)\,ds \le \frac{1}{2}\|u_0\|^2,$$

where in the last inequality we just integrate the stability estimate (a) in the previous theorem over the interval $(0,t)$. Consequently,

$$t\|u'\|^2(t) + 2\int_0^t s\|u''\|^2(s)\,ds \le \frac{1}{2}\|u_0\|^2. \tag{7.1.27}$$

In particular, we have:

$$(I)\ \ \|u'\|(t) \le \frac{1}{\sqrt{2t}}\|u_0\|, \qquad (II)\ \ \Big(\int_0^t s\|u''\|^2(s)\,ds\Big)^{1/2} \le \frac{1}{2}\|u_0\|. \tag{7.1.28}$$

Now using the differential equation $\dot{u} = u''$, and integrating over (ε, t) (same relabeling as above), we obtain

$$\begin{aligned}\int_\varepsilon^t \|\dot{u}\|(s)ds = \int_\varepsilon^t \|u''\|(s)ds = \int_\varepsilon^t 1\cdot\|u''\|(s)ds = \int_\varepsilon^t \frac{1}{\sqrt{s}}\cdot\sqrt{s}\|u''\|(s)ds \\ \le \Big(\int_\varepsilon^t s^{-1}\,ds\Big)^{1/2}\cdot\Big(\int_\varepsilon^t s\|u''\|^2(s)\,ds\Big)^{1/2} \le \frac{1}{2}\sqrt{\ln\Big(\frac{t}{\varepsilon}\Big)}\|u_0\|,\end{aligned}$$

where in the last inequality we have used the estimate (II) in (7.1.28). □

Problem 7.1. *Prove that*

$$\|u''\|(t) \le \frac{1}{\sqrt{2}\,t}\|u_0\|. \tag{7.1.29}$$

Hint: Multiply $\dot{u} - u''$ by $t^2\,(u'')^2$, and note that $u'' = \dot{u} = 0$ on $\partial\Omega$, or alternatively: differentiate $\dot{u} - u'' = 0$ with respect to t and multiply the resulting equation by $t^2\,\dot{u}$.

7.1.2 FEM for the heat equation

We consider the one-dimensional heat equation with Dirichlet boundary conditions

$$\begin{cases} \dot{u} - u'' = f, & 0 < x < 1, \quad t > 0, \\ u(0,t) = u(1,t) = 0, & t > 0, \\ u(x,0) = u_0(x), & 0 < x < 1. \end{cases} \tag{7.1.30}$$

The *Variational formulation* for this problem reads as follows: For every time interval $I_n = (t_{n-1}, t_n]$, find $u(x,t),\ x \in (0,1),\ t \in I_n$, such that

$$\int_{I_n}\int_0^1 (\dot{u}v + u'v')dxdt = \int_{I_n}\int_0^1 fvdxdt, \quad \forall v: \ v(0,t) = v(1,t) = 0. \tag{VF}$$

A piecewise linear Galerkin finite element method (FEM): $cG(1) - cG(1)$ is then formulated as: for each time interval $I_n := (t_{n-1}, t_n]$, with $t_n - t_{n-1} = k_n$, let

$$U(x,t) = U_{n-1}(x)\Psi_{n-1}(t) + U_n(x)\Psi_n(t), \tag{7.1.31}$$

where

$$\Psi_n(t) = \frac{t - t_{n-1}}{k_n}, \qquad \Psi_{n-1}(t) = \frac{t_n - t}{k_n}, \tag{7.1.32}$$

and

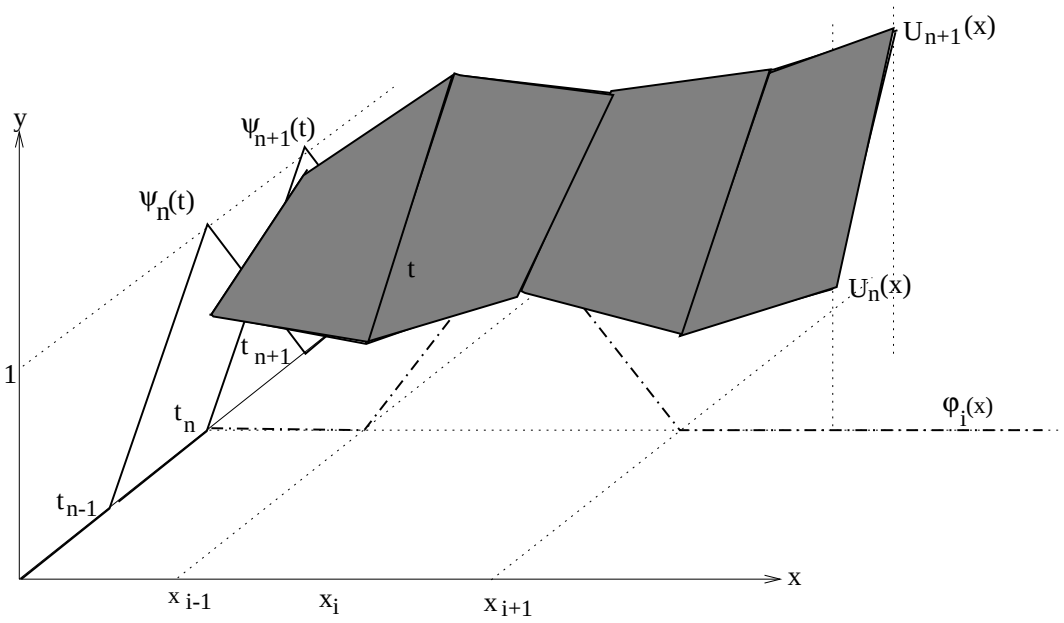
$$U_{\tilde{n}}(x) = U_{\tilde{n},1}\varphi_1(x) + U_{\tilde{n},2}\varphi_2(x) + \ldots + U_{\tilde{n},m}\varphi_m(x), \quad \tilde{n} = n-1,\ n \tag{7.1.33}$$

with φ_j being the usual piecewise linear, continuous finite element basis functions (hat-functions) corresponding to a partition of $\Omega = (0,1)$, with $0 = x_0 < \cdots < x_\ell < x_{\ell+1} < \cdots < x_{m+1} = 1$, and $\varphi_j(x_i) := \delta_{ij}$. Now the Galerkin method (FEM) is to determine the unknown coefficients $U_{n,\ell}$ in the above representation for U (i.e. for U being a piecewise linear, continuous function, in both the space and time variables) that satisfies the following discrete variational formulation: Find $U(x,t)$ given by (7.1.31) such that

$$\int_{I_n}\int_0^1 (\dot{U}\varphi_i + U'\varphi_i')\, dxdt = \int_{I_n}\int_0^1 f\varphi_i\, dxdt, \qquad i = 1, 2, \ldots, m. \tag{7.1.34}$$

Note that, on $I_n = (t_{n-1}, t_n]$ and with $U_n(x) := U(x,t_n)$ and $U_{n-1}(x) := U(x,t_{n-1})$,

$$\dot{U}(x,t) = U_{n-1}(x)\dot{\Psi}_{n-1}(t) + U_n(x)\dot{\Psi}_n(t) = \frac{U_n - U_{n-1}}{k_n}. \tag{7.1.35}$$



Further differentiating (7.1.31) with respect to x we have

$$U'(x,t) = U'_{n-1}(x)\Psi_{n-1}(t) + U'_n(x)\Psi_n(t). \tag{7.1.36}$$

Inserting (7.1.35) and (7.1.36) into (7.1.34) we get using the identities, $\int_{I_n} dt = k_n$ and $\int_{I_n} \Psi_n dt = \int_{I_n} \Psi_{n-1} dt = k_n/2$ that,

$$\begin{aligned} \underbrace{\int_0^1 U_n \varphi_i dx}_{M\cdot U_n} - \underbrace{\int_0^1 U_{n-1}\varphi_i\, dx}_{M\cdot U_{n-1}} + \underbrace{\int_{I_n} \Psi_{n-1}\, dt}_{k_n/2} \underbrace{\int_0^1 U'_{n-1}\varphi'_i\, dx}_{S\cdot U_{n-1}} \\ + \underbrace{\int_{I_n} \Psi_n\, dt}_{k_n/2} \underbrace{\int_0^1 U'_n \varphi'_i\, dx}_{S\cdot U_n} = \underbrace{\int_{I_n}\int_0^1 f\varphi_i\, dxdt}_{F_n}. \end{aligned} \tag{7.1.37}$$

This can be written in a compact form as the *Crank-Nicolson system*

$$\Big(M + \frac{k_n}{2}S\Big)U_n = \Big(M - \frac{k_n}{2}S\Big)U_{n-1} + F_n, \tag{CNS}$$

with the solution U_n given by the data U_{n-1} and F, viz

$$U_n = \underbrace{\Big(M + \frac{k_n}{2}S\Big)^{-1}}_{B^{-1}}\underbrace{\Big(M - \frac{k_n}{2}S\Big)}_{A}U_{n-1} + \underbrace{\Big(M + \frac{k_n}{2}S\Big)^{-1}}_{B^{-1}}F_n, \qquad (7.1.38)$$

where M and S (computed below) are known as the *mass-matrix* and *stiffness-matrix*, respectively, and

$$U_n = \begin{bmatrix} U_{n,1} \\ U_{n,2} \\ \dots \\ U_{n,m} \end{bmatrix}, \qquad F = \begin{bmatrix} F_{n,1} \\ F_{n,2} \\ \dots \\ F_{n,m} \end{bmatrix}, \qquad F_{n,i} = \int_{I_n}\int_0^1 f\varphi_i \, dx\, dt. \quad (7.1.39)$$

Thus, given the source term f we can determine the vector F_n and then, for each $n = 1, \dots N$, given the vector U_{n-1} (the initial value is given by $U_{0,j} := u_0(x_j)$) we may use the CNS to compute $U_{n,\ell}$, $\ell = 1, 2, \dots m$ (m nodal values of U at the x_j : s, and at the time level t_n).

Problem 7.2. *Derive a corresponding system of equations, as above, for $cG(1) - dG(0)$: with the discontinuous Galerkin approximation ($dG(0)$) in time with piecewise constants.*

We now return to the computation of the matrix entries for M and S, for a uniform partition (all subintervals are of the same length) of the interval $I = (0, 1)$. Note that differentiating (7.1.33) with respect to x, yields

$$U_n'(x) = U_{n,1}\varphi_1'(x) + U_{n,2}\varphi_2'(x) + \ldots + U_{n,m}\varphi_m'(x). \qquad (7.1.40)$$

Hence, for $i = 1, \dots, m$, the rows in the system of equations are given by

$$\int_0^1 U_n'\varphi_i' = \Big(\int_0^1 \varphi_i'\varphi_1'\Big)U_{n,1} + \Big(\int_0^1 \varphi_i'\varphi_2'\Big)U_{n,2} + \ldots + \Big(\int_0^1 \varphi_i'\varphi_m'\Big)U_{n,m},$$

which can be written in matrix form as

$$SU_n = \begin{bmatrix} \int_0^1 \varphi_1'\varphi_1' & \int_0^1 \varphi_1'\varphi_2' & \dots & \int_0^1 \varphi_1'\varphi_m' \\ \int_0^1 \varphi_2'\varphi_1' & \int_0^1 \varphi_2'\varphi_2' & \dots & \int_0^1 \varphi_2'\varphi_m' \\ \dots & \dots & \dots & \dots \\ \int_0^1 \varphi_m'\varphi_1' & \int_0^1 \varphi_m'\varphi_2' & \dots & \int_0^1 \varphi_m'\varphi_m' \end{bmatrix} \begin{bmatrix} U_{n,1} \\ U_{n,2} \\ \dots \\ U_{n,m} \end{bmatrix}. \qquad (7.1.41)$$

Thus, S is just the stiffness matrix $\mathbf{A}_{unif}$ computed in Chapter 2:

$$S = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & -1 & 2 & -1 \\ 0 & \dots & \dots & \dots & -1 & 2 \end{bmatrix}. \tag{7.1.42}$$

A non-uniform partition yields a matrix of the form $\mathbf{A}$ in Chapter 2.
Similarly, recalling the notation for the mass matrix M introduced in (7.1.37), we have that for $i = 1, \dots, m$

$$[MU_n]_i = \int_0^1 U_n \varphi_i. \tag{7.1.43}$$

Hence, to compute the mass matrix M one should drop all derivatives from the general form of the matrix for S given by (7.1.41). In other words unlike the form $[SU_n]_i = \int_0^1 U_n' \varphi_i'$, MU_n does not involve any derivatives, neither in U_n nor in φ_i. Consequently

$$M = \begin{bmatrix} \int_0^1 \varphi_1\varphi_1 & \int_0^1 \varphi_1\varphi_2 & \dots & \int_0^1 \varphi_1\varphi_m \\ \int_0^1 \varphi_2\varphi_1 & \int_0^1 \varphi_2\varphi_2 & \dots & \int_0^1 \varphi_2\varphi_m \\ \dots & \dots & \dots & \dots \\ \int_0^1 \varphi_m\varphi_1 & \int_0^1 \varphi_m\varphi_2 & \dots & \int_0^1 \varphi_m\varphi_m \end{bmatrix}. \tag{7.1.44}$$

For a uniform partition, we have computed this mass matrix in Chapter 5:

$$M = h\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 & \dots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & \dots & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \dots & \dots & \dots & \frac{1}{6} & \frac{2}{3} \end{bmatrix} = \frac{h}{6}\begin{bmatrix} 4 & 1 & 0 & 0 & \dots & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 & 4 & 1 \\ 0 & \dots & \dots & \dots & 1 & 4 \end{bmatrix}.$$

7.1.3 Error analysis

In this section we outline a general framework for the error estimation procedure (this procedure is applicable for higher space dimensions as well). We denote the spatial domain by Ω with boundary $\partial\Omega$. In the one dimensional case treated here this means that Ω is an arbitrary interval $\Omega := [a, b]$ with $\partial\Omega = \{a, b\}$. For each $n = 1, 2, \ldots, N$, we consider the slab $S_n := \Omega \times I_n$, with $I_n = (t_{n-1}, t_n]$, and rewrite the variational formulation for the equation (7.1.30) as: to find the exact solution $u(x,t) \in \mathcal{H}_0^1 := C([0,T]; H_0^1([a,b]))$ such that for each $n = 1, 2, \ldots, N$,

$$\int_{I_n}\int_\Omega (\dot{u}v + u'v')dxdt = \int_{I_n}\int_\Omega fv\,dxdt, \quad \forall v \in C(I_n; H_0^1([a,b])), \quad (7.1.45)$$

with $H_0^1([a,b])$ being the usual L_2-based Sobolev space over $[a,b]$ (consisting of square integrable functions vanishing at $x = a$, $x = b$, and having square integrable derivatives). We may formulate a corresponding finite element method for the equation (7.1.30) as the continuous, piecewise linear space-time approximation in (7.1.31)-(7.1.34). We may alternatively construct the so called $cG(1)dG(q)$ method: a finite element method based on approximation using continuous piecewise linear functions in space and discontinuous piecewise polynomials of degree q in time. To this approach we let $S_n = \Omega \times I_n$, and define the trial space as

$$W_k^{(q)} = \{v(x,t) : v|_{S_n} \in W_{k,n}^{(q)},\ n = 1, \ldots, N\}, \quad (7.1.46)$$

where

$$W_{k,n}^{(q)} = \{v(x,t) : v(x,t) = \sum_{j=0}^{q} t^j \varphi_j(x) :\ \varphi_j \in V_n^0,\ (x,t) \in S_n\}.$$

Here $V_n^0 = V_{h_n}^0$ is the space of continuous piecewise linear functions on a subdivision $\mathcal{T}_n$ of Ω with mesh function h_n, vanishing on the boundary. The functions in $W_k^{(q)}$ may be discontinuous in time at the discrete time levels t_n. To account for this, we use the same notation as in the previous chapter:

$$w_n^\pm = \lim_{s \to 0^\pm} w(t_n + s), \qquad [w_n] := w_n^+ - w_n^-. \quad (7.1.47)$$

Now we may formulate the $cG(1)dG(q)$ method as follows: Let $U_0^- := u_0$, and find $U \in W_k^{(q)}$ such that for $n = 1, \ldots, N$,

$$\int_{I_n}\int_\Omega (\dot{U}v + U'v')\,dx\,dt + \int_\Omega [U_{n-1}]v_{n-1}^+\,dx = \int_{I_n}\int_\Omega fv\,dx\,dt \quad \text{for all } v \in W_{k,n}^{(q)}. \tag{7.1.48}$$

Subtracting (7.1.48) from (7.1.45) we obtain the *time-discontinuous Galerkin orthogonality* relation for the error

$$\int_{I_n}\int_\Omega (\dot{e}v + e'v')\,dxdt + \int_\Omega [e_{n-1}]v_{n-1}^+\,dx = 0, \quad \text{for all } v \in W_{kh}^{(q)}. \tag{7.1.49}$$

Obviously in the time-continuous case the jump term in (7.1.48) disappears and we end up with $cG(1)cG(q)$ method, viz:

$$\int_{I_n}\int_\Omega (\dot{U}v + U'v')\,dx\,dt = \int_{I_n}\int_\Omega fv\,dx\,dt \quad \text{for all } v \in W_{k,n}^{(q)}. \tag{7.1.50}$$

Now, subtracting (7.1.50) from (7.1.45) we obtain the *Galerkin orthogonality* relation for the error $e = u - U$ as

$$\int_{I_n}\int_\Omega (\dot{e}v + e'v')\,dxdt = 0, \quad \text{for all } v \in W_{kn}^{(q)}. \tag{7.1.51}$$

A priori and a posteriori error estimates for the heat equation are obtained combining the results in Chapters 5 and 6. To become familiar with some standard techniques, below we shall demonstrate the a posteriori approach.

Theorem 7.5 (cG(1)cG(1) a posteriori error estimate)**.** *Assume:*

- $h_n = h,\ \forall n$
- $u_0 \in V_{h_1}^0 = V_h^0$ *(so that* $e(0) = 0$*).*

Let $\varepsilon = k_N$ *be small, and let* $0 < t < T$*. Let* u *and* U *be the exact and approximate solutions for the heat equation* (7.1.30)*, respectively, i.e.* U *is the finite element approximation satisfying in either one of the equations* (7.1.34) *(cG(1)cG(1) or* (7.1.50) *(cG(1)cG(q). Then the error* $e = u - U$ *is bounded as*

$$\|e(T)\| \le \left(2\sqrt{\ln\frac{T}{\varepsilon}}\right) \max_{[0,T]} \|(k + h^2)r(U)\|, \tag{7.1.52}$$

where $r(U)$ *is the residual and* k *and* h *are temporal and spatial mesh function, respectively.*

Proof. Following the general framework for a posteriori error estimates, we let $\varphi(x,t)$ be the solution of the *dual problem*

$$\begin{cases} -\dot{\varphi} - \varphi'' = 0, & \text{in } \Omega \times (0,T), \\ \varphi = 0, & \text{on } \partial\Omega \times (0,T), \\ \varphi = e, & \text{in } \Omega \qquad \text{for } t = T, \end{cases} \tag{7.1.53}$$

where $e = e(T) = e(\cdot,T) = u(\cdot,T) - U(\cdot,T)$, $T = t_N$. By a change of variables, letting $w(x,s) = \varphi(x,T-s)$, $(s>0)$; the *backward dual problem* (7.1.53) is rewritten in the *forward form*:

$$\begin{cases} \dot{w} - w'' = 0, & \text{in } \Omega \times (0,T), \\ w = 0, & \text{on } \partial\Omega \times (o,t), \\ w = e, & \text{in } \Omega \qquad \text{for } s = 0. \end{cases} \tag{7.1.54}$$

For this problem we have shown in the energy estimate theorem 7.4 that

$$\int_\varepsilon^T \|\dot{w}\|\, ds \le \frac{1}{2}\sqrt{\ln\frac{T}{\varepsilon}}\, \|e\|, \tag{7.1.55}$$

and consequently, if $s = T - t$, then $ds = -dt$ and $\dot{w}(x,s) = -\dot{\varphi}(x,T-s)$. Hence, since $-\varphi'' = \dot{\varphi}$, we have for φ, that

$$\int_0^{T-\varepsilon} \|\dot{\varphi}\|\, dt \le \frac{1}{2}\sqrt{\ln\frac{T}{\varepsilon}}\, \|e\|, \quad \text{and} \quad \int_0^{T-\varepsilon} \|\varphi''\|\, dt \le \frac{1}{2}\sqrt{\ln\frac{T}{\varepsilon}}\|e\|. \tag{7.1.56}$$

Now assume $h_n = h$, $n = 1,2,\ldots$, and let $u_0 \in V_{h_1}^0 = V_h^0$, then, since $(-\dot{\varphi} - \varphi'') = 0$, integration by parts in t and x yields

$$\begin{aligned} \|e(T)\|^2 &= \int_\Omega e(T)\cdot e(T)\, dx + \int_0^T \int_\Omega e(-\dot{\varphi} - \varphi'')\, dxdt \\ &= \int_\Omega e(T)\cdot e(T)\, dx - \int_\Omega e(T)\cdot\varphi(T)\, dx + \int_\Omega \underbrace{e(0)\cdot\varphi(0)}_{=0;(u_0\in V_{h_1}^0)}\, dx \\ &\quad + \int_0^T \int_\Omega (\dot{e}\varphi + e'\varphi')\, dxdt. \end{aligned}$$

Using Galerkin orthogonality and integration by parts in x, we get for $v \in W_k^{(0)}$ (i.e. piecewise constant in time and continuous, piecewise linear in space),

$$\|e(T)\|^2 = \int_0^T \int_\Omega \dot{e}(\varphi - v) + e'(\varphi - v)' \, dxdt. \tag{7.1.57}$$

But we have

$$\begin{aligned}
\int_\Omega \dot{e}(\varphi - v) + e'(\varphi - v)' \, dx &= \int_\Omega (\dot{u} - \dot{U})(\varphi - v) \, dx \\
&+ \sum_{i=1}^{M+1} \int_{x_{i-1}}^{x_i} (u' - U')(\varphi - v)' \, dx \\
&= \int_\Omega (\dot{u} - \dot{U})(\varphi - v) \, dx + \sum_{i=1}^{M+1} [(u' - U')(\varphi - v)]_{x_{i-1}}^{x_i} \\
&- \sum_{i=1}^{M+1} \int_{x_{i-1}}^{x_i} (u'' - U'')(\varphi - v) \, dx \\
&= \int_\Omega (f - \dot{U})(\varphi - v) \, dx + \sum_{i=1}^{M} [U_i'](\varphi - v)(x_i).
\end{aligned}$$

Thus

$$\begin{aligned}
\|e(T)\|^2 = \int_0^T \int_\Omega (f - \dot{U})(\varphi - v) \, dx + \int_0^T \sum_{i=1}^{M} [U_i'](\varphi - v)(x_i) \\
\int_0^T \int_\Omega r_1(U)(\varphi - v) \; dxdt + \int_0^T \sum_{i=1}^{M} [U_i'](\varphi - v)(x_i).
\end{aligned} \tag{7.1.58}$$

where we have used $\dot{e} - e'' = \dot{u} - u'' - \dot{U} + U'' = f - \dot{U} + U'' = f - \dot{U} := r_1(U)$ which is the residual. Next, with mesh functions $h = h(x)$ and $k = k(t)$ in x and t, respectively we may derive (using the fact that v is piecewise constant in time and continuous, piecewise linear in space) an interpolation estimate of the form:

$$||(k + h^2)^{-1}(\varphi - v)||_{L_2} \leq C_i(||\dot{\varphi}||_{L_2} + ||\varphi''||_{L_2}). \tag{7.1.59}$$

To estimate the first term in (7.1.58) (see Eriksson et al [12] Chap 16) we

have using the estimates (7.1.56) that

$$\begin{aligned}\int_0^T \int_\Omega r_1(U)(\varphi - v)\, dxdt \leq & C_i \int_0^{T-k_N} \|(k+h^2)r_1(U)\|(\|\dot{\varphi}\| + \|\varphi''\|)\, dt \\ & + \int_{T-k_N}^{T} \|r_1(U)\|\|\varphi - v\|\, dt \\ \leq & C_i \max_{[0,T-k_N]} \|(k+h^2)r_1(U)\| \cdot \sqrt{\ln \frac{T}{k_N}} \|e(T)\| \\ & + \max_{[T-k_N,T]} \Big(\|r_1(U)\| \cdot \|\varphi - v\| \Big) \cdot k_N,\end{aligned}$$

where the last term can be estimated using the stability estimate (7.1.2) for the dual problem with $f = 0$ as

$$\Big(\|r_1(U)\| \cdot \|\varphi - v\| \Big) \cdot k_N \leq C \|k_N r_1(U)\| \|\varphi\| \leq C \|k_N r_1(U)\| \|e(T)\|. \quad (7.1.60)$$

Summing up we get

$$\int_0^T \int_\Omega r_1(U)(\varphi - v)\, dxdt \leq C_i \max_{[0,T]} \|(k+h^2)r_1(U)\| \cdot \sqrt{\ln \frac{T}{k_N}} \cdot \|e(T)\|. \quad (7.1.61)$$

The second term in (7.1.58) can be similarly estimated as

$$\int_0^T \sum_{i=1}^{M} [U_i'](\varphi - v)(x_i) \leq C_i \max_{[0,T]} \|(k+h^2)r_2(U)\| \cdot \sqrt{\ln \frac{T}{k_N}} \cdot \|e(T)\|, \quad (7.1.62)$$

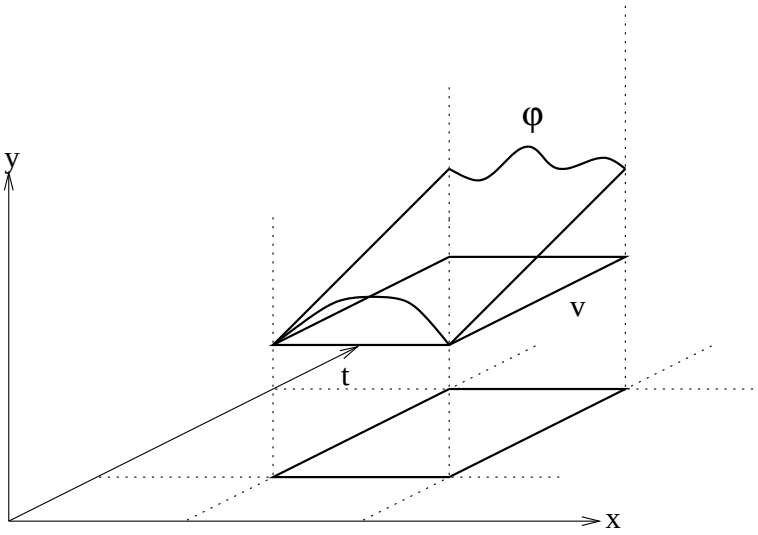
where $r_2(U) := h_i^{-1}|[U_i']|$, for $x \in I_i$. This gives our final estimate and the proof is complete.

$$\|e(T)\| \leq C \Big(\sqrt{\ln \frac{T}{k_N}} \Big) \max_{[0,T]} \|(k+h^2)r(U)\|, \quad (7.1.63)$$

where $r(U) = r_1(U) + r_2(U)$. □

• **Adaptivity Algorithm.** Roughly speaking, adaptivity procedure in a posteriori error estimates can be outlined as follows: Consider, as an example, the Poisson equation

$$-u'' = f \quad \text{in } \Omega, \qquad u = 0, \quad \text{on } \partial\Omega \quad (7.1.64)$$



where for the error $u - U$, we have derived the a posteriori error estimate

$$\|e'\| \leq C\|h\, r(U)\|, \tag{7.1.65}$$

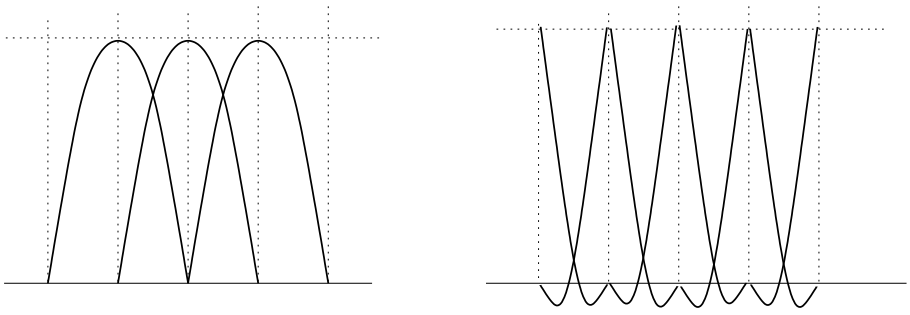
with $r(U) = |f| + \max_{I_k} |[u']|$ and $[\cdot]$ denoting the jump (over the endpoints of a partition interval I_k). Then, the adaptivity is stated as the following, 3-steps, algorithm:

(1) Choose an arbitrary mesh-size $h = h(x)$ and a tolerance $\mathrm{Tol} > 0$.

(2) Given h, compute the corresponding U (also denoted u_h).

(3) If $C\|hr(U)\| \leq \mathrm{Tol}$, then accept U. Otherwise choose a new (refined) $h = h(x)$ and return to the step (2) above. □

• **Higher order elements** $cG(2)$. Piecewise polynomials of degree 2 is determined by the values of the approximate solution at the, e.g. mid-point and end-points of each subinterval. The construction is through using the Lagrange interpolation basis $\lambda_0(x)$, $\lambda_1(x)$, and $\lambda_2(x)$, introduced in the interpolation chapter; as in the graphs below:



Example 7.2 (Error estimates in a simple case for cG(2)). *For the Poisson equation* $-u'' = f,\ \ 0 < x < 1$, *associated with a Dirichlet (or Neumann) boundary condition we have the following* $cG(2)$ *error estimates*

$$\|(u-U)'\| \leq C\|h^2 D^3 u\|. \tag{7.1.66}$$

$$\|u-U\| \leq C \max\Big(h\|h^2 D^3 u\|\Big). \tag{7.1.67}$$

$$\|u-U\| \leq C\|h^2 r(U)\|, \qquad \textit{where} \quad |r(U)| \leq Ch. \tag{7.1.68}$$

The proof of the estimates (7.1.66)-(7.1.68) are rather involved (see Eriksson et al [12] for the details) and therefore are omitted. These estimates can be extended to the space-time discretization of the heat equation.

Example 7.3 (The equation of an elastic beam).

$$\begin{cases} (au'')'' = f, & \Omega = (0,1), \\ u(0) = 0, & u'(0) = 0, \qquad (\textit{Dirichlet}) \\ u''(1) = 0, & (au'')'(1) = 0, \quad (\textit{Neumann}) \end{cases} \tag{7.1.69}$$

where a *is the bending stiffness,* au' *is the moment,* f *is the load function, and* $u = u(x)$ *is the vertical deflection.*
A variational formulation for this equation can be written as

$$\int_0^1 au''v''dx = \int_0^1 fvdx, \qquad \forall v, \qquad \textit{such that} \quad (0) = v'(0) = 0. \tag{7.1.70}$$

Here, considering piecewise linear finite element functions is inadequate.

7.1.4 Exercises

Problem 7.3. *Work out the details with piecewise cubic polynomials having continuous first derivatives: i.e., two degrees of freedom on each node.*
Hint: A cubic polynomial in (a, b) *is uniquely determined by* $\varphi(a), \varphi'(a), \varphi(b)$ *and* $\varphi'(b)$.

Problem 7.4. *Let* $\|\cdot\|$ *denote the* $L_2(0,1)$*-norm. Consider the problem*

$$\begin{cases} -u'' = f, & 0 < x < 1, \\ u'(0) = v_0, & u(1) = 0. \end{cases}$$

a) Show that $|u(0)| \leq \|u'\|$ *and* $\|u\| \leq \|u'\|$.

b) Use a) to show that $\|u'\| \leq \|f\| + |v_0|$.

Problem 7.5. *Assume that* $u = u(x)$ *satisfies*

$$\int_0^1 u'v'dx = \int_0^1 fv\,dx, \quad \text{for all } v(x) \text{ such that } v(0) = 0. \tag{7.1.71}$$

Show that $-u'' = f$ *for* $0 < x < 1$ *and* $u'(1) = 0$.
Hint: See previous chapters.

Problem 7.6 (Generalized Poincare)**.** *Show that for a continuously differentiable function* v *defined on* $(0,1)$ *we have that*

$$||v||^2 \leq v(0)^2 + v(1)^2 + ||v'||^2.$$

Hint: Use partial integration for $\int_0^{1/2} v(x)^2\,dx$ *and* $\int_{1/2}^1 v(x)^2\,dx$ *and note that* $(x - 1/2)$ *has the derivative 1.*

Problem 7.7. *Let* $\|\cdot\|$ *denote the* $L_2(0,1)$*-norm. Consider the following heat equation*

$$\begin{cases} \dot{u} - u'' = 0, & 0 < x < 1, & t > 0, \\ u(0,t) = u_x(1,t) = 0, & & t > 0, \\ u(x,0) = u_0(x), & 0 < x < 1. & \end{cases}$$

a) Show that the norms: $||u(\cdot,t)||$ *and* $||u_x(\cdot,t)||$ *are non-increasing in time.* $||u|| = \left(\int_0^1 u(x)^2 \, dx \right)^{1/2}$.

b) Show that $||u_x(\cdot,t)|| \to 0$, *as* $t \to \infty$.

c) Give a physical interpretation for a) and b).

Problem 7.8. *Consider the inhomogeneous problem 7.8:*

$$\begin{cases} \dot{u} - \varepsilon u'' = f, & 0 < x < 1, \quad t > 0, \\ u(0,t) = u_x(1,t) = 0, & t > 0, \\ u(x,0) = u_0(x), & 0 < x < 1. \end{cases}$$

where $f = f(x,t)$.
a) Show the stability estimate

$$||u(\cdot,t)|| \leq \int_0^t ||f(\cdot,s)|| \, ds.$$

b) Show that for the corresponding stationary ($\dot{u} \equiv 0$*) problem we have*

$$||u'|| \leq \frac{1}{\varepsilon} ||f||.$$

Problem 7.9. *Give an a priori error estimate for the following problem:*

$$(au_{xx})_{xx} = f, \quad 0 < x < 1, \qquad u(0) = u'(0) = u(1) = u'(1) = 0,$$

where $a(x) > 0$ *on the interval* $I = (0,1)$.

7.2 The wave equation in 1d

The theoretical study of the wave equation, being of hyperbolic type, has some basic differences compared to that of the heat equation which is a parabolic equation. Some important aspects in this regard are given in part 2 of these Lecture notes concerning higher spatial dimensions. Nevertheless, in our study here, the structure of the finite element methods for the wave equation is, mainly, the same procedure as for the heat equation outlined in the previous section. To proceed we start with an example of the homogeneous wave equation, by considering the initial-boundary value problem

$$\begin{cases} \ddot{u} - u'' = 0, & 0 < x < 1 \quad t > 0 \quad (DE) \\ u(0,t) = 0, \quad u(1,t) = 0 & t > 0 \quad (BC) \\ u(x,0) = u_0(x), \quad \dot{u}(x,0) = v_0(x), & 0 < x < 1. \quad (IC) \end{cases} \tag{7.2.1}$$

The most important property of the wave equation is the *conservation of energy*:

Theorem 7.6. *For the wave equation* (7.2.1) *we have that*

$$\frac{1}{2}||\dot{u}||^2 + \frac{1}{2}||u'||^2 = \frac{1}{2}||v_0||^2 + \frac{1}{2}||u_0'||^2 = \textit{Constant}, \tag{7.2.2}$$

where

$$||w||^2 = ||w(\cdot,t)||^2 = \int_0^1 |w(x,t)|^2\,dx. \tag{7.2.3}$$

Proof. We multiply the equation by $\dot{u}$ and integrate over $I = (0,1)$ to get

$$\int_0^1 \ddot{u}\,\dot{u}dx - \int_0^1 u''\,\dot{u}\,dx = 0. \tag{7.2.4}$$

Using integration by parts and the boundary data we obtain

$$\begin{aligned} \int_0^1 \frac{1}{2}\frac{\partial}{\partial t}\Big(\dot{u}\Big)^2\,dx + \int_0^1 u'\,(\dot{u})'\,dx - \Big[u'(x,t)\dot{u}(x,t)\Big]_0^1 \\ = \int_0^1 \frac{1}{2}\frac{\partial}{\partial t}\Big(\dot{u}\Big)^2\,dx + \int_0^1 \frac{1}{2}\partial\partial t\Big(u'\Big)^2\,dx \\ = \frac{1}{2}\frac{d}{dt}\Big(||\dot{u}||^2 + ||u'||^2\Big) = 0. \end{aligned} \tag{7.2.5}$$

Thus, we have that the quantity

$$\frac{1}{2}||\dot{u}||^2 + \frac{1}{2}||u'||^2 = \text{Constant, independent of } t. \tag{7.2.6}$$

Therefore the total energy is conserved. We recall that $\frac{1}{2}||\dot{u}||^2$ is the kinetic energy, and $\frac{1}{2}||u'||^2$ is the potential (elastic) energy.

□

Problem 7.10. *Show that* $\|(\dot{u})'\|^2 + \|u''\|^2 =$ *constant, independent of* t. *Hint: Differentiate the equation with respect to* x *and multiply by* $\dot{u}, \ldots$. *Alternatively: Multiply (DE):* $\ddot{u} - u'' = 0$, *by* $-(\dot{u})''$ *and integrate over* I.

Problem 7.11. *Derive a total conservation of energy relation using the Robin type boundary condition:* $\dfrac{\partial u}{\partial n} + u = 0$.

7.2.1 Wave equation as a system of hyperbolic PDEs

We rewrite the wave equation as a system of hyperbolic differential equations. To this approach, we consider solving

$$\begin{cases} \ddot{u} - u'' = 0, & 0 < x < 1, & t > 0, \\ u(0,t) = 0, & u'(1,t) = g(t), & t > 0, \\ u(x,0) = u_0(x), & \dot{u}(x,0) = v_0(x), & 0 < x < 1, \end{cases} \tag{7.2.7}$$

where we let $\dot{u} = v$, and reformulate the problem as:

$$\begin{cases} \dot{u} - v = 0, & \text{(Convection)} \\ \dot{v} - u'' = 0, & \text{(Diffusion).} \end{cases} \tag{7.2.8}$$

We may now set $w = (u, v)^t$ and rewrite the system (7.2.8) as $\dot{w} + Aw = 0$:

$$\dot{w} + Aw = \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ -\frac{\partial^2}{\partial x^2} & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{7.2.9}$$

In other words, the matrix differential operator is given by

$$A = \begin{pmatrix} 0 & -1 \\ -\frac{\partial^2}{\partial x^2} & 0 \end{pmatrix}.$$

Note that, in scalar form, this equation was studied in full detail in Chapter 6. Therefore, in this setting, the study of the wave equation is strongly connected with that of *systems of initial value problems*.

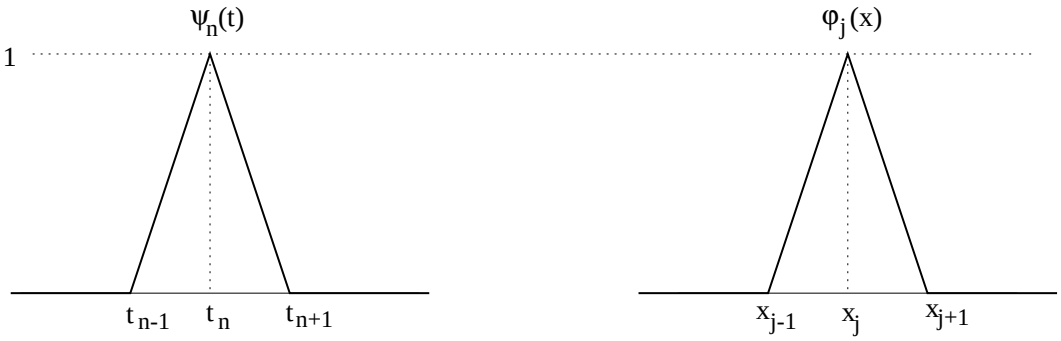
7.2.2 The finite element discretization procedure

We follow the same procedure as in the case of the heat equation, and let $S_n = \Omega \times I_n$, $n = 1, 2, \ldots, N$, with $I_n = (t_{n-1}, t_n]$. Then, for each n we define, on S_n, the piecewise linear approximations

$$\begin{cases} U(x,t) = U_{n-1}(x)\Psi_{n-1}(t) + U_n(x)\Psi_n(t), \\ V(x,t) = V_{n-1}(x)\Psi_{n-1}(t) + V_n(x)\Psi_n(t), \end{cases} \quad 0 < x < 1, \quad t \in I_n, \tag{7.2.10}$$

where, e.g.

$$\begin{cases} U_{\tilde{n}}(x) = U_{\tilde{n},1}(x)\varphi_1(x) + \ldots + U_{\tilde{n},m}(x)\varphi_m(x), \;\; \tilde{n} = n-1,\, n \\ V_{\tilde{n}}(x) = V_{\tilde{n},1}(x)\varphi_1(x) + \ldots + V_{\tilde{n},m}(x)\varphi_m(x), \;\; \tilde{n} = n-1,\, n. \end{cases} \tag{7.2.11}$$



For $\dot{u} - v = 0$ and $t \in I_n$ we write the general variational formulation

$$\int_{I_n}\int_0^1 \dot{u}\varphi \, dxdt - \int_{I_n}\int_0^1 v\varphi \, dxdt = 0, \quad \text{for all } \varphi(x,t). \tag{7.2.12}$$

Likewise, $\dot{v} - u'' = 0$ yields a variational formulation, viz

$$\int_{I_n}\int_0^1 \dot{v}\varphi \, dxdt - \int_{I_n}\int_0^1 u''\varphi \, dxdt = 0. \tag{7.2.13}$$

Integrating by parts in x, in the second term, and using the boundary condition $u'(1,t) = g(t)$ we get

$$\int_0^1 u''\varphi dx = [u'\varphi]_0^1 - \int_0^1 u'\varphi'\,dx = g(t)\varphi(1,t) - u'(0,t)\varphi(0,t) - \int_0^1 u'\varphi'\,dx.$$

Inserting the right hand side in (7.2.13) we get for all φ with $\varphi(0,t) = 0$:

$$\int_{I_n}\int_0^1 \dot{v}\varphi\,dxdt + \int_{I_n}\int_0^1 u'\varphi'\,dxdt = \int_{I_n} g(t)\varphi(1,t)\,dt. \tag{7.2.14}$$

The corresponding $cG(1)cG(1)$ finite element method reads as: For each n, $n = 1, 2, \ldots, N$, find continuous piecewise linear functions $U(x,t)$ and $V(x,t)$, in a partition, $0 = x_0 < x_1 < \cdots < x_m = 1$ of $\Omega = (0,1)$, such that

$$\begin{aligned}\int_{I_n}\int_0^1 \frac{U_n(x) - U_{n-1}(x)}{k_n}\varphi_j(x)\,dxdt \\ - \int_{I_n}\int_0^1 \Big(V_{n-1}(x)\Psi_{n-1}(t) + V_n(x)\Psi_n(t)\Big)\varphi_j(x)\,dxdt = 0, \\ \text{for } j = 1, 2, \ldots, m,\end{aligned} \tag{7.2.15}$$

and

$$\begin{aligned}\int_{I_n}\int_0^1 \frac{V_n(x) - V_{n-1}(x)}{k_n}\varphi_j(x)\,dxdt \\ + \int_{I_n}\int_0^1 \Big(U'_{n-1}(x)\Psi_{n-1}(t) + U'_n(x)\Psi_n(t)\Big)\varphi'_j(x)\,dxdt \\ = \int_{I_n} g(t)\varphi_j(1)\,dt, \qquad \text{for } j = 1, 2, \ldots, m,\end{aligned} \tag{7.2.16}$$

where $\dot{U}$, U', $\dot{V}$, and V' are computed using (7.2.10) with

$$\psi_{n-1}(t) = \frac{t_n - t}{k_n}, \qquad \psi_{n-1}(t) = \frac{t - t_{n-1}}{k_n}, \qquad k_n = t_n - t_{n-1}.$$

Thus, the equations (7.2.15) and (7.2.16) are reduced to the *iterative forms*:

$$\begin{aligned}&\underbrace{\int_0^1 U_n(x)\varphi_j(x)dx}_{MU_n} - \frac{k_n}{2}\underbrace{\int_0^1 V_n(x)\varphi_j(x)dx}_{MV_n} \\ &= \underbrace{\int_0^1 U_{n-1}(x)\varphi_j(x)dx}_{MU_{n-1}} + \frac{k_n}{2}\underbrace{\int_0^1 V_{n-1}(x)\varphi_j(x)\,dx}_{MV_{n-1}}, \quad j = 1, 2, \ldots, m,\end{aligned}$$

and

$$\underbrace{\int_0^1 V_n(x)\varphi_j(x)dx}_{MV_n} + \frac{k_n}{2}\underbrace{\int_0^1 U_n'(x)\varphi_j'(x)\,dx}_{SU_n}$$
$$= \underbrace{\int_0^1 V_{n-1}(x)\varphi_j(x)\,dx}_{MV_{n-1}} - \frac{k_n}{2}\underbrace{\int_0^1 U_{n-1}'(x)\varphi_j'(x)\,dx}_{SU_{n-1}} + g_n, \quad j = 1, 2, \ldots, m,$$

respectively, where we used (7.2.11) and as we computed earlier

$$S = \frac{1}{h}\begin{bmatrix} 2 & -1 & \ldots & 0 \\ -1 & 2 & -1 & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \quad M = \frac{h}{6}\begin{bmatrix} 4 & 1 & \ldots & 0 \\ 1 & 4 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & 1 & 4 & 1 \\ 0 & \ldots & 1 & 2 \end{bmatrix},$$

where

$$g_n = (0, \ldots, 0, g_{n,m})^T, \quad \text{where} \quad g_{n,m} = \int_{I_n} g(t)\,dt.$$

In compact form the vectors U_n and V_n are determined by solving the linear system of equations:

$$\begin{cases} MU_n - \frac{k_n}{2}MV_n = MU_{n-1} + \frac{k_n}{2}MV_{n-1} \\ \frac{k_n}{2}SU_n + MV_n = -\frac{k_n}{2}SU_{n-1} + MV_{n-1} + g_n, \end{cases} \tag{7.2.17}$$

which is a system of $2m$ equations with $2m$ unknowns:

$$\underbrace{\begin{bmatrix} M & -\frac{k_n}{2}M \\ \frac{k_n}{2}S & M \end{bmatrix}}_{A}\underbrace{\begin{bmatrix} U_n \\ V_n \end{bmatrix}}_{W} = \underbrace{\begin{bmatrix} M & \frac{k_n}{2}M \\ -\frac{k_n}{2}S & M \end{bmatrix}\begin{bmatrix} U_{n-1} \\ V_{n-1} \end{bmatrix} + \begin{bmatrix} 0 \\ g_n \end{bmatrix}}_{b},$$

with $W = A \setminus b$, $U_n = W(1:m)$ and $V_n = W(m+1:2m)$.

7.2.3 Exercises

Problem 7.12. *Derive the corresponding linear system of equations in the case of time discretization with dG(0).*

Problem 7.13 (discrete conservation of energy)**.** *Show that cG(1)-cG(1) for the wave equation in system form with $g(t) = 0$, conserves energy: i.e.*

$$\|U_n'\|^2 + \|V_n\|^2 = \|U_{n-1}'\|^2 + \|V_{n-1}\|^2. \tag{7.2.18}$$

Hint: Multiply the first equation by $(U_{n-1} + U_n)^t SM^{-1}$ and the second equation by $(V_{n-1}+V_n)^t$ and add up. Use then, e.g., the fact that $U_n^t S U_n = \|U_n'\|^2$, where

$$U_n = \begin{pmatrix} U_{n,1} \\ U_{n,2} \\ \dots \\ U_{n,m} \end{pmatrix}, \text{ and } U_n = U_n(x) = U_{n,1}(x)\varphi_1(x) + \ldots + U_{n,m}(x)\varphi_m(x).$$

Problem 7.14. *Consider the wave equation*

$$\begin{cases} \ddot{u} - u'' = 0, & x \in R, \quad t > 0, \\ u(x,0) = u_0(x), & x \in R, \\ \dot{u}(x,0) = v_0(x), & x \in R. \end{cases} \tag{7.2.19}$$

Plot the graph of $u(x, 2)$ in the following cases.
a) $v_0 = 0$ and

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

b) $u_0 = 0$, and

$$v_0(x) = \begin{cases} -1, & -1 < x < 0, \\ 1, & 0 < x < 1, \\ 0, & |x| > 0. \end{cases}$$

Problem 7.15. *Compute the solution for the wave equation*

$$\begin{cases} \ddot{u} - 4u'' = 0, & x > 0, & t > 0, \\ u(0,t) = 0, & & t > 0, \\ u(x,0) = 0, & \dot{u}(x,0) = 0, & x > 0. \end{cases} \tag{7.2.20}$$

Plot the solutions for the three cases $t = 0.5$, $t = 1$, $t = 2$, and with

$$u_0(x) = \begin{cases} 1, & x \in [2,3] \\ 0, & else \end{cases} \tag{7.2.21}$$

Problem 7.16. *Apply cG(1) time discretization directly to the wave equation by letting*

$$U(x,t) = U_{n-1}\Psi_{n-1}(t) + U_n(x)\Psi_n(t), \qquad t \in I_n. \tag{7.2.22}$$

Note that $\dot{U}$ is piecewise constant in time and comment on:

$$\underbrace{\int_{I_n}\int_0^1 \ddot{U}\varphi_j \, dxdt}_{?} + \underbrace{\int_{I_n}\int_0^1 u'\varphi_j' \, dxdt}_{\frac{k}{2}S(U_{n-1}+U_n)} = \underbrace{\int_{I_n} g(t)\varphi_j(1)dt}_{g_n}, \quad j = 1, 2, \ldots, m.$$

Problem 7.17. *Construct a FEM for the problem*

$$\begin{cases} \ddot{u} + \dot{u} - u'' = f, & 0 < x < 1, & t > 0, \\ u(0,t) = 0, & u'(1,t) = 0, & t > 0, \\ u(x,0) = 0, & \dot{u}(x,0) = 0, & 0 < x < 1. \end{cases} \tag{7.2.23}$$

Problem 7.18. *Determine the solution for the wave equation*

$$\begin{cases} \ddot{u} - c^2u'' = f, & x > 0, & t > 0, \\ u(x,0) = u_0(x), & u_t(x,0) = v_0(x), & x > 0, \\ u_x(1,t) = 0, & & t > 0, \end{cases}$$

in the following cases:

a) $f = 0$.

b) $f = 1, \quad u_0 = 0, \quad v_0 = 0$.

7.3 Convection - diffusion problems

Most multi-physical phenomena are described by the convection, diffusion and absorption: Problems of fluid- and gas dynamics, chemical reaction-diffusion, electromagnetic fields, charged particles collisions, etc, are often modeled as convection-diffusion and absorption type problems. In Chapter 5 we illustrated the finite element procedure for the one-dimensional stationary convection-diffusion (see Example 5.2). Here we shall give the derivation for a time-dependent convection diffusion problem in one-dimensional case. In the previous sections we illustrated combined space-time discretization. Here, we focus on a certain space-time (or only space) discretizations that is of vital interest for convection-dominated convection-diffusion equations: namely the *Streamline Diffusion Method* (SDM), which is also more adequate for piecewise discontinuous Galerkin approximations in time.

The higher dimensional case will be considered in part 2 of these notes.

•**A convection-diffusion model problem** We illustrate the convection-diffusion phenomenon by an example:

Example 7.4 (A convection model)**.** *Consider the traffic flow in a highway, viz the Fig below. Let* $\rho = \rho(x,t)$ *be the density of cars* $(0 \le \rho \le 1)$ *and* $u = u(x,t)$ *the velocity (speed vector) of the cars at the position* $x \in (a,b)$ *and time* t*. For a highway path* (a,b) *the difference between the traffic inflow* $u(a)\rho(a)$ *at the point* $x = a$ *and outflow* $u(b)\rho(b)$ *at* $x = b$ *gives the density variation on the interval* (a,b)*:*

$$\frac{d}{dt}\int_a^b \rho(x,t)dx = \int_a^b \dot{\rho}(x,t)dx = \rho(a)u(a) - \rho(b)u(b) = -\int_a^b (u\rho)'dx$$

or equivalently

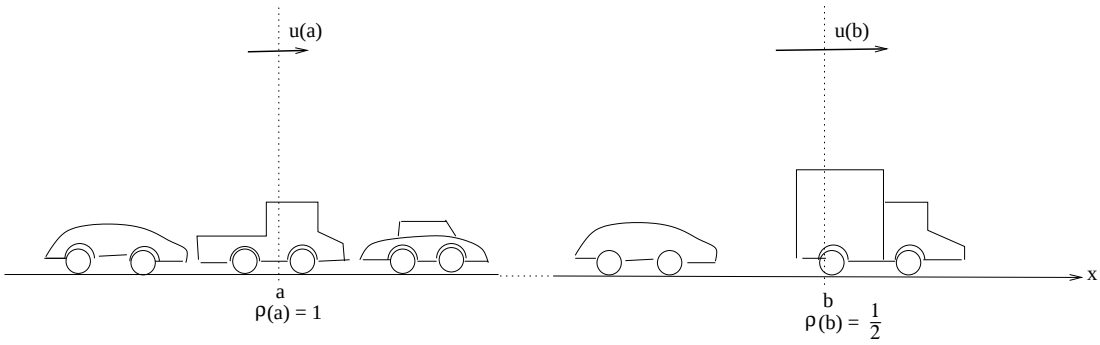
$$\int_a^b \left(\dot{\rho} + (u\rho)'\right) dx = 0. \tag{7.3.1}$$

Since a *and* b *can be chosen arbitrary, thus we have*

$$\dot{\rho} + (u\rho)' = 0. \tag{7.3.2}$$

Let now $u = 1 - \rho$*, (motivate this choice), then* (13.1.2) *is rewritten as*

$$\dot{\rho} + \Big((1-\rho)\rho\Big)' = \dot{\rho} + (\rho - \rho^2)' = 0. \tag{7.3.3}$$



Hence

$$\dot{\rho} + (1 - 2\rho)\rho' = 0 \qquad (A \text{ non-linear convection equation}). \tag{7.3.4}$$

Alternatively, (to obtain a convection-diffusion model), we may assume that $u = c - \varepsilon \cdot (\rho'/\rho),\ c > 0,\ \varepsilon > 0$, *(motivate). Then we get from* (13.1.2) *that*

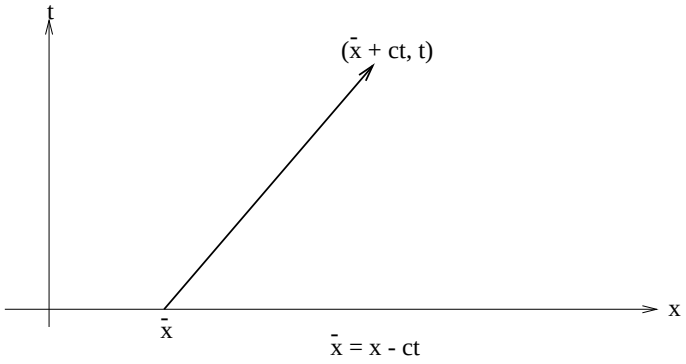
$$\dot{\rho} + \Big((c - \varepsilon\frac{\rho'}{\rho})\rho\Big)' = 0, \tag{7.3.5}$$

i.e.,

$$\dot{\rho} + c\rho' - \varepsilon\rho'' = 0 \quad (A \text{ convection-diffusion equation}). \tag{7.3.6}$$

The equation (13.1.6) *is convection dominated if* $c > \varepsilon$.

For $\varepsilon = 0$ *the solution is given by the exact transport* $\rho(x,t) = \rho_0(x - ct)$, *because then* $\rho =$ *constant on the characteristic's:* $(c, 1)$*-direction.*



Note that differentiating $\rho(x,t) = \rho(\bar{x} + ct, t)$ with respect to t we get

$$\frac{\partial \rho}{\partial x} \cdot \frac{\partial x}{\partial t} + \frac{\partial \rho}{\partial t} = 0, \quad \Longleftrightarrow \quad c\rho' + \dot{\rho} = 0. \tag{7.3.7}$$

Finally, let us rewrite the convection-diffusion equation (13.1.6)*, for ρ, by changing the notation from ρ to u, and replacing c by β, i.e.*

$$\dot{u} + \beta\, u' - \varepsilon\, u'' = 0. \tag{7.3.8}$$

Remark 7.3. *Compare this equation with the Navier-Stokes equations for incompressible flow:*

$$\dot{u} + (\beta \cdot \nabla)u - \varepsilon \Delta u + \nabla P = 0, \quad \wedge \quad div\, u = 0, \tag{7.3.9}$$

where $\beta = u$, $u = (u_1, u_2, u_3)$ is the velocity vector, with u_1 representing the mass, u_2 momentum, and $u_3 =$ energy. Further P is the pressure and $\varepsilon = 1/Re$ with Re denoting the Reynold's number. A typical range for the Reynold's number Re is between 10^5 and 10^7.

Therefore, for $\varepsilon > 0$ and small*, because of difficulties related to boundary layer and turbulence, the Navier-Stokes equations are not easily solvable.*

Example 7.5 (The boundary layer)**.** *Consider the following boundary value problem*

$$(BVP) \quad \begin{cases} u' - \varepsilon u'' = 0, & 0 < x < 1 \\ u(0) = 1, & u(1) = 0. \end{cases} \tag{7.3.10}$$

The exact solution to this problem is given by

$$u(x) = C\Big(e^{1/\varepsilon} - e^{x/\varepsilon}\Big), \qquad with \quad C = \frac{1}{e^{1/\varepsilon} - 1}. \tag{7.3.11}$$

which has an outflow boundary layer of width $\sim \varepsilon$, as seen in the Fig below

7.3.1 Finite Element Method

We shall now study the finite element approximation of the problem (13.1.10). To this end first we represent, as usual, the finite element solution by

$$U(x) = \varphi_0(x) + U_1\varphi_1(x) + \ldots + U_n\varphi_n(x), \tag{7.3.12}$$

where the φ_j:s are the basis function, here continuous piecewise linears (hat-functions) illustrated below:

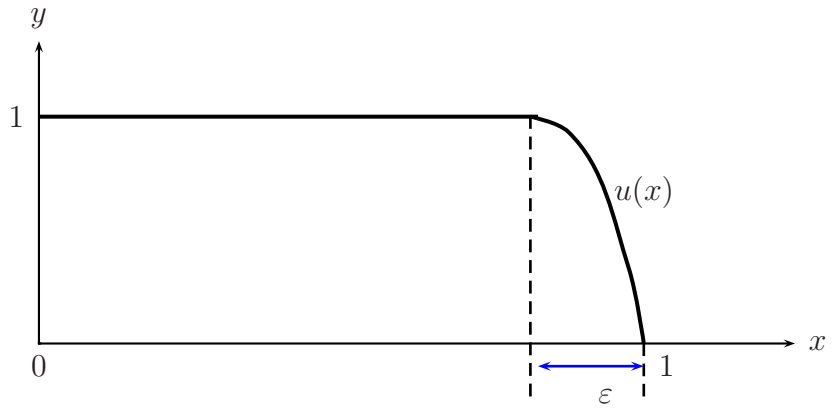
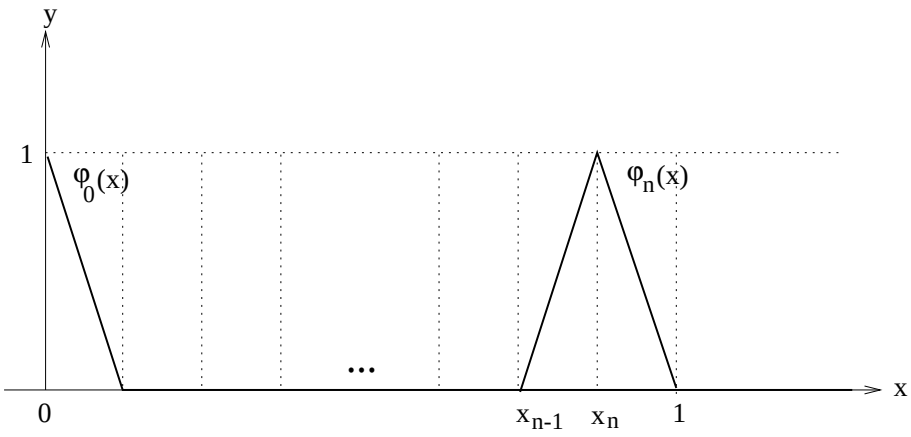


Figure 7.2: A ε boundary layer



Evidently, the corresponding variational formulation yields the FEM:

$$\int_0^1 \Big(U'\varphi_j dx + \varepsilon U'\varphi_j' \Big)\, dx = 0, \quad j = 1, 2, \ldots, n, \tag{7.3.13}$$

which may be represented by the equations

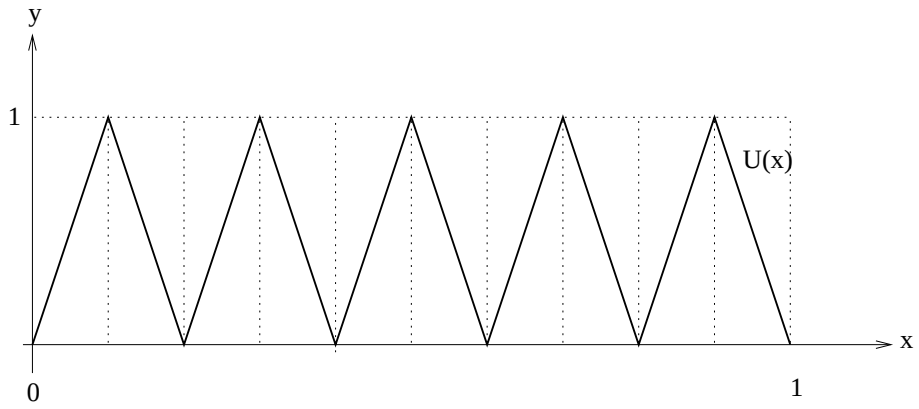
$$\frac{1}{2}\Big(U_{j+1} - U_{j-1} \Big) + \frac{\varepsilon}{h}\Big(2U_j - U_{j-1} - U_{j+1} \Big) = 0, \quad j = 1, 2, \ldots, n, \tag{7.3.14}$$

where $U_0 = 1$ and $U_{n+1} = 0$.

Note that, using *Central -differencing* we may also write

$$\underbrace{\frac{U_{j+1}-U_{j-1}}{2h}}_{\text{corresp. to } u'(x_j)} - \varepsilon \underbrace{\frac{U_{j+1}-2U_j+U_{j-1}}{h^2}}_{\text{corresp. to } u''(x_j)} = 0 \quad \Big(\Longleftrightarrow \frac{1}{h} \times \text{equation(13.1.14)} \Big).$$

Now for ε being very small this gives that $U_{j+1} \approx U_{j-1}$, which results, for even n values alternating 0 and 1 as the solution values at the nodes:



i.e., oscillations in U are transported "upstreams" making U a "globally bad approximation" of u.

A better approach would be to approximate $u'(x_j)$ by an *upwind derivative* as follows

$$u'(x_j) \approx \frac{U_j - U_{j-1}}{h}, \tag{7.3.15}$$

which, formally, gives a better stability, however, with low accuracy.

Remark 7.4. *The example above demonstrates that a high accuracy without stability is indeed useless.*

A more systematic method of making the finite element solution of the fluid problems stable is through using the streamline diffusion method which we, formally, introduce below.

7.3.2 The Streamline-diffusion method (SDM)

The idea is to choose, in the variational formulation, the test functions of the form $(v+\frac{1}{2}\beta hv')$, instead of just v (this would finally correspond to adding an extra diffusion to the original equation in the direction of the stream-lines). Then, e.g. for our model problem we obtain the equation ($\beta \equiv 1$)

$$\int_0^1 \Big[u'\Big(v+\frac{1}{2}hv'\Big) - \varepsilon \cdot u''\Big(v+\frac{1}{2}hv'\Big)\Big]dx = \int_0^1 f\Big(v+\frac{1}{2}hv'\Big)\,dx. \quad (7.3.16)$$

In the case of approximation with piecewise linears, in the discrete version of the variational formulation, we should interpret the term $\int_0^1 U''v'dx$ as a sum viz,

$$\int_0^1 U''v'dx := \sum_j \int_{I_j} U''v'dx = 0. \quad (7.3.17)$$

Then, with piecewise linear test functions, choosing $v = \varphi_j$ we get the discrete term corresponding to the second term in the first integral in (13.1.16) as

$$\int_0^1 U'\frac{1}{2}h\varphi_j'dx = U_j - \frac{1}{2}U_{j+1} - \frac{1}{2}U_{j-1}, \quad (7.3.18)$$

which adding to the obvious approximation of first term in the first integral:

$$\int_0^1 U'\varphi_j dx = \frac{U_{j+1}-U_{j-1}}{2}, \quad (7.3.19)$$

we end up with $(U_j - U_{j-1})$, as an approximation of the first integral in (13.1.16), corresponding to the *upwind scheme*.

Remark 7.5. *The SDM can also be viewed as a least-square method: Let $A = \frac{d}{dx}$, then $A^t = -\frac{d}{dx}$. Now u minimizes the expression $\|w' - f\|$ if $u' = Au = f$. This can be written as*

$$A^tAu = A^tf \iff -u'' = -f, \quad \textit{(the continuous form)}. \quad (7.3.20)$$

While multiplying $u' = Au = f$ by v and integrating over $(0,1)$ we have

$$\int_0^1 U'v'dx = \int_0^1 fv'dx \qquad \textit{(the weak form)}, \quad (7.3.21)$$

where we replaced u' by U'. Thus, the weak form for the discretized equation may be written as

$$(AU, v) = (f, v), \qquad \forall v \in V_h, \tag{7.3.22}$$

and

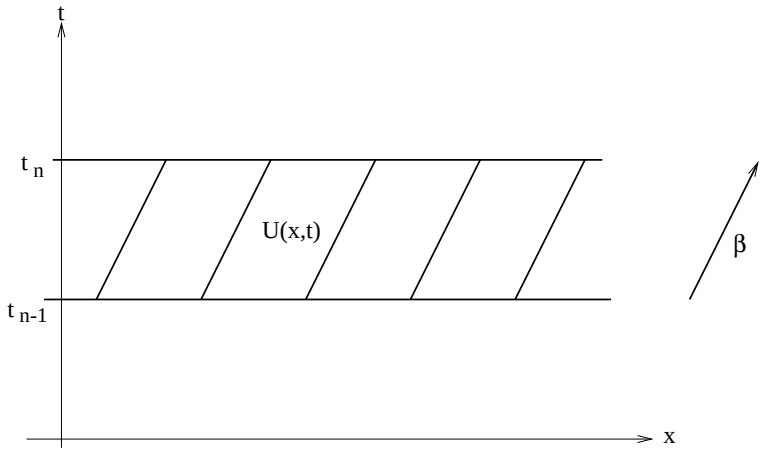
$$(AU, Av) = (f, Av), \qquad \forall v \in V_h. \tag{7.3.23}$$

Thus we have the SDM form

$$(AU, v + \delta Av) = (f, v + \delta Av), \qquad \forall v \in V_h. \tag{7.3.24}$$

For the time-dependent convection equation, the oriented time-space element are used. Consider the time-dependent problem

$$\dot{u} + \beta u' - \varepsilon u'' = f. \tag{7.3.25}$$



Set $U(x,t)$ such that U is piecewise linear in x and piecewise constant in the $(\beta, 1)$-direction. Combine with SDM and add up some artificial viscosity, $\hat{\varepsilon}$, depending on the residual term to get for each time interval I_n, the scheme:

$$\int_{I_n}\int_{\Omega}\Big[(\dot{U} + \beta U)\Big(v + \frac{\beta}{2}h\dot{v}\Big) + \hat{\varepsilon}\; U'v'\Big]\, dxdt = \int_{I_n}\int_{\Omega} f\Big(v + \frac{\beta}{2}hv'\Big)\, dxdt. \quad \square$$

7.3.3 Exercises

Problem 7.19. *Prove that the solution u of the convection-diffusion problem*

$$-u_{xx} + u_x + u = f, quadin\ I = (0,1), \quad u(0) = u(1) = 0,$$

satisfies the following estimate

$$\Big(\int_I u^2\phi\,dx\Big)^{1/2} \le \Big(\int_I f^2\phi\,dx\Big)^{1/2}.$$

where $\phi(x)$ is a positive weight function defined on $(0,1)$ satisfying $\phi_x(x) \le 0$ and $-\phi_x(x) \le \phi(x)$ for $0 \le x \le 1$.

Problem 7.20. *Let ϕ be a solution of the problem*

$$-\varepsilon\phi'' - 3\phi' + 2\phi = e, \qquad \phi'(0) = \phi(1) = 0.$$

Let $\|\cdot\|$ denote the L_2-norm on I. Show that there is a constant C such that

$$|\phi\prime(0)| \le C\|e\|, \qquad \|\varepsilon\phi''\| \le C\|e\|.$$

Problem 7.21. *Use relevant interpolation theory estimates and prove an a priori and an a posteriori error estimate for the $cG(1)$ finite element method for the problem*

$$-u'' + u' = f, \quad in\ I = (0,1), \quad u(0) = u(1) = 0.$$

Problem 7.22. *Prove an a priori and an a posteriori error estimate for the $cG(1)$ finite element method for the problem*

$$-u'' + u' + u = f, \quad in\ I = (0,1), \quad u(0) = u(1) = 0.$$

Problem 7.23. *Consider the problem*

$$-\varepsilon u'' + xu' + u = f, \quad in\ I = (0,1), \qquad u(0) = u'(1) = 0,$$

where ε is a positive constant, and $f \in L_2(I)$. Prove that

$$||\varepsilon u''|| \le ||f||.$$

Problem 7.24. *We modify the problem 7.23 above according to*

$$-\varepsilon u'' + c(x)u' + u = f(x) \quad 0 < x < 1, \qquad u(0) = u'(1) = 0,$$

where ε is a positive constant, the function c satisfies $c(x) \geq 0$, $c'(x) \leq 0$, and $f \in L_2(I)$. Prove that there are positive constants C_1, C_2 and C_3 such that

$$\sqrt{\varepsilon}||u'|| \leq C_1||f||, \quad ||cu'|| \leq C_2||f||, \qquad \text{and} \quad \varepsilon||u''|| \leq C_3||f||,$$

where $||\cdot||$ is the $L_2(I)$-norm.

Problem 7.25. *Consider the convection-diffusion-absorption problem*

$$-\varepsilon u_{xx} + u_x + u = f, \quad \text{in} \quad I = (0,1), \quad u(0) = 0, \qquad \sqrt{\varepsilon}u_x + u(1) = 0,$$

where ε is a positive constant, and $f \in L_2(I)$. Prove the following stability estimates for the solution u

$$\|\sqrt{\varepsilon}u_x\| + \|u\| + |u(1)| \leq C\|f\|,$$

$$\|u_x\| + \|\varepsilon u_{xx}\| \leq C\|f\|,$$

where $\|\cdot\|$ denotes the $L_2(I)$-norm, $I = (0,1)$, and C is an appropriate constant.

Chapter 8

Piecewise polynomials in several dimensions

8.1 Introduction

•Variational formulation in $\mathbb{R}^2$

All the previous studies in the 1 - dimensional case can be extended to $\mathbb{R}^n$, then the *mathematics of computation* becomes much more cumbersome. On the other hand, the two and three dimensional cases are the most relevant cases from both physical as well as practical point of views. A typical problem to study is, e.g.

$$\begin{cases} -\Delta u + au = f, & \mathbf{x} := (x, y) \in \Omega \subset R^2 \\ u(x, y) = 0, & (x, y) \in \partial\Omega. \end{cases} \tag{8.1.1}$$

The discretization procedure, e.g. with piecewise linears, would require the extensions of the interpolation estimates from the intervals in $1D$ to higher dimensions. Other basic concepts such as Cauchy-Shwarz and Poincare inequalities are also extended to the correspoding inequalities in $\mathbb{R}^n$. Due to the integrations involved in the variational formulation, a frequently used difference, from the 1-dimensional case, is in the performance of the partial integrations which is now replaced by the following well known formula:

Lemma 8.1 (Green's formula)**.** *Let $u \in C^2(\Omega)$ and $v \in C^1(\Omega)$, then*

$$\begin{aligned}\iint_\Omega \Big(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\Big) v dx dy = \int_{\partial\Omega} \Big(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\Big) \cdot \mathbf{n}(x,y) v ds \\ - \iint_\Omega \Big(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\Big) \cdot \Big(\frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}\Big) dx dy,\end{aligned} \tag{8.1.2}$$

where $\mathbf{n}(x,y)$ is the outward unit normal at the boundary point $\mathbf{x} = (x,y) \in \partial\Omega$ and ds is a curve element on the boundary $\partial\Omega$. In concise form

$$\int_\Omega (\Delta u) v dx = \int_\Omega (\nabla u \cdot \mathbf{n}) v ds - \int_\Omega \nabla u \cdot \nabla v dx. \tag{8.1.3}$$

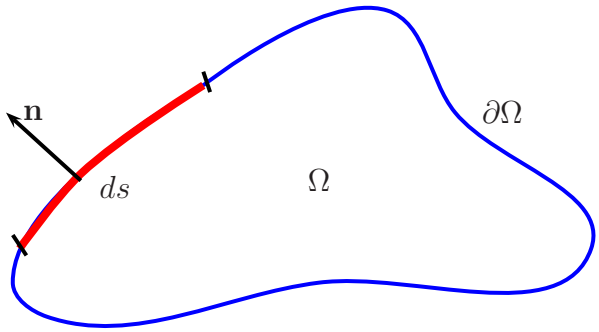


Figure 8.1: A smooth domain Ω with an outward unit normal $\mathbf{n}$

In the case that Ω is a rectangular domain. Then we have that

$$\begin{aligned}\iint_\Omega \frac{\partial^2 u}{\partial x^2} v dx dy &= \int_0^b \int_0^a \frac{\partial^2 u}{\partial x^2}(x,y) \cdot v(x,y) dx dy = [P.I.] \\ &= \int_0^b \Big(\Big[\frac{\partial u}{\partial x}(x,y) \cdot v(x,y)\Big]_{x=0}^a - \int_0^a \frac{\partial u}{\partial x}(x,y) \cdot \frac{\partial v}{\partial x}(x,y) dx \Big) dy \\ &= \int_0^b \Big(\frac{\partial u}{\partial x}(a,y) \cdot v(a,y) - \frac{\partial u}{\partial x}(0,y) \cdot v(0,y) \Big) dy - \\ &- \iint_\Omega \frac{\partial u}{\partial x} \cdot \frac{\partial v}{\partial x}(x,y) dx dy.\end{aligned}$$

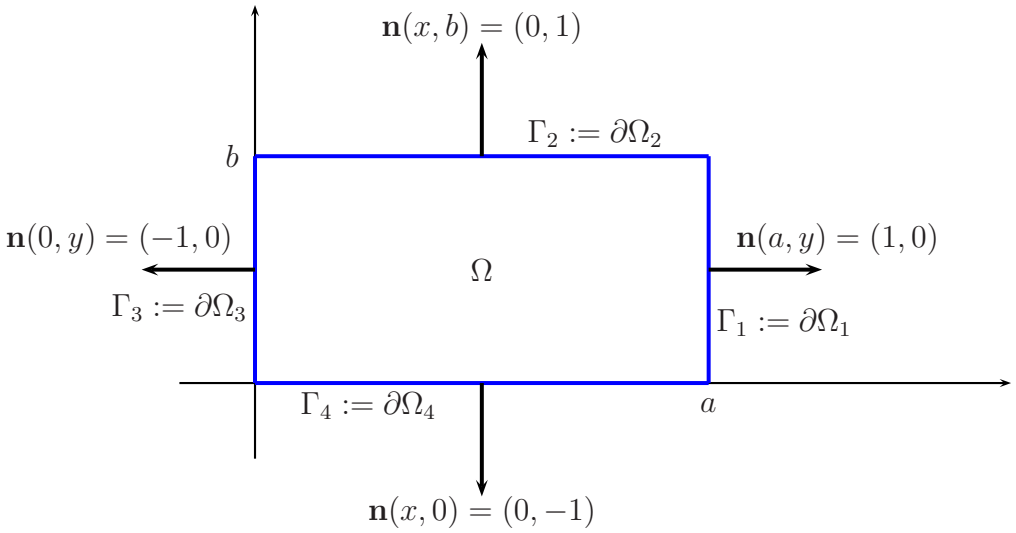


Figure 8.2: A rectangular domain Ω with its outward unit normals

Now we have on $\Gamma_1 : \mathbf{n}(a,y) = (1,0)$

on $\Gamma_2 : \mathbf{n}(x,b) = (0,1)$

on $\Gamma_3 : \mathbf{n}(0,y) = (-1,0)$

on $\Gamma_4 : \mathbf{n}(x,0) = (0,-1)$

Thus the first integral on the right hand side can be written as

$$\int_{\partial\Omega} \Big(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\Big) \cdot \mathbf{n}(x,y) v ds = \Big(\int_{\Gamma_1} + \int_{\Gamma_3}\Big)\Big(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\Big) \cdot \mathbf{n}(x,y) v(x,y) ds$$

and hence

$$\iint_\Omega \frac{\partial^2 u}{\partial x^2} dxdy = \int_{\Gamma_1 \cup \Gamma_3} \Big(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\Big) \cdot \mathbf{n}(x,y) v(x,y) ds - \iint_\Omega \frac{\partial u}{\partial x} \cdot \frac{\partial v}{\partial x} dxdy$$

Similarly, for the y-direction we get

$$\iint_\Omega \frac{\partial^2 u}{\partial y^2} v dxdy = \int_{\Gamma_2 \cup \Gamma_4} \Big(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\Big) \cdot \mathbf{n}(x,y) v(x,y) ds - \iint_\Omega \frac{\partial u}{\partial y} \cdot \frac{\partial v}{\partial y} dxdy.$$

Now adding up these two recent relations gives the desired result. The case of general domain Ω, is a routine proof in the calculus of several variables. □

8.2 Piecewise linear approximation in 2 D

The objective in this part is the study of piecewise polynomial approximations for the solutions for differential equations in two dimensional spatial domains. In this setting, and for simplicity, we focus on piecewise linear polynomials and polygonal domains. Thus we shall deal with triangular mesh without any concerns about curved boundary.

8.2.1 Basis functions for the piecewise linears in 2 D

We recall that in the 1-dimensional case a function which is linear on a subinterval is uniquely determined by its values at the endpoints. (There is only one straight line connecting two points)

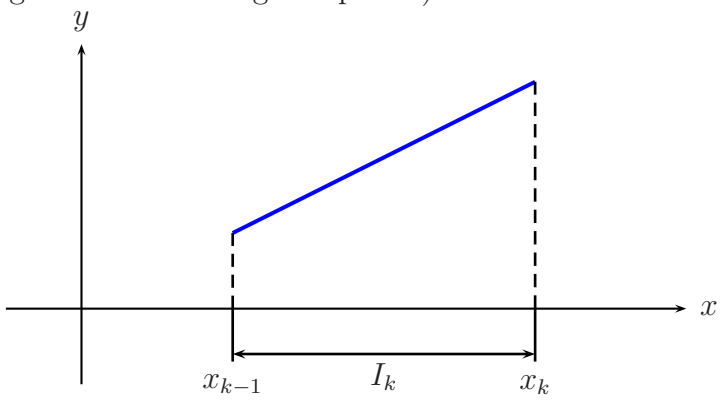
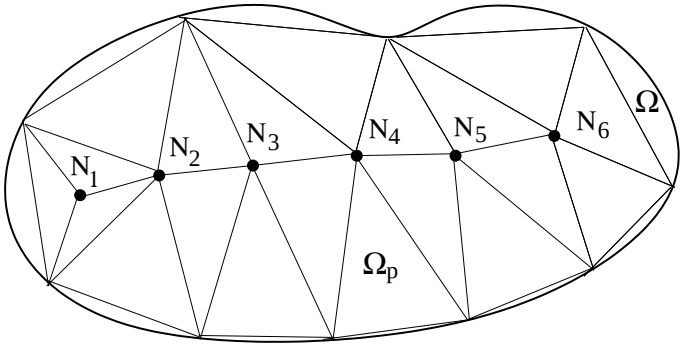


Figure 8.3: A picewise linear function on a subinterval $I_k = (x_{k-1}, x_k)$.

Similarly a plane in $\mathbb{R}^3$ is uniquely determined by three points. Therefore it is natural to make partitions of 2-dimensional domains using triangular elements and letting the sides of the triangles to correspond to the endpoints of the intervals in the 1-dimensional case.

The figure illustrates a "partitioning": *triangulation* of a domain Ω with curved boundary where the partitioning is performed only for a polygonal domain Ω_P generated by Ω (a domains with polygonal boundary). Here we have 6 internal nodes $N_i,\ 1 \leq i \leq 6$ and Ω_p is the *polygonal* domain inside Ω, which is triangulated. The figure 1.4 illustrates a piecewise linear function



on a single triangle which is determined by its values at the vertices of the triangle.

Now for every linear function U on Ω_p we have

$$U(\mathbf{x}) = U_1\varphi_1(\mathbf{x}) + U_1\varphi_2(\mathbf{x}) + \ldots + U_6\varphi_6(\mathbf{x}), \tag{8.2.1}$$

where $U_i = U(N_i), \quad i = 1, 2, \ldots, 6$ are numbers (nodal values) and $\varphi_i(N_i) = 1$, while $\varphi_i(N_j) = 0$ for $j \neq i$. Further $\varphi_i(\mathbf{x})$ is linear in $\mathbf{x}$ in every triangle/element. In other words

$$\varphi_i(N_j) = \left\{ \begin{array}{ll} 1, & j = i \\ 0, & j \neq i \end{array} \right\} = \delta_{ij} \quad \text{(affin)} \tag{8.2.2}$$

and, for instance with the Dirichlet boundary condition we take $\varphi_i(\mathbf{x}) = 0$ on $\partial\Omega_p$.

In this way given a differential equation, to determine the approximate solution U is now reduced to find the values (numbers) $U_1, U_2, \ldots, U_6$, satisfying the corresponding variational formulation. For instance if we chosse $\mathbf{x} = N_5$, then $U(N_5) = U_1\varphi_1(N_5) + U_2\varphi_2(N_5) + \ldots + U_5\varphi_5(N_5) + U_6\varphi_6(N_5)$, where $\varphi_1(N_5) = \varphi_2(N_5) = \varphi_3(N_5) = \varphi_4(N_5) = \varphi_6(N_5) = 0$ and $\varphi_5(N_5) = 1$, and hence

$$U(N_5) = U_5\varphi_5(N_5) = U_5 \tag{8.2.3}$$

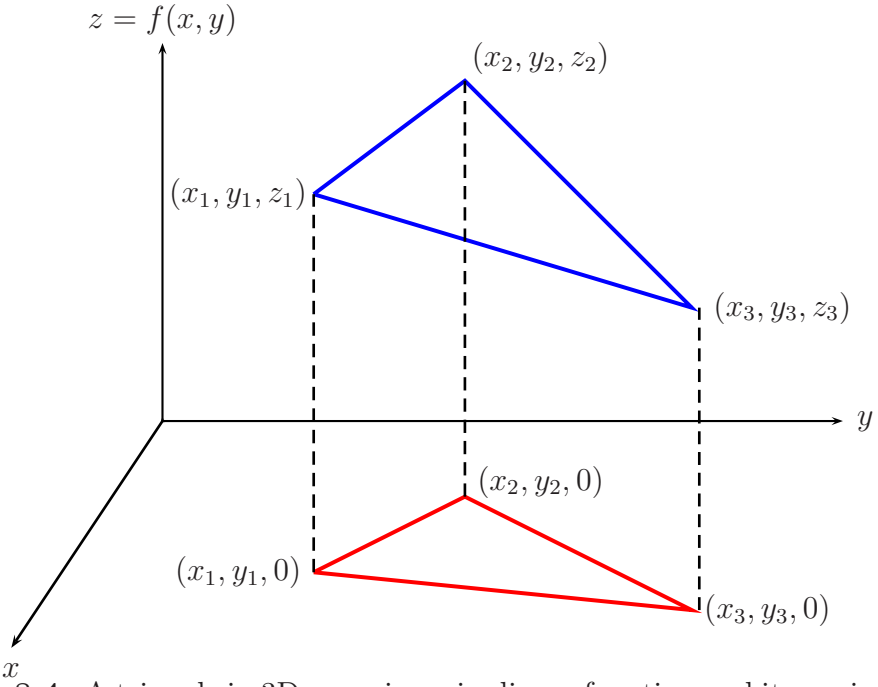
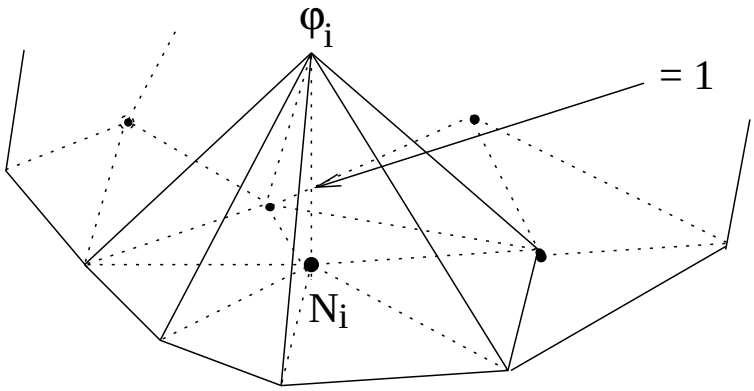


Figure 8.4: A triangle in 3D as a piecewise linear function and its projection in 2D.



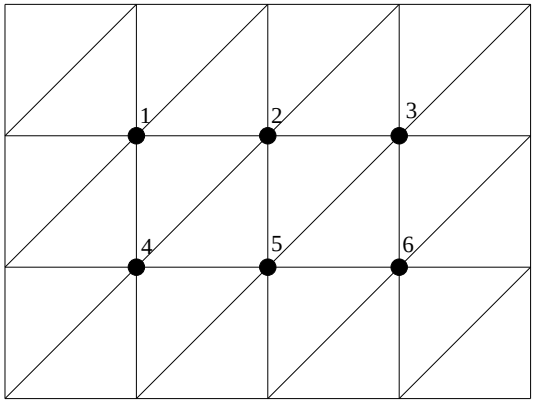
Example 8.1. *, let* $\Omega = \{(x, y) : 0 < x < 4, 0 < y < 3\}$ *and make a FEM discretization of the following boundary value problem:*

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{8.2.4}$$

The variational formulation reads as follows: Find a function u vanishing at the boundary $\Gamma = \partial\Omega$ of Ω, such that

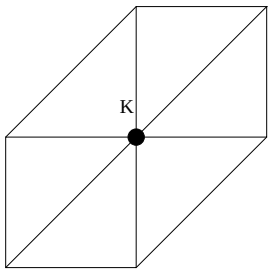
$$\iint_\Omega (\nabla u \cdot \nabla v) dxdy = \iint_\Omega f v dxdy, \qquad \forall v \in H_0^1(\Omega). \tag{8.2.5}$$

Note that $H_0^1(\Omega)$ is the space of continuously differetiable functions in Ω which are vanishing at the boundary $\partial\Omega$. Now we shall make a test function space of piecewise linears. To this approach we triangulate Ω as in the figure below and let



$V_h^0 = \{v \in \mathcal{C}(\Omega) : v$ *is linear on each sub-triangle and is 0 at the boundary.*$\}$

Since such a function is uniquely determined by its values at the vertices of the triangles and 0 on the boundary, so indeed in our example we have only 6 inner vertices of interest. Now precisely as in the "$1-D$" case we construct basis functions. (6 of them in this particular case), with values 1 at one of the nodes and zero at the others. Then we get the two-dimensional telt functions *as shown in the figure above.*



8.2.2 Error estimates for piecewise linear interpolation

In this section we make a straightforward generalization of the one dimenensional linear interpolation estimate on an interval in the maximum norm to a two dimensional linear interpolation on a triangle. As in the 1D case, our estimate indicates that the interpolation error depends on the second order, this time, partial derivatives of the functions being interpolated, i.e., the *curvature* of the functions, mesh size and also the shape of the triangle. The results are also extended to other L_p, $1 \leq p < \infty$ norms as well as higher dimensions than 2.

To continue we assume a triangulation $\mathcal{T} = \{K\}$ of a two dimensional polygonal domain Ω. We let v_i, $i = 1, 2, 3$ be the vertices of the triangle K. Now we consider a continuous function f defined on K and define the linear interpolant $\pi_h f \in \mathcal{P}^1(K)$ by

$$\pi_h f(v_i) = f(v_i), \qquad i = 1, 2, 3. \tag{8.2.6}$$

This is illustrated in the figure on the next page. We shall now state some basic interpolation results that we frequently use in the error estimates. The proofs of these results are given in CDE, by Eriksson et al.

Theorem 8.1. *If f has contionuous second order partial derivatives, then*

$$\|f - \pi_h f\|_{L_\infty(K)} \leq \frac{1}{2} h_K^2 ||D^2 f||_{L_\infty(K)}, \tag{8.2.7}$$

$$||\nabla(f - \pi_h f)||_{L_\infty(K)} \leq \frac{3}{\sin(\alpha_K)} h_K ||D^2 f||_{L_\infty(K)}, \tag{8.2.8}$$

where h_K is the largest side of K, α_K is the smallest angle of K, and

$$D^2 f = \Big(\sum_{i,j=1}^{2} (\frac{\partial^2 f}{\partial x_i \partial x_j})^2 \Big)^{1/2}.$$

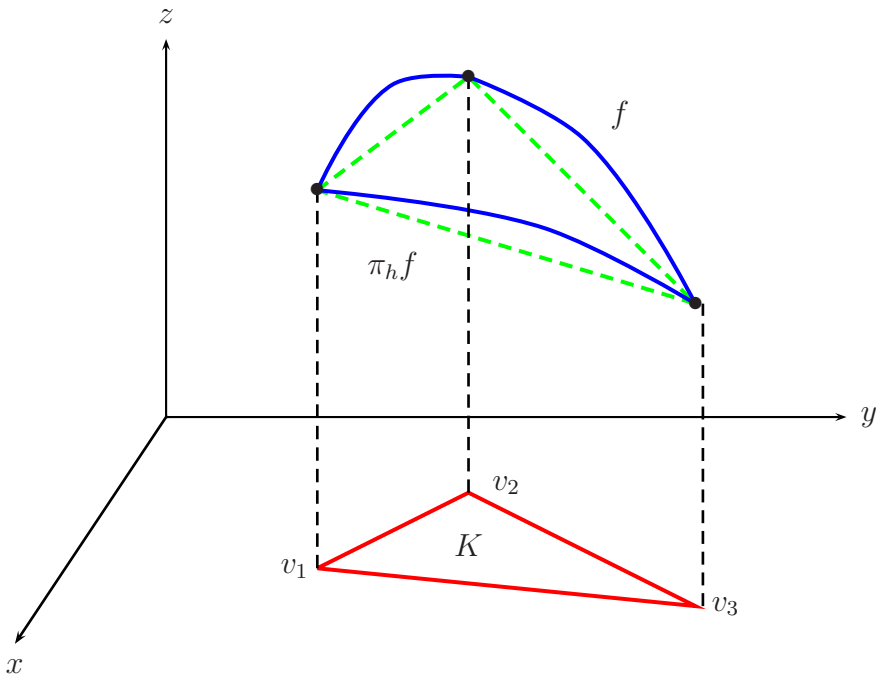


Figure 8.5: The nodal interpolant of f in 2D case

Remark 8.1. *Note that the gradient estimate* (8.2.8) *deteriotes for small* $\sin(\alpha_K)$*; i.e. for the thinner triangle* K*. This phenomenon is avoided assuming a quasi-uniform triangulation, where there is a minimum angle condition for the triangles viz,*

$$\sin(\alpha_K) \geq C, \qquad \textit{for some constant } C. \tag{8.2.9}$$

8.2.3 The L_2 projection

Definition 8.1. *Let* V_h *be the space of all continuous linear functions on a triangulation* $\mathcal{T}_h = \{K\}$ *of the domain* Ω*. The* L_2 *projection* $P_h u \in V_h$ *of a function* $u \in L_2(\Omega)$ *is defined by*

$$(u - P_h u, v) = 0, \qquad \forall v \in V_h. \tag{8.2.10}$$

This means that, the error $u - P_h u$ is orthogonal to V_h. (8.2.10) yields a linear system of equations for the coefficients of $P_h u$ with respect to the nodal basis of V_h.

Advantages of the L_2 projection to the nodal interpolation

• The L_2 projection $P_h u$ is well defined for $u \in L_2(\Omega)$, whereas the nodal interpolant $\pi_h u$ in general requires u to be continuous. Therefore the L_2 projection is an alternative for the nodal interpolation for, e.g. discontinuous L_2 functions.

• Letting $v \equiv 1$ in (8.2.10) we have that

$$\int_\Omega P_h u \, dx = \int_\Omega u \, dx. \tag{8.2.11}$$

Thus the L_2 projection conserves the *total mass*, whereas, in general, the nodal interpolation operator does not preserve the total mass.

• Finally we have the following error estimate for the L_2 projection:

Theorem 8.2.

$$\|u - \pi_h u\| \leq C_i \|h^2 D^2 u\|. \tag{8.2.12}$$

Proof. We have using (8.2.10) and the Cauchy's inequality that

$$\begin{aligned} \|u - \pi_h u\|^2 &= (u - \pi_h u, u - \pi_h u) \\ &(u - \pi_h u, u - v) + (u - \pi_h u, v - \pi_h u) = (u - \pi_h u, u - v) \\ &\leq \|u - \pi_h u\| \|u - v\|. \end{aligned} \tag{8.2.13}$$

This yields

$$\|u - \pi_h u\| \leq \|u - v\|, \qquad \forall v \in V_h. \tag{8.2.14}$$

Now choosing $v = \pi_h u$ and recalling the interpolation theoren above we get the desired result. □

8.3 Exercises

Problem 8.1. *Show that the function $u : \mathbb{R}^2 \to \mathbb{R}$ given by $u(x) = \log(|x|^{-1}),\ x \neq 0$ is a solution to the Laplace equation $\Delta u(x) = 0$.*

Problem 8.2. *Show that the Laplacian of a C^2 function $u : \mathbb{R}^2 \to \mathbb{R}$ in the polar coordinates is written by*

$$\Delta u = \frac{1}{r} \frac{\partial}{\partial r} \Big(r \frac{\partial u}{\partial r} \Big) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2}. \tag{8.3.1}$$

Problem 8.3. *Show using* (8.3.1) *that the function* $u = a\log(r) + b$ *where* a *and* b *are arbitrary constants is a solution of the Laplace equation* $\Delta u(x) = 0$ *for* $x \neq 0$. *Are there any other solutions of the Laplace equation in* $\mathbb{R}^2$ *which are invariant under rotation (i.e. it depends only on* $r = |x|$*)?*

Problem 8.4. *For a given triangle* K, *determine the relation between the smallest angle* τ_K, *the triangle diameter* h_K *and the diameter* ρ_K *of the largest inscribed circle.*

Problem 8.5. *Prove that a linear function in* $\mathbb{R}^2$ *is uniquely determined by its at three points as long as they don't lie on a straight line.*

Problem 8.6. *Let* K *be a triangle with nodes* $\{a^i\}$, $i = 1, 2, 3$ *and let the midpoints of the edges be denoted* $\{a^{ij}, 1 \leq i < j \leq 3\}$.
a) Show that a function $v \in \mathcal{P}^1(K)$ *is uniqely determined by the degrees of freedom* $\{v(a^{ij}), 1 \leq i < j \leq 3\}$.
b) Are functions continuous in the corresponding finite element space of piecewise linear functions?

Problem 8.7. *Prove that if* K_1 *and* K_2 *are two neighboring triangles and* $w_1 \in \mathcal{P}^2(K_1)$ *and* $w_2 \in \mathcal{P}^2(K_2)$ *agree at three nodes on the common boundary (e.g., two endpoints and a midpoint), then* $w_1 \equiv w_2$ *on the common boundary.*

Problem 8.8. *Prove that a linear function is uniquely determined by its values at three points, as long as they don't lie on a straight line.*

Problem 8.9. *Assume that the triangle* K *has nodes at* $\{v^1, v^2, v^3\}$, $v^i = (v_1^i, v_2^i)$, *the element nodal basis is the set of functions* $\lambda_i \in \mathcal{P}^1(K)$, $i = 1, 2, 3$ *such that*

$$\lambda_i(v^j) = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

Compute the explicit formulas for λ_i.

Problem 8.10. *Let* K *be a triangular element. Show the following identities, for* j, $k = 1, 2$, *and* $x \in K$,

$$\sum_{i=1}^{3} \lambda_i(x) = 1, \qquad \sum_{i=1}^{3} (v_j^i - x_j)\lambda_i(x) = 0, \tag{8.3.2}$$

$$\sum_{i=1}^{3} \frac{\partial}{\partial x_k}\lambda_i(x) = 0, \qquad \sum_{i=1}^{3}(v_j^i - x_j)\frac{\partial \lambda_i}{\partial x_k} = \delta_{jk}, \tag{8.3.3}$$

where $v^i = (v_1^i, v_2^i)$, $i = 1, 2, 3$ *are the vertices of* K, $x = (x_1, x_2)$ *and* $\delta_{jk} = 1$ *if* $j = k$ *and* $\delta_{jk} = 0$ *otherwise.*

Problem 8.11. *Using* (8.3.2)*, we obtain a representation for the interpolation error of the form*

$$f(x) - \pi_h f(x) = -\sum_{i=1}^{3} r_i(x)\lambda_i(x). \tag{8.3.4}$$

Prove that the remainder term $r_i(x)$ *can be estimated as*

$$|r_i(x)| \leq \frac{1}{2} h_K \|D^2 f\|_{L_\infty(K)}, \qquad i = 1, 2, 3. \tag{8.3.5}$$

Hint: (I) Note that $|v^i - x| \leq h_K$*. (II) Start applying Cauchy's inequality to show that*

$$\sum_{ij} x_i c_{ij} x_j = \sum_i x_i \sum_j c_{ij} x_j.$$

Problem 8.12. τ_K *is the smallest angle of a triangular element* K*. Show that*

$$\max_{x \in K} |\nabla \lambda_i(x)| \leq \frac{2}{h_K \sin(\tau_K)}.$$

Problem 8.13. *The Euler equation for an incompressible inviscid fluid of density can be written as*

$$u_t + (u \cdot \nabla)u + \nabla p = f, \qquad \nabla \cdot u = 0, \tag{8.3.6}$$

where $u(x, t)$ *is the velocity and* $p(x, t)$ *the pressure of the fluid at the pint* x *at time* t *and* f *is an applied volume force (e.g., a gravitational force). The second equation* $\nabla \cdot u = 0$ *expresses the incopressibility. Prove that the first equation follows from the Newton's law.*

Hint: Let $u = (u_1, u_2)$ *with* $u_i = u_i(x(t), t)$, $i = 1, 2$ *and use the chain rule to derive* $\dot{u}_i = \frac{\partial u_i}{\partial x_1}u_1 + \frac{\partial u_i}{\partial x_2}u_2 + \frac{\partial u_i}{\partial t}$, $\quad i = 1, 2$.

Problem 8.14. *Prove that if* $u : \mathbb{R}^2 \to \mathbb{R}^2$ *satisfies* $rot\, u := \left(\frac{\partial u_2}{\partial x_1}, -\frac{\partial u_1}{\partial x_2}\right) = 0$ *in a convex domian* $\Omega \subset \mathbb{R}^2$*, then there is a scalar function* φ *defined on* Ω *such that* $u = \nabla\varphi$ *in* Ω*.*

Problem 8.15. *Prove that* $\int_\Omega rot\, u\, dx = \int_\Gamma \mathbf{n} \times u\, ds$*, where* Ω *is a subset of* $\mathbb{R}^3$ *with boundary* Γ *with outward unit normal* $\mathbf{n}$*.*

Chapter 9

Riesz and Lax-Milgram Theorems

9.1 Preliminaries

In part I, we proved under certain assumptions that to solve a boundary value problem (BVP) is equavalent to a corresponding variational formulation (VF) which in turn is equivalent to a minimization problem (MP):

$$\text{BVP} \iff \text{VF} \iff \text{MP}.$$

More precisely we had the following 1-dimensional boundary value problem:

$$(BVP): \qquad \begin{cases} -\Big(a(x)u'(x)\Big)' = f(x), \quad 0 < x < 1 \\ u(0) = u(1) = 0, \end{cases} \tag{9.1.1}$$

with the corresponding variational formulation, viz

(VF): Find $u(x)$, with $u(0) = u(1) = 0$, such that

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \quad \forall v \in H_0^1, \tag{9.1.2}$$

where $H_0^1 := H_0^1(0,1)$ is the Sobolev space of all square integrable functions having square integrable first order derivatives on $(0,1)$ and vanishing at the boundary of the interval $(0,1)$:

$$H_0^1 = \Big\{ v : \int_0^1 \Big(v(x)^2 + v'(x)^2 \Big) dx < \infty, \quad v(0) = v(1) = 0 \Big\}, \tag{9.1.3}$$

and a minimization problem as:

(MP): Find $u(x)$, with $u(0) = u(1) = 0$, such that $u(x)$ minimizes the *functional* F given by

$$F(v) = \frac{1}{2} \int_0^1 v'(x)^2 dx - \int_0^1 f(x) v(x) dx. \tag{9.1.4}$$

Recalling Poincare inequality we may actually take instead of H_0^1, the space

$$\mathcal{H}_0^1 = \Big\{ f : [0,1] \to \mathbb{R} : \int_0^1 f'(x)^2 dx < \infty, \wedge f(0) = f(1) = 0 \Big\}. \tag{9.1.5}$$

Let now V be a *vector space* of function on $(0,1)$ and define a *bilinear form* on V; $a(\cdot,\cdot) : V \times V \to \mathbb{R}$, i.e. for $\alpha, \beta, x, y \in \mathbb{R}$ and $u, v, w \in V$, we have

$$\begin{cases} a(\alpha u + \beta v, w) = \alpha \cdot a(u,w) + \beta \cdot a(v,w) \\ a(u, xv + yw) = x \cdot a(u,v) + y \cdot a(u,w) \end{cases} \tag{9.1.6}$$

Example 9.1. *Let $V = \mathcal{H}_0^1$ and define*

$$a(u,v) := (u,v) := \int_0^1 u'(x) v'(x) dx, \tag{9.1.7}$$

then $(\cdot,\cdot)$ is symmetric, i.e. $(u,v) = (v,u)$, bilinear (obvious), and positive definite in the sense that

$$(u,u) \geq 0, \quad \text{and } (u,u) = 0 \Longleftrightarrow u \equiv 0.$$

Note that

$$(u,u) = \int_0^1 u'(x)^2 dx = 0 \Longleftrightarrow u'(x) = 0,$$

thus $u(x)$ is constant and since $u(0) = u(1) = 0$ we have $u(x) \equiv 0$.

Definition 9.1. *A linear function $L : V \to \mathbb{R}$ is called a linear form on V: If*

$$L(\alpha u + \beta v) = \alpha L(u) + \beta L(v). \tag{9.1.8}$$

Example 9.2. *Let*

$$\ell(v) = \int_0^1 f v \, dx, \qquad \forall v \in \mathcal{H}_0^1, \tag{9.1.9}$$

Then our (VF) can be restated as follows: Find $u \in \mathcal{H}_0^1$ such that

$$(u, v) = \ell(v), \qquad \forall v \in \mathcal{H}_0^1. \tag{9.1.10}$$

Generalizing the above example we get the following *abstract problem*: Find $u \in V$, such that

$$a(u, v) = L(v), \qquad \forall v \in V. \tag{9.1.11}$$

Definition 9.2. *Let $\|\cdot\|_V$ be a norm corresponding to a scalar product $(\cdot, \cdot)_V$ defined on $V \times V$. Then the bilinear form $a(\cdot, \cdot)$ is called* coercive *(* V-elliptic*), and $a(\cdot, \cdot)$ and $L(\cdot)$ are continuous, if there are constants c_1, c_2 and c_2 such that:*

$$a(v, v) \geq c_1 \|v\|_V^2, \quad \forall v \in V \qquad (\textit{coercivity}) \tag{9.1.12}$$

$$|a(u, v)| \leq c_2 \|u\|_V \|v\|_V, \quad \forall u, v \in V \qquad (a \textit{ is continuous}) \tag{9.1.13}$$

$$|L(v)| \leq c_3 \|v\|_V, \quad \forall v \in V \qquad (L \textit{ is continuous}). \tag{9.1.14}$$

Note. Since L is linear, we have using the relation (9.1.14) above that

$$|L(u) - L(v)| = |L(u - v)| \leq c_3 \|u - v\|_V,$$

which shows that $L(u) \Longrightarrow L(v)$ as $u \Longrightarrow v$, in V. Thus L is continuous. Similarly the relation $|a(u, v)| \leq c_1 \|u\|_V \|v\|_V$ implies that the bilinear form $a(\cdot, \cdot)$ is continuous in each component.

Definition 9.3. *The energy norm on V is defined by $\|v\|_a = \sqrt{a(v, v)}$, $v \in V$.*

Recalling the relations (9.1.12) and (9.1.13) above, the energy norm satisfies

$$c_1 \|v\|_V^2 \leq a(v, v) = \|v\|_a^2 \leq c_2 \|v\|_V^2. \tag{9.1.15}$$

Hence, the energy norm $\|v\|_a$ is equivalent to the abstract $\|v\|_V$ norm.

Example 9.3. *For the scalar product*

$$(u,v)=\int_0^1 u'(x)v'(x)dx, \quad in \quad \mathcal{H}_0^1, \tag{9.1.16}$$

and the norm

$$\|u\|=\sqrt{(u,u)}, \tag{9.1.17}$$

the relations (9.1.12) *and* (9.1.13) *are valid with* $c_1=c_2\equiv 1$ *: More closely we have in this case that*

(i): $(v,v)=\|v\|^2$ *is an identity, and*

(ii): $|(u,v)|\le\|u\|\|v\|$ *is the Cauchy's inequality sketched below:*

Proof of the Cauchy's inequality. Using the obvious inequality $2ab\le a^2+b^2$, we have

$$2|(u,w)|\le\|u\|^2+\|w\|^2. \tag{9.1.18}$$

We let $w=(u,v)\cdot v/\|v\|^2$, then

$$2|(u,w)|=2\Big|\Big(u,(u,v)\frac{v}{\|v\|^2}\Big)\Big|\le\|u\|^2+|(u,v)|^2\frac{\|v\|^2}{\|v\|^4} \tag{9.1.19}$$

Thus

$$2\frac{|(u,v)|^2}{\|v\|^2}\le\|u\|^2+|(u,v)|^2\frac{\|v\|^2}{\|v\|^4}, \tag{9.1.20}$$

which multiplying by $\|v\|^2$, gives

$$2|(u,v)|^2\le\|u\|^2\cdot\|v\|^2+|(u,v)|^2, \tag{9.1.21}$$

and hence

$$|(u,v)|^2\le\|u\|^2\cdot\|v\|^2, \tag{9.1.22}$$

and the proof is complete. □

Definition 9.4. *A* Hilbert space *is a complete linear space with a scalar product.*

To define complete linear space we first need to define a *Cauchy sequence* of *real* or *complex* numbers.

Definition 9.5. *A sequence $\{z_k\}_{k=1}^{\infty}$ is a Cauchy sequence if for every $\varepsilon > 0$, there is an integer $N > 0$, such that $m, n > N \Rightarrow |z_m - z_n| < \varepsilon$.*

Now we state, without proof, a classical theorem of analysis:

Theorem 9.1. *Every Chaucy sequence in $\mathbb{C}$ is convergent. More precisely: If $\{z_k\}_{k=1}^{\infty} \subset \mathbb{C}$ is a Cauchy sequence, then there is a $z \in \mathbb{C}$, such that for every $\epsilon > 0$, there is an integer $M > 0$, such that $m \geq M \Rightarrow |z_m - z| < \varepsilon$.*

Definition 9.6. *A linear space V (vector space) with the norm $\|\cdot\|$ is called complete if every Cauchy sequence in V is convergent. In other words: For every $\{v_k\}_{k=1}^{\infty}$ with the property that for every $\varepsilon > 0$ there is an integer $N > 0$, such that $m, n > N \Rightarrow \|v_m - v_n\| < \varepsilon$, (i.e. for every Cauchy sequence) there is a $v \in V$ such that for every $\varepsilon > 0$ there is an integer $M > 0$ such that $m \geq M \Rightarrow |v_m - v| < \varepsilon$.*

Theorem 9.2. *$\mathcal{H}_0^1 = \{f : [0,1] \to \mathbb{R} : \int_0^1 f'(x)^2 dx < \infty, \wedge f(0) = f(1) = 0\}$ is a complete Hilbert space with the norm*

$$\|u\| = \sqrt{(u,u)} = \Big(\int_0^1 u'(x)^2 dx \Big)^{1/2}. \tag{9.1.23}$$

Lemma 9.1 (Poincare's inequality in 1D)**.** *If $u(0) = u(L) = 0$ then*

$$\int_0^L u(x)^2 dx \leq C_L \int_0^L u'(x)^2 dx, \tag{9.1.24}$$

where C_L is a constant independent of $u(x)$ but depends on L.

Proof. Using the Cauchy-Schwarz inequality we have

$$\begin{aligned} u(x) &= \int_0^x u'(y) dy \leq \int_0^x |u'(y)| dy \leq \int_0^L |u'(y)| \cdot 1 dy \\ &\leq \Big(\int_0^L u'(y)^2 dy \Big)^{1/2} \Big(\int_0^L 1^2 dy \Big)^{1/2} = \sqrt{L} \Big(\int_0^L u'(y)^2 dy \Big)^{1/2}. \end{aligned} \tag{9.1.25}$$

Consequently

$$u(x)^2 \leq L \int_0^L u'(y)^2 dy, \tag{9.1.26}$$

and hence

$$\int_0^L u(x)^2 dx \le L \int_0^L \Big(\int_0^L u'(y)^2 dy \Big) dx = L^2 \int_0^L u'(x)^2 dx, \tag{9.1.27}$$

i.e. $C_L = L$. Thus Poincare inequality deteriorates in unbounded domains. □

Definition 9.7. *We define a functional ℓ as a mapping from a (linear) function space V into $\mathbb{R}$, i.e.,*

$$\ell : V \to \mathbb{R}. \tag{9.1.28}$$

- *A funcitonal ℓ is called <u>linear</u> if*

$$\begin{cases} \ell(u+v) = \ell(u) + \ell(v) & \text{for all } \ u, v \in V \\ \ell(\alpha u) = \alpha \cdot \ell(u) & \text{for all } \ u \in V \text{ and } \ \alpha \in \mathbb{R}. \end{cases} \tag{9.1.29}$$

- *A functional is called <u>bounded</u> if there is a constant C such that*

$$|\ell(u)| \le C \cdot \|u\| \quad \text{for all } u \in V \quad (C \text{ is independnet of } u)$$

Example 9.4. *If $f \in L^2(0,1)$, i.e. $\int_0^1 f(x)^2 dx$ is bounded, then*

$$\ell(v) = \int_0^1 u(x)v(x)dx \tag{9.1.30}$$

is a bounded linear functional.

Problem 9.1. *Show that ℓ, defined in example above is linear.*

Problem 9.2. *Prove using Cauchy's and Poincare's inequalities that ℓ, defined as in the above example , is bounded in $\mathcal{H}_0^1$.*

9.2 Riesz and Lax-Milgram Theorems

Abstract formulations: Recalling that

$$(u,v) = \int_0^1 u'(x)v'(x)dx \qquad \text{and} \quad \ell(v) = \int_0^1 u(x)v(x)dx,$$

we may redefine our variational formulation (VF) and minimization problem (MP) in an abstract forn as (V) and (M), respectively:

(V) Find $u \in \mathcal{H}_0^1$, such that $(u, v) = \ell(v)$ for all $v \in \mathcal{H}_0^1$.

(M) Find $u \in \mathcal{H}_0^1$, such that $F(u) = \min_{v \in \mathcal{H}_0^1} F(v)$ with $F(v) = \frac{1}{2}\|v\|^2 - \ell(v)$.

Theorem 9.3. *There exists a unique solution for the, equivalent, problems (V) and (M).*

Proof. That (V) and (M) are equvalent is trivial and shown as in part I. Now, we note that there exists a real number σ such that $F(v) > \sigma$ for all $v \in \mathcal{H}_0^1$, (otherwise it is not possible to minimize F): namely we can write

$$F(v) = \frac{1}{2}\|v\|^2 - \ell(v) \geq \frac{1}{2}\|v\|^2 - \gamma\|v\|, \tag{9.2.1}$$

where γ is the constant bounding ℓ, i.e. $|\ell(v)| \leq \gamma\|v\|$. But since

$$0 \leq \frac{1}{2}(\|v\| - \gamma)^2 = \frac{1}{2}\|v\|^2 - \gamma\|v\| + \frac{1}{2}\gamma^2, \tag{9.2.2}$$

thus evidently we have

$$F(v) \geq \frac{1}{2}\|v\|^2 - \gamma\|v\| \geq -\frac{1}{2}\gamma^2. \tag{9.2.3}$$

Let now σ^* be the largest real number σ such that

$$F(v) > \sigma \quad \text{for all } v \in \mathcal{H}_0^1. \tag{9.2.4}$$

Take now a sequence of functions $\{u_k\}_{k=0}^{\infty}$, such that

$$F(u_k) \longrightarrow \sigma^*. \tag{9.2.5}$$

To show that there *exists* a *unique* solution for (V) and (M) we shall use the following two fundamental results:

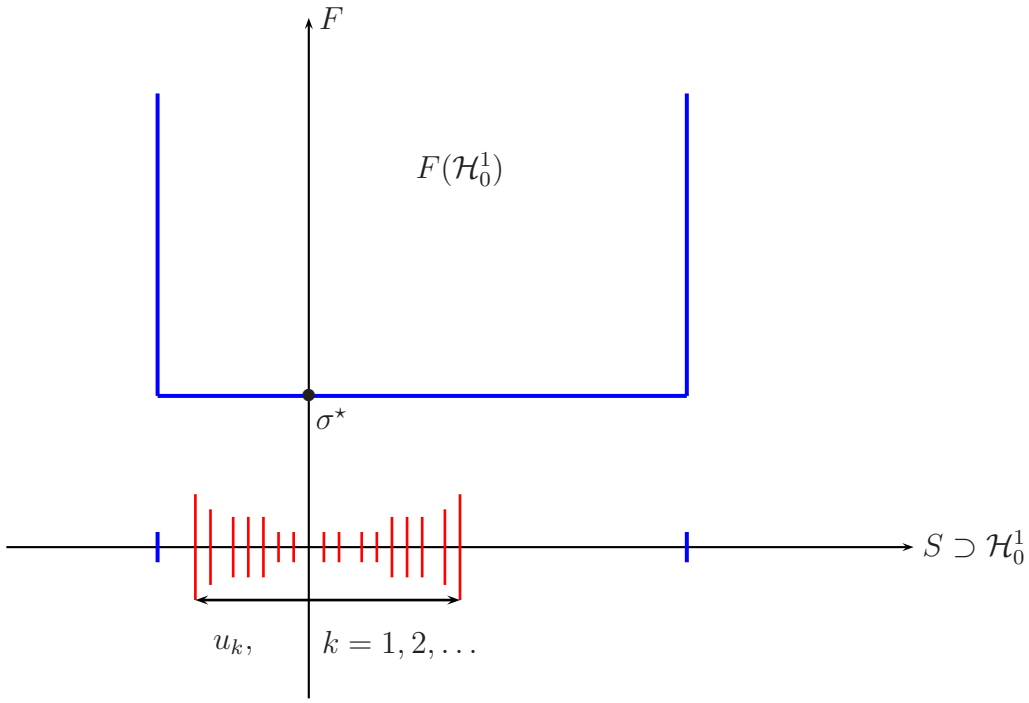


Figure 9.1: The axiom of choice for existence of a solution of (V) and (M).

(i) It is always possible to find a sequence $\{u_k\}_{k=0}^{\infty}$, such that $F(u_k) \to \sigma^\bullet$ (because $\mathbb{R}$ is complete.)

(ii) The parallelogram law (elementary linear algebra).

$$\|a+b\|^2 + \|a-b\|^2 = 2\|a\|^2 + 2\|b\|^2.$$

Using (ii) and the linearity of ℓ we can write

$$\begin{aligned}\|u_k - u_j\|^2 &= 2\|u_k\|^2 + 2\|u_j\|^2 - \|u_k + u_j\|^2 - 4\ell(u_k) - 4\ell(u_j) + 4\ell(u+v)\\ &= 2\|u_k\|^2 - 4\ell(u_k) + 2\|u_j\|^2 - 4\ell(u_j) - \|u_k + u_j\|^2 + 4\ell(u_k + u_j)\\ &= 4F(u_k) + 4F(u_j) - 8F\Big(\frac{u_k+u_j}{2}\Big),\end{aligned}$$

where we have used the definition of $F(v) = \frac{1}{2}\|v\|^2 - \ell(v)$ with $v = u_k,\ u_j,$ and $v = (u_k + u_j)/2$, respectivey. In particular by linearity of ℓ:

$$-\|u_k+u_j\|^2 + 4\ell(u_k+u_j) = -4\Big\|\frac{u_k+u_j}{2}\Big\|^2 + 8\ell\Big(\frac{u_k+u_j}{2}\Big) = -8F\Big(\frac{u_k+u_j}{2}\Big).$$

Now since $F(u_k) \to \sigma^*$ and $F(u_j) \to \sigma^*$, then

$$\|u_k - u_j\|^2 \leq 4F(u_k) + 4F(u_j) - 8\sigma^* \to 0, \quad \text{as} \quad k, j \to \infty.$$

Thus we have shown that $\{u_k\}_{k=0}^{\infty}$ is a Cauchy sequence. Since $\{u_k\} \subset \mathcal{H}_0^1$ and $\mathcal{H}_0^1$ is complete thus $\{u_k\}_{k=1}^{\infty}$ is a convergent sequence. Hence

$$\exists u \in \mathcal{H}_0^1, \quad \text{such that} \quad u = \lim_{k\to\infty} u_k.$$

By the continuity of F we get that

$$\lim_{k\to\infty} F(u_k) = F(u). \tag{9.2.6}$$

Now (9.2.5) and (9.2.6) yield $F(u) = \sigma^*$ and by (9.2.4) and the definition of σ^* we end up with

$$F(u) < F(v), \qquad \forall v \in \mathcal{H}_0^1. \tag{9.2.7}$$

This in our minimization problem (M). And since (M) $\Leftrightarrow$ (V) we conclude that:

there is a unique $u \in \mathcal{H}_0^1$ *, such that* $\ell(v) = (u, v) \quad \forall v \in \mathcal{H}_0^1$. □

Summing up we have proved that:

Proposition 9.1. *Every bounded linear functional can be represented as a scalar product with a given function u. This u is the unique solution for both (V) and (M).*

Theorem 9.4 (Riesz representation theorem)**.** *If V is a Hilbert space with the scalar product (u, v) and norm $\|u\| = \sqrt{(u, u)}$, and $\ell(v)$ is a bounded linear functional on V, then there is a unique $u \in V$, such that $\ell(v) = (u, v), \quad \forall v \in V$.*

Theorem 9.5 (Lax-Milgram theorem)**.** *(A general version of Riesz theorem) Assume that $\ell(v)$ is a bounded linear functional on V and $a(u, v)$ is bilinear bounded and elliptic in V, then there is a unique $u \in V$, such that*

$$a(u, v) = \ell(v), \quad \forall v \in V. \tag{9.2.8}$$

Remark 9.1. Bilinear *means that $a(u,v)$ satisfies the same properties as a scalar product, however it need not! to be* symmetric.

Bounded *means:*

$$|a(u,v) \leq \beta \|u\| \, \|v\|, \quad \textit{for some constant } \beta > 0. \tag{9.2.9}$$

Elliptic *means:*

$$a(v,v) \geq \alpha \|v\|^2, \quad \textit{for some } \alpha > 0. \tag{9.2.10}$$

Note

$$\textit{If } a(u,v) = (u,v), \textit{ then } \alpha = \beta = 1.$$

9.3 Exercises

Problem 9.3. *Verify that the assumptions of the Lax-Milgram theorem are satisfied for the following problems with appropriate assumptions on α and f.*

(I)
$$\begin{cases} -u'' + \alpha u = f, & \textit{in } (0,1), \\ u(0) = u'(1) = 0, & \alpha = 0 \textit{ and } 1. \end{cases}$$

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

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

Problem 9.4. *Let Ω be a bounded domain in $\mathbb{R}^d$ with boundary Γ, show that there is a constant C such that for all $v \in H^1(\Omega)$,*

$$\|v\|_{L_2(\Gamma)} \leq C \|v\|_{H^1(\Omega)}, \tag{9.3.1}$$

where $\|v\|^2_{H^1(\Gamma)} = \|v\|^2 + \|\nabla v\|^2$. Hint: Use the following Green's formula

$$\int_\Omega v^2 \Delta \varphi = \int_\Gamma v^2 \partial_n \varphi - \int_\Omega 2v \nabla v \cdot \nabla \varphi, \tag{9.3.2}$$

with $\partial_n \varphi = 1$. (9.3.1) is knowm as trace inequality, or trace theorem.

Problem 9.5. *Let u be the solution of the following Neumann problem:*

$$\begin{cases} -\Delta u = f, & \text{in } \Omega \subset \mathbb{R}^d, \\ -\partial_n u = ku, & \text{on } \Gamma = \partial\Omega, . \end{cases}$$

where $\partial_n u = n \cdot \nabla u$ with n being outward unit normal to Γ and $k \geq 0$. a) Show the stability estimate

$$\|u\|_\Omega \leq C_\Omega(\|u\|_\Gamma + \|\nabla u\|_\Omega).$$

b) Use the estimate in a) to show that $\|u\|_\Gamma \to 0$ as $k \to \infty$.

Problem 9.6. *Using the trace inequality, show that the solution for the problem*

$$\begin{cases} -\Delta u + u = 0, & \text{in } \Omega \\ \partial_n u = g, & \text{on } \Gamma, \end{cases}$$

satisfies the inequality

$$\|v\|^2 + \|\nabla v\|^2 \leq C\|g\|^2_{L_2(\Gamma)}.$$

Problem 9.7. *Consider the boundary value problem*

$$\begin{cases} \Delta u = 0, & \text{in } \Omega \subset \mathbb{R}^2, \\ \partial_n u + u = g, & \text{on } \Gamma = \partial\Omega, \quad n \text{ is outward unit normal to } \Gamma. \end{cases}$$

a) Show the stability estimate

$$||\nabla u||^2_{L_2(\Omega)} + \frac{1}{2}||u||^2_{L_2(\Gamma)} \leq \frac{1}{2}||g||^2_{L_2(\Gamma)}.$$

b) Discuss, concisely, the conditions for applying the Lax-Milgram theorem to this problem.

Chapter 10

The Poisson Equation

In this chapter we shall extend the study in Chapter 4 in Part I to solve the Poisson equation

$$\begin{cases} -\Delta u = f, & \text{in } \Omega \in \mathbb{R}^d, \quad d = 2, 3 \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{10.0.1}$$

where Ω is a bounded domain in $\mathbb{R}^d$, with $d = 2$ or $d = 3$, with polygonal boundary $\Gamma = \partial\Omega$. For the presentation of problems from science and industry that are modeled by the Poisson's equation we refer to Eriksson et al. Computational Differential Equations [] and Folland: An introduction to Fourier Analysis and its Applications []. Below we shall prove stability results and derive a priori and a posteriori error estimates for the problem (10.0.1)

10.1 Stability

To derive stability estimates for (10.0.1) we shall assume an underlying general vector space V (to be specified below) of functions. We multiply the equation by u and integrate over Ω to obtain

$$-\int_\Omega (\Delta u) u dx = \int_\Omega f\ u dx, \quad x \in \Omega \quad \text{and } u \in V. \tag{10.1.1}$$

Using Green's formula and the boundary condition: $u = 0$ on Γ, we get that

$$\|\nabla u\|^2 = \int_\Omega fu \leq \|f\| \, \|u\|, \tag{10.1.2}$$

where $\|\cdot\|$ denotes the usual $L_2(\Omega)$-norm.

Lemma 10.1 (Poincaré inequality; the 2D-version)**.** *For the solution u of the problem* (10.0.1) *in a bounded domain $\Omega \in \mathbb{R}^2$, There exisists a constant C_Ω, independet of u such that*

$$\|u\| \leq C_\Omega \|\nabla u\| \tag{10.1.3}$$

Proof. Let φ be a function such that $\Delta\varphi = 1$ in Ω, and $2|\nabla\varphi| \leq C_\Omega$ in Ω, (it is easy to construct such a functionφ), then again by the use of Green's formula and the boundary condition we get

$$\|u\|^2 = \int_\Omega u^2 \Delta\varphi = -\int_\Omega 2u(\nabla u \cdot \nabla\varphi) \leq C_\Omega \|u\| \, \|\nabla u\|. \tag{10.1.4}$$

Thus

$$\|u\| \leq C_\Omega \|\nabla u\|. \tag{10.1.5}$$

Now combining with the inequality (10.1.2) we get that the following *weak stability estimate* holds

$$\|\nabla u\| \leq C_\Omega \|f\|. \tag{10.1.6}$$

□

Problem 10.1. *Derive corresponding estimates for following Neumann problem:*

$$\begin{cases} -\Delta u + u = f, & \text{in } \Omega \\ \frac{\partial u}{\partial n} = 0, & \text{on } \Gamma = \partial\Omega. \end{cases} \tag{10.1.7}$$

10.2 Error Estimates for FEM

We start with the *variational formulation* for the problem (10.0.1), through multiplying the equation by a test function, integrating over Ω and using the

Green's formula: Find a solution $u(x)$ such that $u(x) = 0$ on $\Gamma = \partial\Omega$ and

$$(VF): \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx, \quad \forall v \text{ such that } v = 0 \text{ on } \Gamma. \quad (10.2.1)$$

We prepare for a finite element method where we shall approximate the exact solution $u(x)$ by a suitable discrete solution $U(x)$. To this approach let $\mathcal{T} = \{K : \cup K = \Omega\}$ be a triangulation of the domain Ω and $\varphi_j, j = 1, 2, \ldots, n$ be the corresponding basis functions, such that $\varphi_j(x)$ is continuous, linear in x on each K and

$$\varphi_j(N_i) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \quad (10.2.2)$$

where $N_1, N_2, \ldots, N_n$ are the inner nodes in the triangulation.

Now we set the approximate solution $U(x)$ to be a linear combination of the basis functions $\varphi_j,\ j = 1, \ldots, n$:

$$U(x) = U_1\varphi_1(x) + U_2\varphi_2(x) + \ldots + U_n\varphi_n(x), \quad (10.2.3)$$

and seek the coefficients $U_j = U(N_j)$, i.e., the nodal values of $U(x)$, at the nodes $N_j,\ 1 \leq i \leq n$, so that

$$(FEM) \qquad \int_\Omega \nabla U \cdot \nabla \varphi_i \, dx = \int_\Omega f \cdot \varphi_i \, dx, \quad i = 1, 2, \ldots n, \quad (10.2.4)$$

or equivalently

$$(V_h^0) \qquad \int_\Omega \nabla U \cdot \nabla v \, dx = \int_\Omega f \cdot v \, dx, \quad \forall v \in V_h^0. \quad (10.2.5)$$

We recall that

$V_h^0 = \{v(x) : v \text{ is continuous, piecewise linear}(\text{on } \mathcal{T}),\ \text{and } v = 0 \text{ on } \Gamma = \partial\Omega\}$.

Note that every $v \in V_h^0$ can be represented by

$$v(x) = v(N_1)\varphi_1(x) + v(N_2)\varphi_2(x) + \ldots + v(N_n)\varphi_n(x). \quad (10.2.6)$$

Theorem 10.1 (a priori error estimate for the gradient $\nabla u - \nabla U$)**.** *Let $e = u - U$ represent the error in the above piecelinear, continuous finite element estimate approximation of the solution for* (10.0.1)*, let $\nabla e = \nabla u - \nabla U = \nabla(u - U)$. Then we have the following estimate for the gradient of the error*

$$\|\nabla e\| = \|\nabla(u - U) \leq C\|h\, D^2 u\|. \quad (10.2.7)$$

Proof. For the error $e = u - U$ we have $\nabla e = \nabla u - \nabla U = \nabla(u - U)$. Subtracting (10.2.5) from the (10.2.1) we obtain the *Galerkin Orthogonality:*

$$\int_\Omega (\nabla u - \nabla U)\nabla v \, dx = \int_\Omega \nabla e \cdot \nabla v \, dx = 0, \qquad \forall v \in V_h^0. \tag{10.2.8}$$

Further we may write

$$\|\nabla e\|^2 = \int_\Omega \nabla e \cdot \nabla e \, dx = \int_\Omega \nabla e \cdot \nabla (u - U) \, dx = \int_\Omega \nabla e \cdot \nabla u \, dx - \int_\Omega \nabla e \cdot \nabla U \, dx.$$

Now using the Galerkin orthogonality (10.2.8), since $U(x) \in V_h^0$ we have the last integral above: $\int_\Omega \nabla e \cdot \nabla U \, dx = 0$. Hence removing the vanishing ∇U-term and inserting $\int_\Omega \nabla e \cdot \nabla v \, dx = 0, \ \forall v \in V_h^0$ we have that

$$\|\nabla e\|^2 = \int_\Omega \nabla e \cdot \nabla u dx - \int_\Omega \nabla e \cdot \nabla v dx = \int_\Omega \nabla e \cdot \nabla (u - v) dx \leq \|\nabla e\| \cdot \|\nabla (u - v)\|.$$

Thus

$$\|\nabla (u - U)\| \leq \|\nabla (u - v)\|, \quad \forall v \in V_h^0, \tag{10.2.9}$$

that is, measuring in the L_2-norm the finite element solution U is closer to u than any other v in V_h^0.

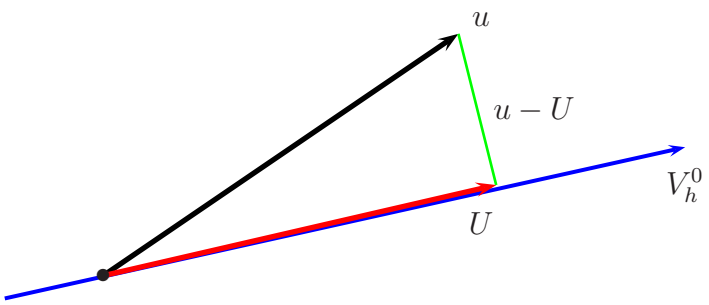


Figure 10.1: The orthogonal (L_2) projection of u on V_h^0.

In other words the error $u - U$ is orthogonal to V_h^0.

It is possible to show that there is a $v \in V_h^0$ (an interpolant), such that

$$\|\nabla(u - v)\| \leq C\|h\, D^2 u\|, \tag{10.2.10}$$

where $h = h(x) = \text{diam}(K)$ for $x \in K$ and C is a constant, independent of h. This is the case, for example, if v interpolates u at the nodes N_i

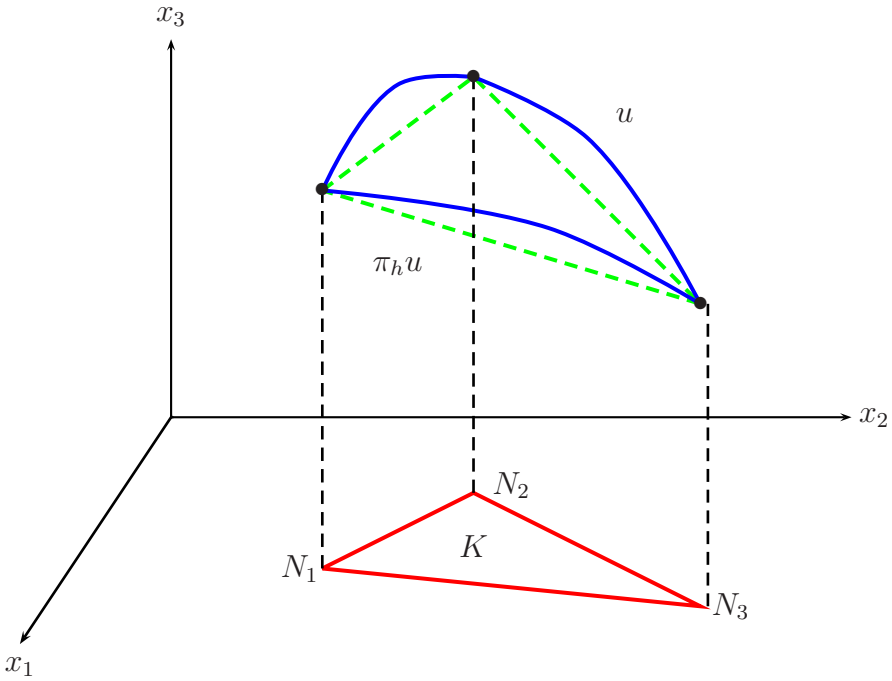


Figure 10.2: The nodal interpolant of u in 2D case

Combining (10.2.9) and (10.2.10) we get

$$\|\nabla e\| = \|\nabla(u - U) \leq C\|h\, D^2 u\|, \tag{10.2.11}$$

which is indicating that the error is small if $h(x)$ is sufficiently small depending on $D^2 u$. See the Fig. below

□

To prove an a priori error estimate for the solution we shall use the following result:

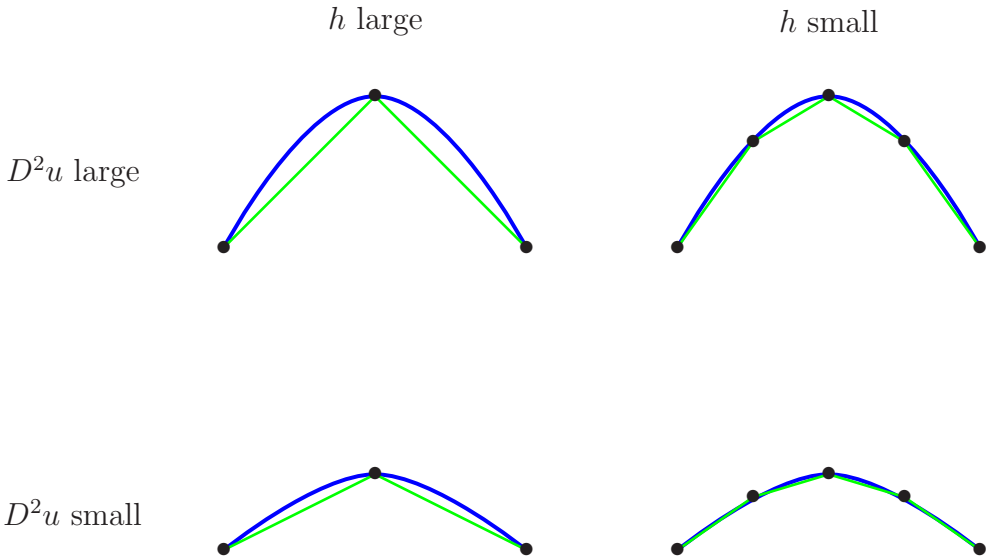


Figure 10.3: The adaptivity priciple: to refine mesh for large D^2u

Lemma 10.2 (regularity lemma)**.** *Assume that Ω has no re-intrents. We have for $u \in H^2(\Omega)$; with $u = 0$ or $(\frac{\partial u}{\partial n} = 0)$ on $\partial\Omega$. that*

$$\|D^2u\| \leq c_\Omega \cdot \|\Delta u\|, \tag{10.2.12}$$

where

$$D^2u = (u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2)^{1/2}. \tag{10.2.13}$$

We postpone the proof of this lemma and first derive the error estimate:

Theorem 10.2 (a priori error estimate for the solution $e = u - U$)**.** *For a general mesh we have the following a priori error estimate for the solution of the Poisson equation* (10.0.1)*:*

$$\|e\| = \|u - U\| \leq C^2\, C_\Omega\, (\max_\Omega h) \cdot \|h\, D^2u\|. \tag{10.2.14}$$

Proof. Let φ be the solution of the *dual problem*

$$\begin{cases} -\Delta\varphi = e, & \text{in } \Omega \\ \varphi = 0, & \text{on } \partial\Omega \end{cases} \tag{10.2.15}$$

Then we have using Green's formula

$$\begin{aligned}\|e\|^2 &= \int_\Omega e(-\Delta\varphi)dx = \int_\Omega \nabla e\cdot\nabla\varphi\ dx, \int_\Omega \nabla e\cdot\nabla(\varphi - v)\,dx \\ &\le \|\nabla e\|\cdot\|\nabla(\varphi - v)\|, \quad \forall v \in V_h^0,\end{aligned} \tag{10.2.16}$$

where in the last equality we have used the Galerkin orthogonality. We now choose v such that

$$\|\nabla(\varphi - v)e\| \le C\|h\cdot D^2\varphi C\| \le C(\max_\Omega h)\|h\cdot D^2\varphi C\|. \tag{10.2.17}$$

Applying the lemma to φ, we get

$$\|D^2\varphi\| \le C_\Omega\cdot\|\Delta\varphi\| = C_\Omega\|e\|. \tag{10.2.18}$$

Now (10.2.11)-(10.2.18) implies that

$$\begin{aligned}\|e\|^2 &\le \|\nabla e\|\cdot\|\nabla(\varphi - v)\| \le \|\nabla e\|\cdot C\max_\Omega h\,\|D^2\varphi\| \\ &\le \|\nabla e\|\cdot C\max_\Omega h C_\Omega\|e\| \le C^2\,C_\Omega\max_\Omega h\|e\|\|h\,D^2u\|.\end{aligned} \tag{10.2.19}$$

Thus we have obtained the desired result: *a priori error estimate:*

$$\|e\| = \|u - U\| \le C^2\,C_\Omega\,(\max_\Omega h)\cdot\|h\,D^2u\|. \tag{10.2.20}$$

□

Corollary 10.1 (strong stability estimate)**.** *Using the Lemma, for a uniform (constant h), the a priori error estimate* (10.2.20) *can be written as an stability estimate viz,*

$$\|u - U\| \le C^2\,C_\Omega^2\,(\max_\Omega h)^2\,\|f\|. \tag{10.2.21}$$

Theorem 10.3 (a posteriori error estimate)**.** *For the solution of the Poisson equation* (10.0.1) *we have that*

$$\|u - U\| \leq C \, \|h^2 r\|, \tag{10.2.22}$$

where U is the continuous piecewise linear finite element approximation and $r = f + \Delta_n U$ is the residual with Δ_n being discrete Laplacian defined by

$$(\Delta_n U, v) = \sum_{K \in \mathcal{T}_h} (\nabla U, \nabla v)_K. \tag{10.2.23}$$

Proof. We consider the following dual problem

$$\begin{cases} -\Delta\varphi(x) = e(x), & x \in \Omega, \\ \varphi(x) = 0, \quad x \in \partial\Omega, & e(x) = u(x) - U(x). \end{cases} \tag{10.2.24}$$

Thus $e(x) = 0$, $\forall x \in \partial\Omega$. Using (10.2.24) and the Green's formula, the L_2-norm of the error can be written as:

$$\|e\|^2 = \int_\Omega e \cdot e \, dx = \int_\Omega e(-\Delta\varphi) dx = \int_\Omega \nabla e \cdot \nabla\varphi \, dx. \tag{10.2.25}$$

Thus by the Galerkin orthogonality: $\int_\Omega \nabla e \cdot \nabla v dx = 0$, $\forall v \in V_h^0$, and the boundary data: $\varphi(x)$, $\forall x \in \partial\Omega$ we can write

$$\begin{aligned} \|e\|^2 &= \int_\Omega \nabla e \cdot \nabla\varphi \, dx - \int_\Omega \nabla e \cdot \nabla v dx = \int_\Omega \nabla e \cdot \nabla(\varphi - v) \, dx \\ &= \int_\Omega (-\Delta e)(\varphi - v) \, dx \leq \|h^2 r\| \cdot \|h^{-2}(\varphi - v)\| \\ &\leq C \cdot \|h^2 r\| \cdot \|\Delta\varphi\| \leq C \cdot \|h^2 r\| \cdot \|e\|, \end{aligned} \tag{10.2.26}$$

where we use the fact that the $-\Delta e = -\Delta u + \Delta U = f + \Delta U$ is the residual r and v is an interpolant of φ. Thus, for this problem, the final *a posteriori error estimate* is:

$$\|u - U\| \leq C \, \|h^2 r\|. \tag{10.2.27}$$

Observe that for piecewise linear approximations $\Delta U = 0$ on each element K and hence $r \equiv f$ and our a posteriori error estimate above can be viewed as a *strong stability estimate* viz,

$$\|e\| \leq C \, \|h^2 f\|. \tag{10.2.28}$$

Note that now is $\nabla e(\varphi - v) \neq 0$ on the enter-element boundaries. □

Problem 10.2. *Show that* $\|(u-U)'\| \leq C\|hr\|$.

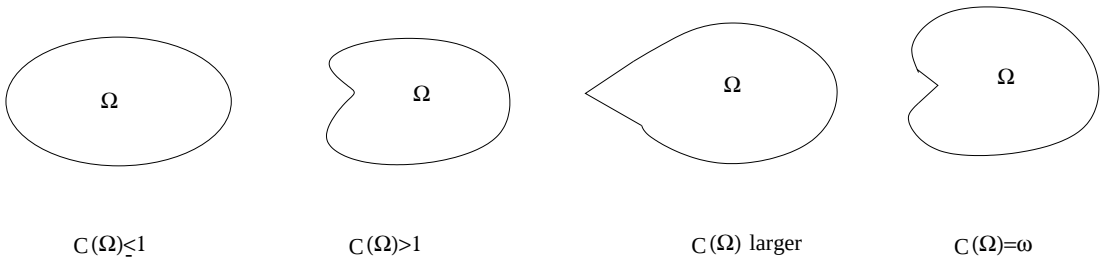
Problem 10.3. *Verify that for* v *being the interpolant of* φ*, we have*

$$\|e\| \leq C\,\|h^2 f\| \times \begin{cases} \|h^{-2}\,(\varphi - v)\| & \leq C\,\|\Delta\varphi\|, \quad \textit{and} \\ \|h^{-1}(\varphi - v)\| & \leq C\,\|\nabla\varphi\|. \end{cases} \tag{10.2.29}$$

Problem 10.4. *Derive the corresponding estimate to (10.2.27) in the 1-dimensional case (*$d = 1$*).*

Now we return to the proof of Lemma:

proof of regularity lemma. First note that for convex Ω, the constant $C_\Omega \leq 1$ in lemma, otherwise the constant $C_\Omega > 1$ and increases from left to right for the Ω:s below.



Let now Ω be a rectangular domain and set $u = 0$ on $\partial\Omega$. We have then

$$\|\Delta u\|^2 = \int_\Omega (u_{xx} + u_{yy})^2 dxdy = \int_\Omega (u_{xx}^2 + 2u_{xx}u_{yy} + u_{yy}^2)\,dxdy. \tag{10.2.30}$$

Further applying Green's formula:

$$\int_\Omega (\Delta u)v\,dx = \int_\Gamma (\nabla u \cdot n)v\,ds - \int_\Omega \nabla u \cdot \nabla v\,dx$$

to our rectangular domain Ω we have

$$\int_\Omega u_{xx}u_{yy}dxdy = \int_{\partial\Omega} u_x(u_{yy}\cdot n_x)ds - \int_\Omega u_x \underbrace{u_{yyx}}_{=u_{xyy}} dxdy \tag{10.2.31}$$

using Green's formula once again (with "$v = u_x$", "$\Delta u = u_{xyy}$") we get

$$\int_\Omega u_x u_{xyy} dxdy = \int_{\partial\Omega} u_x (u_{yx} \cdot n_y) ds - \int_\Omega u_{xy} u_{xy}\, dxdy, \tag{10.2.32}$$

which inserting in (10.2.31) gives that

$$\int_\Omega u_{xx} u_{yy}\, dxdy = \int_{\partial\Omega} (u_x u_{yy} n_x - u_x u_{yx} n_y) ds + \int_\Omega u_{xy} u_{xy}\, dxdy. \tag{10.2.33}$$

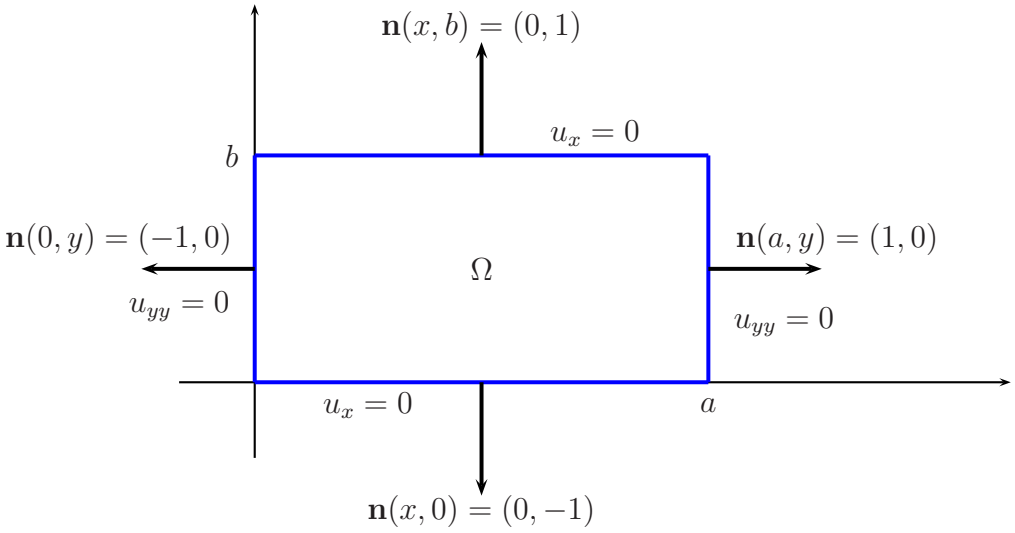


Figure 10.4: A rectangular domain Ω with its outward unit normals

Now, as we can see from the figure that $(u_x u_{yy} n_x - u_x u_{yx} n_y) = 0$, on $\partial\Omega$ and hence we have

$$\int_\Omega u_{xx} u_{yy} dxdy = \int_\Omega u_{xy} u_{xy} dxdy = \int_\Omega u_{xy}^2\, dxdy. \tag{10.2.34}$$

Thus, in this case,

$$\|\Delta u\|^2 = \int_\Omega (u_{xx} + u_{yy})^2 dxdy = \int_\Omega (u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2) dxdy = \|D^2 u\|^2,$$

and the proof is complete by a constant $\equiv 1$.

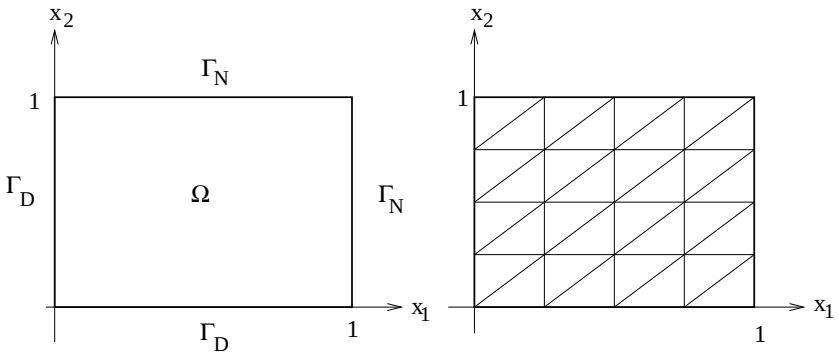
□

10.3 Exercises

Problem 10.5. *Consider the following two dimensional problem:*

$$\begin{cases} -\Delta u = 1, & \text{in } \Omega \\ u = 0, & \text{on } \Gamma_D \\ \frac{\partial u}{\partial n} = 0, & \text{on } \Gamma_N \end{cases} \tag{10.3.1}$$

See figure below

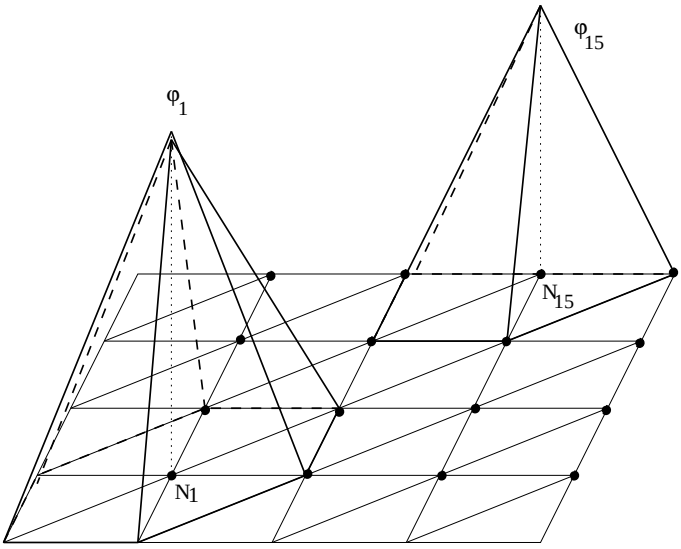


Triangulate Ω as in the figure and let

$$U(x) = U_1\varphi_1(x) + \ldots + U_{16}\varphi_{16}(x),$$

where $x = (x_1, x_2)$ and φ_j, $j = 1, \ldots 16$ are the basis functions, see Fig. below, and determine $U_1, \ldots U_{16}$ so that

$$\int_\Omega \nabla U \cdot \nabla \varphi_j dx = \int_\Omega \varphi_j dx, \quad j = 1, 2, \ldots, 16.$$



Problem 10.6. *Generalize the procedure in the previous problem to the following case*

$$\begin{cases} -\nabla(a\nabla u) = f, & \text{in } \Omega \\ u = 0, & \text{on } \Gamma_D \\ a\frac{\partial u}{\partial n} = 7, & \text{on } \Gamma_N \end{cases} \text{, where } \begin{cases} a = 1 \quad \text{for } x_1 < \frac{1}{2} \\ a = 2 \quad \text{for } x_1 > \frac{1}{2} \\ f = x_2. \quad \text{mesh-size=}h. \end{cases}$$

Problem 10.7. *Consider the Dirichlet problem*

$$-\nabla \cdot (a(x)\nabla u) = f(x), \quad x \in \Omega \subset \mathbb{R}^2, \qquad u = 0, \text{ for } x \in \partial\Omega.$$

Assume that c_0 and c_1 are constants such that $c_0 \leq a(x) \leq c_1$, $\forall x \in \Omega$ and let $U = \sum_{j=1}^{N} \alpha_j w_j(x)$ be a Galerkin approximation of u in a finite dimensional subspace M of $H_0^1(\Omega)$. Prove the a priori error estimate

$$||u - U||_{H_0^1(\Omega)} \leq C \inf_{\chi \in M} ||u - \chi||_{H_0^1(\Omega)}.$$

Problem 10.8. *Consider the following Schrödinger equation*

$$\dot{u} + i\Delta u = 0, \quad \text{in } \Omega, \qquad u = 0, \quad \text{on } \partial\Omega,$$

where $i = \sqrt{-1}$ and $u = u_1 + iu_2$. a) Show that the the L_2 norm of the solution, i.e., $\int_\Omega |u|^2$ is time independent.

Hint: Multiply the equation by $\bar{u} = u_1 - iu_2$*, integrate over* Ω *and consider the real part.*

b) Consider the corresponding eigenvalue problem*, of finding* $(\lambda, u \neq 0)$*, such that*

$$-\Delta u = \lambda u \quad \text{in } \Omega, \qquad u = 0, \quad \text{on } \partial\Omega.$$

Show that $\lambda > 0$*, and give the relation between* $||u||$ *and* $||\nabla u||$ *for the corresponding eigenfunction* u*.*

c) What is the optimal constant C *(expressed in terms of smallest eigenvalue* λ_1*), for which the inequality* $||u|| \leq C||\nabla u||$ *can fullfil for all functions* u*, such that* $u = 0$ *on* $\partial\Omega$*?*

Problem 10.9. *Determine the stiffness matrix and load vector if the* $cG(1)$ *finite element method applied to the Poisson's equation on a triangulation with triangles of side length* $1/2$ *in both* x_1*- and* x_2*-directions:*

$$\begin{cases} -\Delta u = 1, & \text{in} \quad \Omega = \{(x_1, x_2) : \ 0 < x_1 < 2, \ \ 0 < x_2 < 1\}, \\ u = 0, & \text{on} \quad \Gamma_1 = \{(0, x_2)\} \cup \{(x_1, 0)\} \cup \{(x_1, 1)\}, \\ \frac{\partial u}{\partial n} = 0, & \text{on} \quad \Gamma_2 = \{(2, x_2) : 0 \leq x_2 \leq 1\}. \end{cases}$$

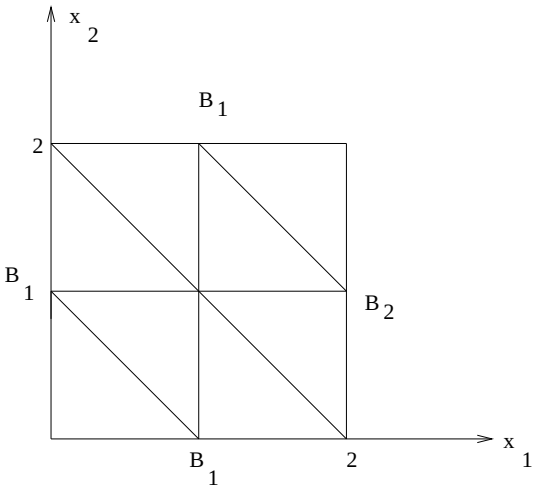
Problem 10.10. *Let* $\Omega = (0,2) \times (0,2)$*,* $B_1 = \partial\Omega \setminus B_2$ *and* $B_2 = \{2\} \times (0,2)$*. Determine the stiffness matrix and load vector in the* $cG(1)$ *solution for the problem*

$$\begin{cases} -\frac{\partial^2 u}{\partial x_1^2} - 2\frac{\partial^2 u}{\partial x_2^2} = 1, & \text{in } \Omega = (0,2) \times (0,2), \\ u = 0, \quad \text{on } B_1, & \frac{\partial u}{\partial x_1} = 0, \quad \text{on } B_2, \end{cases}$$

with piecewise linear approximation applied on the triangulation below:

Problem 10.11. *Determine the stiffness matrix and load vector if the* $cG(1)$ *finite element method with piecewise linear approximation is applied to the following Poisson's equation with mixed boundary conditions:*

$$\begin{cases} -\Delta u = 1, & \text{on} \quad \Omega = (0,1) \times (0,1), \\ \frac{\partial u}{\partial n} = 0, & \text{for} \quad x_1 = 1, \\ u = 0, & \text{for} \quad x \in \partial\Omega \setminus \{x_1 = 1\}, \end{cases}$$

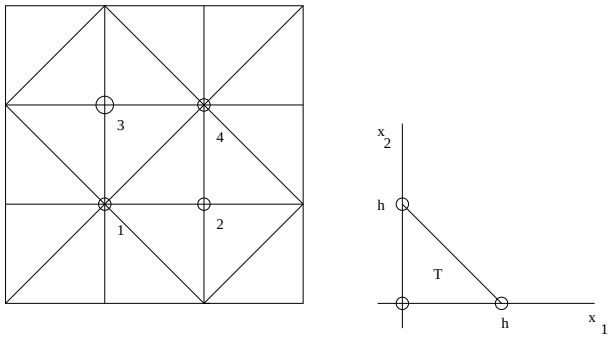


on a triangulation with triangles of side length $1/4$ *in the* x_1*-direction and* $1/2$ *in the* x_2*-direction.*

Problem 10.12. *Formulate the* $cG(1)$ *method for the boundary value problem*

$$-\Delta u + u = f, \quad x \in \Omega; \qquad u = 0, \quad x \in \partial\Omega.$$

Write down the matrix form of the resulting equation system using the following uniform mesh:



Chapter 11

The heat equation in $\mathbb{R}^N$

In this chapter we shall study the stability of the heat equation in $\mathbb{R}^d$, $d \geq 2$. The one-dimesional case is studied in Part I. Here our concern will be those aspects of the stability estimates for the higher dimensional case that are not a direct consequence of the study of the one-dimesional problem. The finite element error analysis in the higher dimensions are derived in a similar way as the corresponding 1D case. Here we omit the detailed error estimates and instead refer the reader to the text book CDE, Eriksson et al. [].

The initial boundary value problem for the heat equation can be formulated as

$$\begin{cases} \dot{u} - \Delta u = 0, & \text{in } \Omega \subset R^d, d = 1, 2, 3) \quad (DE) \\ u = 0, & \text{on } \Gamma := \partial\Omega, \quad (BC) \\ u(0, x) = u_0, & \text{for } x \in \Omega, \quad (IC) \end{cases} \tag{11.0.1}$$

where $\dot{u} = \frac{\partial u}{\partial t}$.

The equation (11.0.1) is of parabolic type with signifacnt *smoothing* and *stability* properties. It can also be used as a model for a variety of physical phenomena involving *diffusion processes*. We shall not go in detail of the physical properties for (11.0.1), instead we focus only on the stability issue.

11.1 Stability

The stability estimates for the heat equation (11.0.1) are summarized in the following theorem:

Theorem 11.1 (Energy estimates)**.** *The solution u of the initial-boundary value problem* (11.0.1) *satisfies the stability estimates*

$$\|u\|(t) \le \|u_0\| \tag{11.1.1}$$

$$\int_0^t \|\nabla u\|^2(s)\,ds \le \frac{1}{2}\|u_0\|^2 \tag{11.1.2}$$

$$\|\nabla u\|(t) \le \frac{1}{\sqrt{2\,t}}\|u_0\| \tag{11.1.3}$$

$$\Big(\int_0^t s\|\Delta u\|^2(s)\,ds\Big)^{1/2} \le \frac{1}{2}\|u_0\| \tag{11.1.4}$$

$$\|\Delta u\|(t) \le \frac{1}{\sqrt{2}\,t}\|u_0\| \tag{11.1.5}$$

$$\int_\varepsilon^t \|\dot u\|(s)\,ds \le \frac{1}{2}\sqrt{\ln\frac{t}{\varepsilon}}\|u_0\|. \tag{11.1.6}$$

Proof. To derive the first two estimates (11.1.1) and (11.1.2) we multiply (11.0.1) by u and integrate over Ω, viz

$$\int_\Omega \dot u u\,dx - \int_\Omega (\Delta u)u\,dx = 0. \tag{11.1.7}$$

Note that $\dot u u = \frac{1}{2}\frac{d}{dt}u^2$ and using Green's formula with the Dirichlet boundary data: $u = 0$ on Γ, we get

$$-\int_\Omega (\Delta u)u\,dx = -\int_\Gamma (\nabla u\cdot\mathbf{n})\,u\,ds + \int_\Omega \nabla u\cdot\nabla u\;dx = \int_\Omega |\nabla u|^2\,dx. \tag{11.1.8}$$

Thus equation (11.1.7) can be written in the following, equivalent, form:

$$\frac{1}{2}\frac{d}{dt}\int_\Omega u^2 dx + \int_\Omega |\nabla u|^2 dx = 0 \iff \frac{1}{2}\frac{d}{dt}\|u\|^2 + \|\nabla u\|^2 = 0, \tag{11.1.9}$$

where $\|\cdot\|$ denotes the $L_2(\Omega)$ norm. We substitute t by s and integrate the equation (11.1.9) over $s \in (0,t)$ to get

$$\frac{1}{2}\int_0^t \frac{d}{ds}\|u\|^2(s)ds + \int_0^t \|\nabla u\|^2(s)ds = \frac{1}{2}\|u\|^2(t) - \frac{1}{2}\|u\|^2(0) + \int_0^t \|\nabla u\|^2 ds = 0.$$

Hence, inserting the initial data $u(0) = u_0$ we have

$$\|u\|^2(t) + 2\int_0^t \|\nabla u\|^2(s)\, ds = \|u_0\|^2. \tag{11.1.10}$$

In particular, we have our first two stability estimates

$$\|u\|(t) \le \|u_0\|, \qquad \text{and} \qquad \int_0^t \|\nabla u\|^2(s)\, ds \le \frac{1}{2}\|u_0\|.$$

To derive (11.1.3) and (11.1.4) we multiply the (DE) in (11.0.1): $\dot{u} - \Delta u = 0$, by $-t \cdot \Delta u$ and integrate over Ω to obtain

$$-t\int_\Omega \dot{u} \cdot \Delta u\, dx + t\int_\Omega (\Delta u)^2\, dx = 0. \tag{11.1.11}$$

Using Green's formula ($u = 0$ on Γ) yields

$$\int_\Omega \dot{u}\, \Delta u\, dx = -\int_\Omega \nabla \dot{u} \cdot \nabla u\, dx = -\frac{1}{2}\frac{d}{dt}\|\nabla u\|^2, \tag{11.1.12}$$

so that (11.1.11) can be written as

$$t\,\frac{1}{2}\,\frac{d}{dt}\|\nabla u\|^2 + t\,\|\Delta u\|^2 = 0. \tag{11.1.13}$$

Now using the relation $t\,\frac{d}{dt}\|\nabla u\|^2 = \frac{d}{dt}(t\|\nabla u\|^2) - \|\nabla u\|^2$, we rewrite the (11.1.13) as

$$\frac{d}{dt}\Big(t\|\nabla u\|^2\Big) + 2t\|\Delta u\|^2 = \|\nabla u\|^2. \tag{11.1.14}$$

Once again we substitute t by s and integrate over $(0,t)$ to get:

$$\int_0^t \frac{d}{ds}\Big(s\|\nabla u\|^2(s)\Big)\, ds + 2\int_0^t s\|\Delta u\|^2(s)ds = \int_0^t \|\nabla u\|^2(s)ds \le \frac{1}{2}\|u_0\|^2,$$

where in the last inequality we use (11.1.2). Consequently

$$t\,\|\nabla u\|^2(t) + 2\int_0^t s\,\|\Delta u\|^2(s)\,ds \le \frac{1}{2}\|u_0\|^2. \tag{11.1.15}$$

In particular, we have:

$$\|\nabla u\|(t) \le \frac{1}{\sqrt{2t}}\|u_0\| \qquad \text{and} \qquad \Big(\int_0^t s\|\Delta u\|^2(s)\,ds\Big)^{1/2} \le \frac{1}{2}\|u_0\|,$$

which are our third and fourth stability estimates (11.1.3) and (11.1.4). The stability estimate (11.1.5) is proved analogously. Now using (11.0.1): ($\dot{u} = \Delta u$) and (11.1.5) we may write

$$\int_\varepsilon^t \|\dot{u}\|(s)ds \le \frac{1}{\sqrt{2}}\|u_0\| \int_\varepsilon^t \frac{1}{s}\,ds = \frac{1}{\sqrt{2}}\ln\frac{t}{\varepsilon}\|u_0\| \tag{11.1.16}$$

or more carefully

$$\begin{aligned}\int_\varepsilon^t \|\dot{u}\|(s)ds = \int_\varepsilon^t \|\Delta u\|(s)ds &= \int_\varepsilon^t 1\cdot\|\Delta u\|(s)ds = \int_\varepsilon^t \frac{1}{\sqrt{s}}\cdot\sqrt{s}\|\Delta u\|(s)ds\\ &\le \Big(\int_\varepsilon^t s^{-1}\,ds\Big)^{1/2}\cdot\Big(\int_\varepsilon^t s\|\Delta u\|^2(s)\,ds\Big)^{1/2}\\ &\le \frac{1}{2}\sqrt{\ln\frac{t}{\varepsilon}}\|u_0\|,\end{aligned}$$

where in the last two inequalities we use Cauchy Schwartz inequality and (11.1.4), respectively. □

Problem 11.1. *Show that* $\|\nabla u(t)\| \le \|\nabla u_0\|$ *(the stability estimate for the gradient). Hint: Multiply* (11.0.1) *by* $-\Delta u$ *and integrate over* Ω.

Is this inequality valid for $u_0 =$ *constant?*

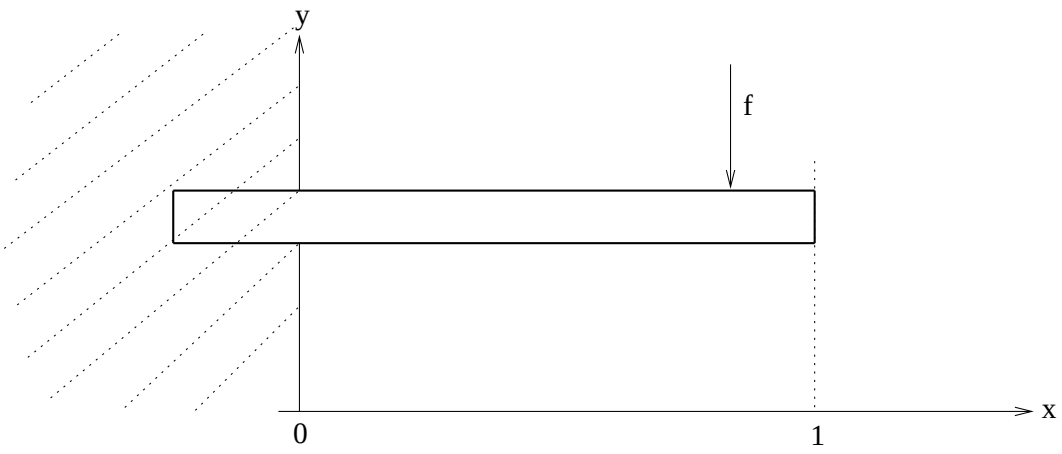
Problem 11.2. *Derive the corresponding estimate for Neuman boundary condition:*

$$\frac{\partial u}{\partial n} = 0. \tag{11.1.17}$$

Problem 11.3. *Prove the stability estimate* (11.1.5).

Example 11.1 (The equation of an elastic beam)**.** *This is an example of a stationary biharmonic equation describing the bending of an elastic beam as a one-dimensional model problem (the relation to the heat coductivity is the even number of spatial diferentiation)*

$$\begin{cases} (au'')'' = f, & \Omega = (0,1), & & \\ u(0) = 0, & u'(0) = 0, & (\textit{Dirichlet}) & \\ u''(1) = 0, & (au'')'(1) = 0, & (\textit{Neumann}) & \end{cases} \tag{11.1.18}$$



where a *is the bending stiffness*

au'' *is the moment*

f *is the function load*

$u = u(x)$ *is the vertical deflection*

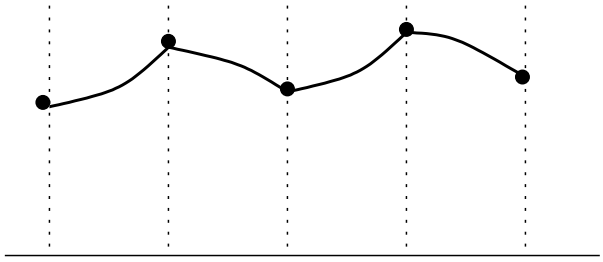
Variational form:

$$\int_0^1 au''v''dx = \int_0^1 fvdx, \quad \forall v(x) \text{ such that } v(0) = v'(0) = 0. \tag{11.1.19}$$

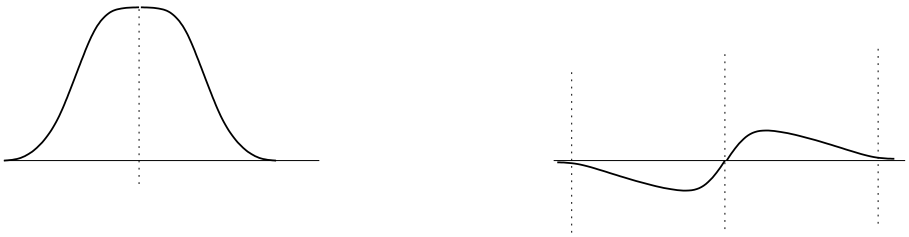
FEM: Piecewise linear functions won't work (inadequate).

11.2 Exercises

Problem 11.4. *Work out the details with piecewise cubic polynomials having continuous first derivatives: i.e., two degrees of freedom on each node.*



A cubic polynomial in (a, b) is uniquely determined by $\varphi(a), \varphi'(a), \varphi(b)$ and $\varphi'(b)$, where the basic functions would have the following form:



Problem 11.5. *Consider the following general form of the heat equation*

$$\begin{cases} u_t(x,t) - \Delta u(x,t) = f(x,t), & \text{for } x \in \Omega,\ \ 0 < t \leq T, \\ u(x,t) = 0, & \text{for } x \in \Gamma,\ \ 0 < t \leq T, \\ u(x,0) = u_0(x), & \text{for } x \in \Omega, \end{cases} \tag{11.2.1}$$

where $\Omega \in \mathbb{R}^2$ with boundary Γ. Let $\tilde{u}$ be the solution of (11.2.1) *with a modified initial data $\tilde{u}_0(x) = u_0(x)\varepsilon(x)$.*

a) Show that $w := \tilde{u} - u$ solves (11.2.1) with initial data $w_0(x) = \varepsilon(x)$.

b) Give estimates for the difference between u and $\tilde{u}$.

c) Prove that the solution of (11.2.1) is unique.

Problem 11.6. *Formulate the equation for $cG(1)dG(1)$ for the two-dimensional heat equation using the discrete Laplacian.*

Problem 11.7. *In two dimensions the heat equation, in the case of radial symmetry, can be formulated as $r\dot{u} - (ru'_r)'_r = rf$, where $r = |x|$ and $w'_r = \frac{\partial w}{\partial r}$.*

a) Verify that $u = \frac{1}{4\pi t}\exp(-\frac{r^2}{4t})$ is a solution for the homogeneous equation ($f = 0$) with the initial data being the Dirac δ function $u(r, 0) = \delta(r)$.

b) Sketching $u(r, t)$ for $t = 1$ and $t = 0.01$, deduce that $u(r, t) \to 0$ as $t \to 0$ for $r > 0$.

c) Show that $\int_{\mathbb{R}^2} u(x, t)\, dx = 2\pi \int_0^\infty u(r, t)\, r\, dr = 1$ for all t.

d) Determine a stationary solution to the heat equation with data

$$f = \begin{cases} 1/(\pi\varepsilon)^2, & \text{for} \quad r < \varepsilon, \\ 0, & \text{otherwise.} \end{cases}$$

e) Determine the fundamental solution corresponding to $f = \delta$, letting $\varepsilon \to 0$.

Problem 11.8. *Consider the Schrödinger equation*

$$i\dot{u} - \Delta u = 0, \quad \text{in } \Omega, \qquad u = 0, \quad \text{on } \partial\Omega.$$

where $i = \sqrt{-1}$ and $u = u_1 + iu_2$.

a) Show that the total probability $\int_\Omega |u|^2$ is independent of the time.

Hint: Multiplying by $\bar{u} = u_1 - iu_2$, and consider the imaginary part

b) Consider the corresponding eigenvalue problem, i.e, find the eigenvalue λ and the corresponding eigenfunction $u \neq 0$ such that

$$-\Delta u = \lambda u \quad \text{in } \Omega, \qquad u = 0, \quad \text{on } \partial\Omega.$$

Show that $\lambda > 0$ and give the relationship between the norms $\|u\|$ and $\|\nabla u\|$ for the corresponding eigenfunction u.

c) Determine (in terms of the smallest eigenvalue λ_1), the smallest possible value for the constant C in the Poincare estimate

$$\|u\| \leq C\|\nabla u\|,$$

derived for all solutions u vanishing at the boundary ($u = 0$, on $\partial\Omega$).

Problem 11.9. *Consider the initial-boundary value problem*

$$\begin{cases} u_t(x,t) - \Delta u(x,t) = f(x,t), & \text{for } x \in \Omega,\ t > 0, \\ u(x,t) = 0, & \text{for } x \in \Gamma,\ t > 0, \\ u(x,0) = u_0(x), & \text{for } x \in \Omega, \end{cases} \tag{11.2.2}$$

a) Prove (with $\|u\| = (\int_\Omega u^2\,dx)^{1/2}$) that

$$\|u(t)\|^2 + \int_0^t \|\nabla u(s)\|^2\,ds \leq \|u_0\|^2 + \int_0^t \|f(s)\|^2\,ds$$
$$\|\nabla u(t)\|^2 + \int_0^t \|\Delta u(s)\|^2\,ds \leq \|\nabla u_0\|^2 + \int_0^t \|f(s)\|^2\,ds$$

b) Formulate $dG(0) - cG(1)$ method for this problem.

Problem 11.10. *Formulate and prove $dG(0) - cG(1)$ a priori and a posteriori error estimates for the two dimentional heat equation (cf. the previous problem) that uses lumped mass and midpoit quadrature rule.*

Chapter 12

The wave equation in $\mathbb{R}^N$

The fundamental study of the wave equation in $\mathbb{R}^n,\ n \geq 2$ is an extension of the results in the one-dimensional case introduced in Part I. Some additional properties in 1D are introduced in Lecture Notes in the Fourier Analysis (see homepage of the authors). The higher dimensional problem is considered in details in our *course text book*: CDE. In the present Chapter we prove the *law of conservation of energy* for the wave equation in $\mathbb{R}^n,\ n \geq 2$, and the full study refer to CDE.

Theorem 12.1 (Conservation of energy)**.** *For the wave equation*

$$\begin{cases} \ddot{u} - \Delta u = 0, quad & in\ \Omega & (DE) \\ u = 0, & on\ \partial\Omega = \Gamma & (BC) \\ (u = u_0) \wedge (\dot{u} = v_0) & in\ \Omega,\ for\ t = 0, & (IC) \end{cases} \tag{12.0.1}$$

where $\ddot{u} = \partial^2 u/\partial t^2$ *we have that*

$$\frac{1}{2}\|\dot{u}\|^2 + \frac{1}{2}\|\nabla u\|^2 = constant,\ independent\ of\ t, \tag{12.0.2}$$

i.e., the total energy is conserved, where $\frac{1}{2}\|\dot{u}\|^2$ *is the kinetic energy, and* $\frac{1}{2}\|\nabla u\|^2$ *is the potential (elastic) energy.*

Proof. We multiply the equation by $\dot{u}$ and integrate over Ω to get

$$\int_\Omega \ddot{u} \cdot \dot{u}\, dx - \int_\Omega \Delta u \cdot \dot{u}\, dx = 0. \tag{12.0.3}$$

Using Green's formula:

$$-\int_\Omega (\Delta u)\dot{u}\, dx = -\int_\Gamma (\nabla u \cdot n)\dot{u}\, ds + \int_\Omega \nabla u \cdot \nabla \dot{u}\, dx, \tag{12.0.4}$$

and the boundary condition $u = 0$ on Γ, (which implies $\dot{u} = 0$ on Γ), we get

$$\int_\Omega \ddot{u} \cdot \dot{u}\, dx + \int_\Omega \nabla u \cdot \nabla \dot{u}\, dx = 0. \tag{12.0.5}$$

Consequently we have that

$$\int_\Omega \frac{1}{2}\frac{d}{dt}(\dot{u}^2)\, dx + \int_\Omega \frac{1}{2}\frac{d}{dt}(|\nabla u|^2)\, dx = 0 \Longleftrightarrow \frac{1}{2}\frac{d}{dt}(\|\dot{u}\|^2 + \|\nabla u\|^2) = 0,$$

and hence

$$\frac{1}{2}\|\dot{u}\|^2 + \frac{1}{2}\|\nabla u\|^2 = \text{constant, independent of}\, t,$$

and we have the desired result. □

12.1 Exercises

Problem 12.1. *Show that*

$$\|\dot{u}\|^2 + \|\nabla u\|^2 = \text{ constant, independent of } t.$$

Hint: Multiply (DE): $\ddot{u} - \Delta u = 0$ *by* $-\Delta\dot{u}$ *and integrate over* Ω.

Alternatively: differentiate the equation with respect to x *and multiply the result by* $\dot{u}$*, and continue!*

Problem 12.2. *Derive a total conservation of energy relation using the Robin type boundary condition:* $\dfrac{\partial u}{\partial n} + u = 0$.

Problem 12.3. *Determine a solution for the following equation*

$$\ddot{u} - \Delta u = e^{it}\delta(x),$$

where $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}$, $i = \sqrt{-1}$, $x = (x_1, x_2, x_3)$ *and* δ *is the Dirac-delta function.*

Hint: Let $u = e^{it}v(x)$, $v(x) = w(r)/r$ *where* $r = |x|$. *Further* $rv = w \to \frac{1}{4\pi}$ *as* $r \to 0$.

Problem 12.4. *Consider the initial boundary value problem*

$$\begin{cases} \ddot{u} - \Delta u + u = 0, & x \in \Omega, \quad t > 0, \\ u = 0, & x \in \partial\Omega, \quad t > 0, \\ u(x,0) = u_0(x), \qquad \dot{u}(x,0) = u_1(x), & x \in \Omega. \end{cases} \tag{12.1.1}$$

Rewrite the problem as a system of two equations with a time derivative of order at most 1. Why this modification is necessary?

Problem 12.5. *Consider the initial boundary value problem*

$$\begin{cases} \ddot{u} - \Delta u = 0, & x \in \Omega, \quad t > 0, \\ u = 0, & x \in \partial\Omega, \quad t > 0, \\ u(x,0) = u_0(x), \qquad \dot{u}(x,0) = u_1(x), & x \in \Omega. \end{cases} \tag{12.1.2}$$

Formulate the $cG(1)$ *method for this problem. Show that the energy is conserved.*

Chapter 13

Convection - diffusion problems

Most of the multi-physical phenomena are described by the convection, diffusion and absorption. Fluid- and gas dynamical problems, chemical reaction-diffusion, electromagnetic fields, collisions in plasma of charged Coulomb particles (electron and ions), particle transport processes both in micro (neutron transport) and macro-dimension (traffic flow with cars as particles) are often modeled as convection diffusion and absorption type problems. In this chapter we shall give a brief review of the problem in the one-dimensional case. The higher dimensional case will be considered in a forthcoming version of this notes.

13.1 A convection-diffusion model problem

We illustrate the convection-diffusion phenomenon by an example:

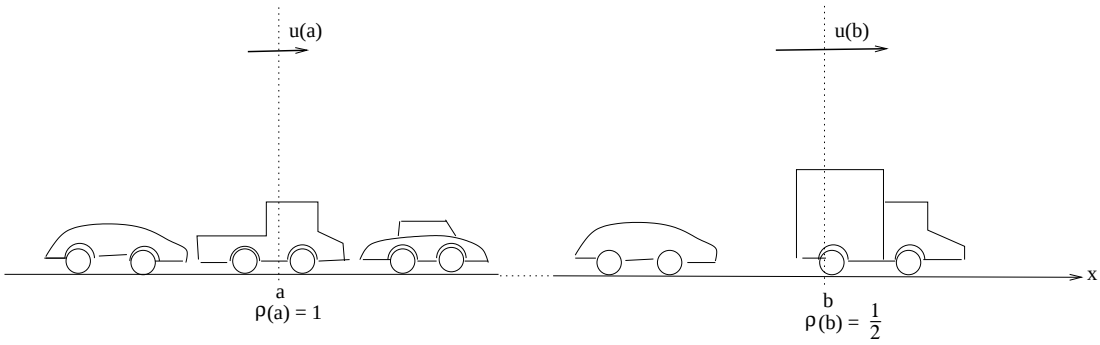
Example 13.1 (A convection model)**.** *Consider the traffic flow in a highway, viz the Fig. below. Let* $\rho = \rho(x,t)$ *be the density of cars* $(0 \leq \rho \leq 1)$ *and* $u = u(x,t)$ *the velocity (speed vector) of the cars at the position* $x \in (a,b)$ *and time* t*. For a highway path* (a,b) *the difference between the traffic inflow* $u(a)\rho(a)$ *at the point* $x = a$ *and outflow* $u(b)\rho(b)$ *at* $x = b$ *gives the density*

variation on the interval (a, b)*:*

$$\frac{d}{dt}\int_a^b \rho(x,t)dx = \int_a^b \dot{\rho}(x,t)dx = \rho(a)u(a) - \rho(b)u(b) = -\int_a^b (u\rho)'dx$$

or equivalently

$$\int_a^b \Big(\dot{\rho} + (u\rho)'\Big)\, dx = 0. \tag{13.1.1}$$



Since a *and* b *can be chosen arbitrary, thus we have*

$$\dot{\rho} + (u\rho)' = 0. \tag{13.1.2}$$

Let now $u = 1 - \rho$*, (motivate this choice), then* (13.1.2) *is rewritten as*

$$\dot{\rho} + \Big((1-\rho)\rho\Big)' = \dot{\rho} + (\rho - \rho^2)' = 0. \tag{13.1.3}$$

Hence

$$\dot{\rho} + (1-2\rho)\rho' = 0 \qquad (A \text{ non-linear convection equation}). \tag{13.1.4}$$

Alternatively, to obtain a convection-diffusion model*), we may assume that* $u = c - \varepsilon \cdot (\rho'/\rho),\ c > 0,\ \varepsilon > 0$*, (motivate). Then we get from* (13.1.2) *that*

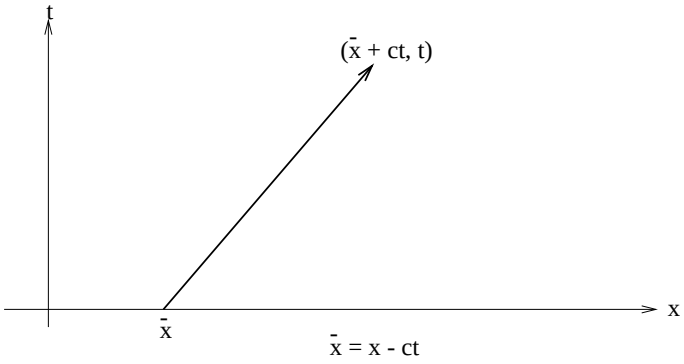
$$\dot{\rho} + \Big((c - \varepsilon\frac{\rho'}{\rho})\rho\Big)' = 0, \tag{13.1.5}$$

i.e.,

$$\dot{\rho} + c\rho' - \varepsilon\rho'' = 0 \quad (A \text{ convection - diffusion equation}). \tag{13.1.6}$$

The equation (13.1.6) *is convection dominated if* $c > \varepsilon$.

For $\varepsilon = 0$ *the solution is given by the exact transport* $\rho(x,t) = \rho_0(x - ct)$, *because then* $\rho =$ *constant on the* $(c,1)$*-direction.*



Note that differentiating $\rho(x,t) = \rho(\bar{x} + ct, t)$ *with respect to* t *we get*

$$\frac{\partial\rho}{\partial x} \cdot \frac{\partial x}{\partial t} + \frac{\partial\rho}{\partial t} = 0, \quad \Longleftrightarrow \quad c\rho' + \dot{\rho} = 0. \tag{13.1.7}$$

Finally, we may rewrite (13.1.6)*: our last convection-diffusion equation for* ρ, *by changing the notation from* ρ *to* u, *and replacing* c *by* β *to get*

$$\dot{u} + \beta \cdot u' - \varepsilon \cdot u'' = 0. \tag{13.1.8}$$

Remark 13.1. *Compare this equation with the Navier-Stokes equations for incompressible flow:*

$$\dot{u} + (\beta \cdot \nabla)u - \varepsilon\Delta u + \nabla P = 0, \quad \wedge \quad div\, u = 0, \tag{13.1.9}$$

where $\beta = u$, $u = (u_1, u_2, u_3)$ *is the velocity vector, with* u_1 *representing the mass,* u_2 *momentum, and* $u_3 =$ *energy. Further* P *is the pressure and* $\varepsilon = \dfrac{1}{Re}$ *with* Re *denoting the Reynold's number.*

Navier-Stokes equations are not easily solvable, for $\varepsilon > 0$ *and* small, *because of difficulties related to boundary layer and turbulence. A typical range for the Reynold's number Re is between* 10^5 *and* 10^7.

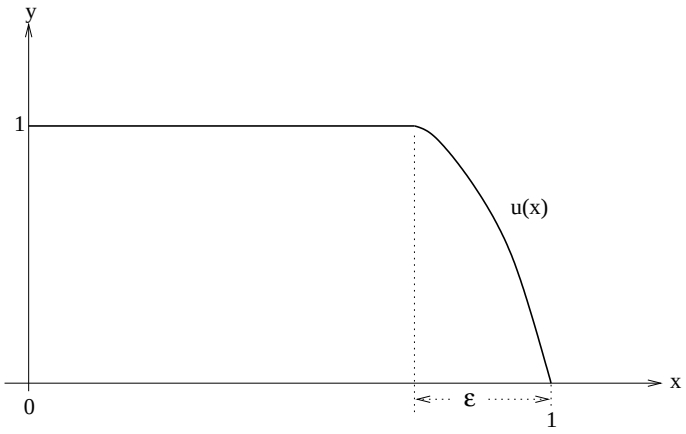
Example 13.2 (The boundary layer)**.** *Consider the following boundary value problem*

$$(BVP) \quad \begin{cases} u' - \varepsilon u'' = 0, & 0 < x < 1 \\ u(0) = 1, & u(1) = 0. \end{cases} \tag{13.1.10}$$

The exact solution to this problem is given by

$$u(x) = C\Big(e^{1/\varepsilon} - e^{x/\varepsilon}\Big), \qquad \text{with} \quad C = \frac{1}{e^{1/\varepsilon} - 1}. \tag{13.1.11}$$

which has an outflow boundary layer of width $\sim \varepsilon$*, as seen in the Fig. below*

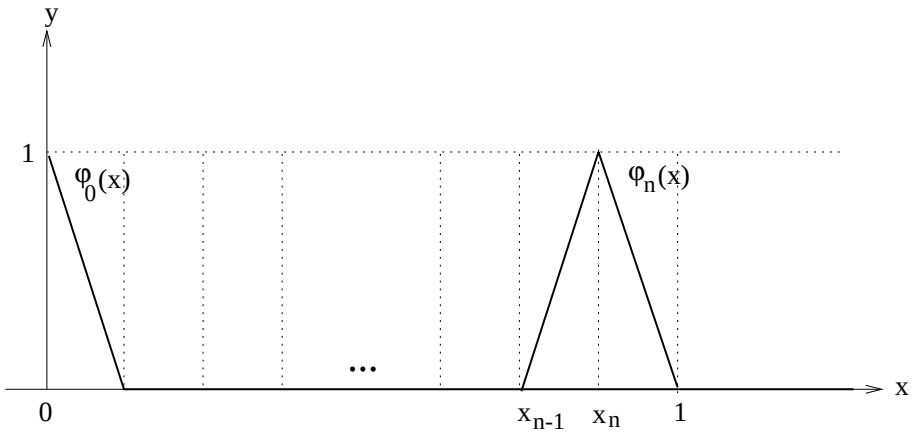


13.1.1 Finite Element Method

We shall now study the finite element solution of the problem (13.1.10). To this end we represent, as usual, the finite element solution by

$$U(x) = \varphi_0(x) + U_1\varphi_1(x) + \ldots + U_n\varphi_n(x), \tag{13.1.12}$$

where the φ_j:s are the piecewise linear basis function illustrated viz Fig. below



Evidently, the corresponding variational formulation is

$$\int_0^1 \Big(U'\varphi_j dx + \varepsilon U'\varphi_j' \Big)\, dx = 0, \quad j = 1, 2, \ldots n. \tag{13.1.13}$$

This yields the equations

$$\frac{1}{2}\Big(U_{j+1} - U_{j-1}\Big) + \frac{\varepsilon}{h}\Big(2U_j - U_{j-1} - U_{j+1}\Big) = 0, \quad j = 1, 2, \ldots, n, \tag{13.1.14}$$

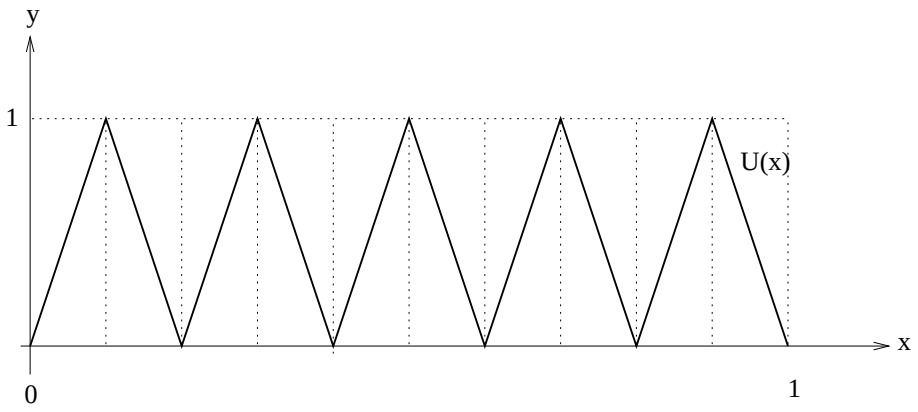
where $U_0 = 1$ and $U_{n+1} = 0$.

Note that, using *Central -differencing* we may also write

$$\underbrace{\frac{U_{j+1} - U_{j-1}}{2h}}_{\text{corresp. to } u'(x_j)} - \varepsilon \underbrace{\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2}}_{\text{corresp. to } u''(x_j)} = 0 \quad \Big(\Longleftrightarrow \frac{1}{h} \times \text{ equation(13.1.14)} \Big).$$

Now for ε being very small this gives that $U_{j+1} \approx U_{j-1}$, which results, for even n values, alternating 0 and 1 as the solution values at the nodes:

i.e., oscillations in U are transported "upstreams" making U a "globally bad approximation" of u.



A better approach would be to approximate $u'(x_j)$ by an *upwind derivative* as follows

$$u'(x_j) \approx \frac{U_j - U_{j-1}}{h}, \tag{13.1.15}$$

which, formally, gives a better stability, however, with low accuracy.

Remark 13.2. *The example above demonstrates that a high accuracy without stability is indeed useless.*

A more systematic method of making the finite element solution of the fluid problems stable is through using the streamline diffusion method which we, formally, introduce in the following subsection.

13.1.2 The Streamline - diffusion method (SDM)

The idea is to choose, in the variational formulation, the test functions of the form $(v+\frac{1}{2}\beta hv')$, instead of just v (this would finally correspond to adding an extra diffusion to the original equation in the direction of the stream-lines). Then, e.g., for our model problem we obtain the equation ($\beta \equiv 1$)

$$\int_0^1 \Big[u'(v + \frac{1}{2}hv') - \varepsilon \cdot u''\Big(v + \frac{1}{2}hv'\Big)\Big]dx = \int_0^1 f\Big(v + \frac{1}{2}hv'\Big)\,dx. \tag{13.1.16}$$

In the case of approximation with piecewise linears, in the discrete version of the variational formulation, we should interpret the term $\int_0^1 U''v'dx$ as a

sum viz,

$$\int_0^1 U''v'dx := \sum_j \int_{I_j} U''v'dx = 0. \tag{13.1.17}$$

Then, with piecewise linear test functions, i.e., choosing $v = \varphi_j$ we get the discrete term corresponding to the second integral in (13.1.16) as

$$\int_0^1 U'\frac{1}{2}h\varphi_j'dx = U_j - \frac{1}{2}U_{j+1} - \frac{1}{2}U_{j-1}, \tag{13.1.18}$$

which adding to the obvious relation

$$\int_0^1 U'\varphi_j dx = \frac{U_{j+1} - U_{j-1}}{2}, \tag{13.1.19}$$

we end up with $(U_j - U_{j-1})$, as an approximation of the first integral in (13.1.16), corresponding to the *upwind scheme*.

Remark 13.3. *The SDM can also be interpreted as a sort of least-square method:*

Let $A = \frac{d}{dx}$ *then* $A^t = -\frac{d}{dx}$. *Now* u *minimizes* $\|w' - f\|$ *if* $u' = Au = f$. *This can be written as*

$$A^tAu = A^tf \quad \Longleftrightarrow \quad -u'' = -f, \qquad \textit{(the continuous form)}. \tag{13.1.20}$$

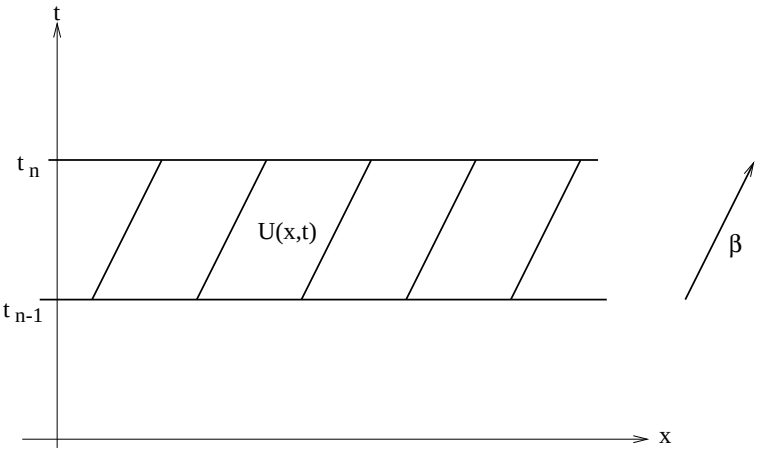
While multiplying $u' = Au = f$ *by* v *and integrating over* $(0,1)$ *we have*

$$\int_0^1 U'v'dx = \int_0^1 fv'dx \qquad (\textit{the weak form}), \tag{13.1.21}$$

where we replaced u' *by* U'.

For the time-dependent convection equation, the oriented time-space element are used. Consider the time-dependent problem

$$\dot{u} + \beta u' - \varepsilon u'' = f. \tag{13.1.22}$$



Set $U(x,t)$ such that U is piecewise linear in x and piecewise constant in the $(\beta,1)$-direction. Combine with SDM and add up some artificial viscosity, $\hat{\varepsilon}$, depending on the residual term to get for each time interval I_n:

$$\int_{I_n}\int_\Omega \Big[(\dot{U}+\beta U)\Big(v+\frac{\beta}{2}h\dot{v}\Big)+\hat{\varepsilon}\ U'v'\Big]\,dxdt = \int_{I_n}\int_\Omega f\Big(v+\frac{\beta}{2}hv'\Big)\,dxdt. \quad \square$$

13.1.3 Exercises

Problem 13.1. *Prove that the solution u of the convection-diffusion problem*

$$-u_{xx}+u_x+u=f, quad in\ I=(0,1), \quad u(0)=u(1)=0,$$

satisfies the following estimate

$$\Big(\int_I u^2\phi\,dx\Big)^{1/2} \le \Big(\int_I f^2\phi\,dx\Big)^{1/2}.$$

where $\phi(x)$ is a positive weight function defined on $(0,1)$ satisfying $\phi_x(x)\le 0$ and $-\phi_x(x)\le\phi(x)$ for $0\le x\le 1$.

Problem 13.2. *Let ϕ be a solution of the problem*

$$-\varepsilon\phi''-3\phi'+2\phi=e, \qquad \phi'(0)=\phi(1)=0.$$

Let $\|\cdot\|$ denote the L_2-norm on I. Show that there is a constant C such that

$$|\phi\prime(0)|\le C\|e\|, \qquad \|\varepsilon\phi''\|\le C\|e\|.$$

Problem 13.3. *Consider the convection-diffusion-absorption problem*

$$-\varepsilon u'' + xu' + u = f, \quad \text{in } I = (0,1), \quad u(0) = u'(1) = 0,$$

where ε is a positive constant, and $f \in L_2(I)$. Prove that

$$\|\varepsilon u''\| \leq \|f\|,$$

where $\|\cdot\|$ denotes the $L_2(I)$-norm.

Problem 13.4. *Use relevant interpolation theory estimates and prove an a priori and an a posteriori error estimate for the $cG(1)$ finite element method for the problem*

$$-u'' + u' = f, \quad \text{in } I = (0,1), \quad u(0) = u(1) = 0.$$

Problem 13.5. *Prove an a priori and an a posteriori error estimate for the $cG(1)$ finite element method for the problem*

$$-u'' + u' + u = f, \quad \text{in } I = (0,1), \quad u(0) = u(1) = 0.$$

Problem 13.6. *Consider the convection-diffusion-absorption problem*

$$-\varepsilon u_{xx} + u_x + u = f, \quad \text{in} \quad I = (0,1), \quad u(0) = 0, \qquad \sqrt{\varepsilon}u_x + u(1) = 0,$$

where ε is a positive constant, and $f \in L_2(I)$. Prove the following stability estimates for the solution u

$$\|\sqrt{\varepsilon}u_x\| + \|u\| + |u(1)| \leq C\|f\|,$$
$$\|u_x\| + \|\varepsilon u_{xx}\| \leq C\|f\|,$$

where $\|\cdot\|$ denotes the $L_2((0,))$-norm and C is an appropriate constant.

Problem 13.7. *Consider the convection problem*

$$\beta \cdot \nabla u + \alpha u = f, \qquad x \in \Omega, \quad u = g, quad x \in \Gamma_-, \tag{13.1.23}$$

Define the outflow Γ_+ ans inflow Γ_- boundaries. Assume that $\alpha - \frac{1}{2}\nabla \cdot \beta \geq c > 0$. Show the following stability estimate

$$c\|u\|^2 \int_{\Gamma_+} n \cdot \beta u^2 \, ds \, dt \leq \|u_0\|^2 + \frac{1}{c}\|f\|^2 + \int_{\Gamma_-} |n \cdot \beta| g^2 \, ds. \tag{13.1.24}$$

Hint: Show first that

$$2(\beta \cdot \nabla u, u) = \int_{\Gamma_+} n \cdot \beta \, u^2 \, ds - \int_{\Gamma_-} \|n \cdot \beta\| \, u^2 \, ds - ((\nabla \cdot \beta)u, u).$$

Formulate the streamline diffusion for this problem.

Problem 13.8. *Consider the convection problem*

$$\begin{aligned} &\dot{u} + \beta \cdot \nabla u + \alpha u = f, && x \in \Omega, \quad t > 0, \\ &u = g, && x \in \Gamma_-, \quad t > 0, \\ &u(x,0) = u_0(x), && x \in \Omega, \end{aligned} \tag{13.1.25}$$

where Γ_+ and Γ_- are defined as above. Assume that $\alpha - \frac{1}{2}\nabla \cdot \beta \geq c > 0$. Show the following stability estimate

$$\begin{aligned} \|u(\cdot,T)\|^2 + c\int_0^T \|u(\cdot,t)\|^2\,dt + \int_0^T \int_{\Gamma_+} n \cdot \beta u^2\,ds\,dt \\ \leq \|u_0\|^2 + \frac{1}{c}\int_0^T \|f(\cdot,t)\|^2\,dt + \int_0^T \int_{\Gamma_-} |n \cdot \beta| g^2\,ds\,dt, \end{aligned} \tag{13.1.26}$$

where $\|u(\cdot,T)\|^2 = \int_\Omega u(x,T)^2\,dx$.

Answers to Exercises

Chapter 1. Exercise Section 1.1.1

1.1 The solution is unique only for $\ell \neq n\pi$, where n is an integer.

1.2 a) No! b) $\int_0^\ell f(x)\,dx = 0$.

1.3 $\sum_{i=1}^{N} c_i = 1$.

1.4 b) $c_1 = 0$, $c_2 = 1$.

1.5 a) $a + 3b = 0$, b) $3a - \pi b = 0$, c) $2a + eb = 0$.

1.6 a) $u_{ss} = 0$; $v(s,t) = u(x,y)$

b) $u(x,y) = f(x-y) + xg(x-y)$.

c) $v_{ss} + v_{tt} = 0$, $v(s,t) = u(x,y)$.

1.7 c) $\dfrac{\sin 2\pi x \sinh 2\pi y}{\sinh 2\pi} - \dfrac{\sin 3\pi x \sinh 3\pi y}{\sinh 3\pi} + \dfrac{2\sin \pi x \sinh \pi(1-y)}{\sinh \pi}$.

Chapter 1. Exercise Section 1.2.1

1.11 u = constant gives circles with center $\left(1, \frac{-1}{u}\right)$ and radius $1/|u|$.

v = constant gives circles with center $\left(\frac{-v}{v-1}, 0\right)$ and radius $1/|v-1|$.

1.15 The term represents heat loss at a rate proportional to the execess temperature over θ_0.

1.18 $a^2 = b^2c^2$.

1.19 $\alpha = \pm c$.

Chapter 2. Exercise Section 2.2

2.3
$$Pu(t) = 0.9991 + 1.083t + 0.4212t^2 + 0.2786t^3.$$

2.5 a. $u(x) = \frac{1}{2}x(1-x)$

b. $R(x) = \pi^2 A \sin \pi x + 4\pi^2 B \sin 2\pi x - 1$

c. $A = 4/\pi^3$ and $B = 0$.

d. -

2.6 a. -

b. $R(x) = (\pi^2+1)A \sin \pi x + (4\pi^2+1)B \sin 2\pi x + (9\pi^2+1)C \sin 3\pi x - x$

c. $A = \dfrac{2}{\pi(\pi^2+1)}$, $B = -\dfrac{1}{\pi(4\pi^2+1)}$ and $C = \dfrac{2}{3\pi(9\pi^2+1)}$.

2.7 a. $u(x) = \frac{1}{6}(\pi^3 - x^3) + \frac{1}{2}(x^2 - \pi^2)$

b. $R(x) = -U''(x) - x + 1 = \frac{1}{4}\xi_0 \cos\frac{x}{2} + \frac{9}{4}\xi_1 \cos\frac{3x}{2}$

c. $\xi_0 = 8(2\pi - 6)/\pi$ and $\xi_1 = \frac{8}{9}(\frac{2}{9} - \frac{2}{3}\pi)/\pi$.

2.8 $U(x) = (16 \sin x + \frac{16}{27} \sin 3x)/\pi^3 + 2x^2/\pi^2$.

Chapter 3. Exercise Section 3.4

3.2 (a) x, (b) 0.

3.3
$$\Pi_1 f(x) = \begin{cases} 4 - 11(x+\pi)/(2\pi), & -\pi \le x \le -\frac{\pi}{2}, \\ 5/4 - (x+\frac{\pi}{2})/(2\pi), & -\frac{\pi}{2} \le x \le 0, \\ 1 - 7x/(2\pi), & 0 \le x \le \frac{\pi}{2}, \\ 3(x-\pi)/(2\pi), & \frac{\pi}{2} \le x \le \pi. \end{cases}$$

3.8 Check the conditions required for a Vector space.

3.9
$$\Pi_1 f(x) = f(a)\frac{2x-a-b}{a-b} + f(\frac{a+b}{2})\frac{2(x-a)}{b-a}.$$

3.11 Hint: Use the procedure in the proof of Theorem 3.1, with somewhat careful estimates at the end.

3.12
$$\pi_4\left(e^{-8x^2}\right) \approx 0.25x^4 - 1.25x^2 + 1.$$

3.13 For example we may choose the following basis:

$$\varphi_{i,j}(x) = \begin{cases} 0, & x \in [x_{i-1}, x_i], \\ \lambda_{i,j}(x), & i = 1, \ldots, m+1, \quad j = 0, 1, 2. \end{cases}$$

$$\lambda_{i,0}(x) = \frac{(x-\xi_i)(x-x_i)}{(x_{i-1}-\xi_i)(x_{i-1}-x_i)}, \quad \lambda_{i,1}(x) = \frac{(x-x_{i-1})(x-x_i)}{(\xi_i-x_{i-1})(\xi_i-x_i)},$$

$$\lambda_{i,2}(x) = \frac{(x-x_{i-1})(x-\xi_i)}{(x_i-x_{i-1})(x_i-\xi_i)}, \qquad \xi_i \in (x_{i-1}, x_i).$$

3.14 This is a special case of problem 3.13.

3.15 Trivial

3.16 Hint: Use Taylor expansion of f about $x = \frac{x_1+x_2}{2}$.

Chapter 4. Exercise Section 4.3

4.2
$$LU = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 2 \\ 0 & -1 & 3 \\ 0 & 0 & 5 \end{bmatrix}.$$

4.3
$$x = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

4.4
$$LDU = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & -4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 14 \end{bmatrix} \begin{bmatrix} 1 & 1 & -3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

4.7 The exact solution is $(1/15, -11/15) = (0.066666, -0.733333)$.
(a) $(u_1^3, u_2^3) = (5/64, -47/64)$, $\rho(J) = 1/4$ and $||e_3||_\infty = 0.011$.
(b) $(u_1^3, u_2^3) = (0.0673828, -0.7331543)$, $\rho(G) = 1/16$ and $||e_3||_\infty = 7 \times 10^{-4}$.
(c) $(u_1^3, u_2^3) = (0.066789, -0.733317)$, $\rho(\omega_0) = 0.017$ and $||e_3||_\infty = 1 \times 10^{-4}$.

Chapter 5. Exercise Section 5.5

5.1 c) $\sin \pi x$, $x \ln x$ and $x(1-x)$ are test functions of this problem. x^2 and $e^x - 1$ are not test functions.

5.3 a) U is the solution for
$$AU = f \Longleftrightarrow 1/h \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} = h \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$
with $h = 1/4$.
b) A is invertible, therefore U is unique.

5.6 a) ξ is the solution for
$$2 \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 7 \end{pmatrix}$$

b) $(\xi_1, \xi_2) = 7(1/2, 1)$ and $U(x) = 7x$ (same as the exact solution).

5.7 a) ξ is the solution for

$$A\xi = f \Longleftrightarrow 1/h \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} \xi_0 \\ \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} -5 \\ 0 \\ 0 \end{pmatrix}$$

with $h = 1/3$. That is: $(\xi_0, \xi_1, \xi_2) = -\frac{1}{3}(15, 10, 5)$.

b) $U(x) = 5x - 5$ (same as the exact solution).

5.8 a) No solution!

b) Trying to get a finite element approximation ends up with the matrix equation

$$A\xi = f \Longleftrightarrow \begin{pmatrix} 2 & -2 & 0 \\ -2 & 4 & -2 \\ 0 & -2 & 2 \end{pmatrix} \begin{pmatrix} \xi_0 \\ \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

where the coefficient matrix is singular ($detA = 0$). There is no finite element solution.

5.9 d) $||U||_E^2 = \xi^T A\xi$ (check spectral theorem, linear algebra!)

5.10 For an $M + 1$ partition (here $M = 2$) we get $a_{ii} = 2/h$, $a_{i,i+1} = -1/h$ except $a_{M+1,M+1} = 1/h - 1$, $b_i = 0,\ i = 1, \ldots, M$ and $b_{M+1} = -1$:

a) $U = (0, 1/2, 1, 3/2)$.

b) e.g, $U_3 = U(1) \to 1$, as $k \to \infty$.

5.11 c) Set $\alpha = 2$ and $\beta = 3$ in the general FEM solution:
$\xi = \frac{\alpha}{3}(-1, 1, 1)^T + \beta(0, 0, 2)^T$:

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = 2/3 \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}.$$

5.12
$$3\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \frac{1}{18}\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\Longleftrightarrow \text{(MATLAB)} \quad \xi_1 = \xi_2 = 0.102.$$

5.13 Just follow the procedure in the theory.

5.15 a priori: $||e||_E \leq ||u - \pi_h u||_E$.
a posteriori: $||e||_E \leq C_i||hR(U)||_{L_2(I)}$.

5.16 a) $||e'||_a \leq C_i||h(aU')'||_{1/a}$.
b) The matrix equation:

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 3 & -2 \\ 0 & 0 & 2 & 4 \end{pmatrix}\begin{pmatrix} \xi_0 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

which yields the approximate solution $U = -3(1/2, 1, 2, 3)^T$.
c) Since a is constant and U is linear on each subinterval we have that

$$(aU')' = a'U' + aU'' = 0.$$

By the a posteriori error estimate we have that $||e'||_a = 0$, i.e. $e' = 0$. Combining with the fact that $e(x)$ is continuous and $e(1) = 0$, we get that $e \equiv 0$, which means that the finite, in this case, coincides with the exact solution.

5.17 a priori: $||e||_{H^1} \leq C_i\Big(||hu''|| + ||h^2u''||\Big)$.
a posteriori: $||e||_{H^1} \leq C_i||hR(U)||_{L_2(I)}$.

5.18 a) a priori: $||e||_E \leq ||u - v||_h(1 + c)$, and a posteriori: $||e||_E \leq C_i||hR(U)||_{L_2(I)}$.
b) Since $c \geq 0$, the a priori error estimate in a) yields optimality for $c \equiv 0$, i.e. in the case of no convection (does this tell anything to you?).

5.19 a priori: $||e||_{H^1} \leq C_i\Big(||hu''||+4||h^2u''||\Big)$, a posteriori estimate is similar (see [5.17]).

Chapter 6. Exercise Section 6.6

6.3 a) $a_{ij} = \frac{j}{j+i} - \frac{1}{j+i+1}, \qquad b_i = \frac{1}{i+1}, \qquad i,j = 1,2,\ldots,$

b) $q = 1: \quad U(t) = 1 + 3t. \qquad q = 2: \quad U(t) = 1 + \frac{8}{11}t + \frac{10}{11}t^2.$

6.5 a) $u(t) = e^{-4t} + \frac{1}{32}(8t^2 - 4t + 1).$

b) $u(t) = e^{\frac{1}{2}t^2} - t + \frac{\sqrt{\pi}}{\sqrt{2}}e^{\frac{1}{2}t^2}erf(\frac{t}{\sqrt{2}}), \quad erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-y^2}\,dy.$

6.6 a) $U_i(x_i) = \frac{[(x_i^3 - x_{i-1}^3)/3] - U_i(x_{i-1})\cdot(2(x_i - x_{i-1}) - 1)}{1+2(x_i - x_{i-1})}$

6.9 a)

Explicit Euler: $U_n = -3U_{n-1}, \quad U_0 = 1.$

Implicit Euler $U_n = \frac{1}{5}U_{n-1}, \quad U_0 = 1.$

Crank-Nicolson: $U_n = \frac{1}{3}U_{n-1}, \quad U_0 = 1.$

b)

Explicit Euler: $|U_n| = \sqrt{1+0.01}|U_{n-1}| \quad \Longrightarrow \quad |U_n| \geq |U_{n-1}|.$

Implicit Euler: $|U_n| = \frac{1}{\sqrt{1+0.01}}|U_{n-1}| \quad \Longrightarrow \quad |U_n| \leq |U_{n-1}|.$

Crank-Nicolson: $|U_n| = |\frac{1-0.2i/2}{1+0.2i/2}||U_{n-1}| = |U_{n-1}|.$

Chapter 7. Exercise Section 7.1.4

7.9 $||e|| \leq ||h^2 u_{xx}||$

Chapter 7. Exercise Section 7.2.3

7.15

$$u(x,t) = \begin{cases} \frac{1}{2}(u_0(x+2t) + u_0(x-2t)), & x \geq 2t \\ \frac{1}{2}(u_0(2t+x) + u_0(2t-x)), & x < 2t \end{cases}$$

7.18 a) $u(x,t) = \frac{1}{2}[u_0(x+ct) + u_0(ct-x)] + \frac{1}{2c}\Big(\int_0^{x+ct} v_0 + \int_0^{ct-x} v_0\Big)$.

b) $u(x,t) = \frac{1}{2c}\int_0^t 2c(t-s)\,ds = t^2/2$.

Chapter 7. Exercise Section 7.3.3

7.21 a priori: $||e||_{H^1} \leq C_i\Big(||hu''|| + ||h^2u''||\Big)$.

a posteriori: $||e||_{H^1} \leq C_i||hR(U)||$.

7.22 a priori: $||e||_E \leq C_i\Big(||hu''|| + ||h^2u''||\Big)$.

a posteriori: $||e||_E \leq C_i||hR(U)||$.

Bibliography

[1] M. Anisworth and J. T. Oden, *posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.

[2] V. I. Arnold, *Ordinary differential Equations*, (Translated from the Russian by Richard A. Silverman), MIT Press, Cambridge, Massachusetts, and London, England, 2en Ed. 1980.

[3] M. Asadzadeh, *Lecture Notes in Fourier Analysis*, (Electronic available in author's website), pp 342.

[4] K. Atkinson, *An Introduction to Numerical Analysis*, 2ed Ed. John Wiley & Sons, Inc, New York, 1989.

[5] K. Böhmer, *Numerical Methods for Nonlinear Elliptic Differential Equations*, Oxford University Press, 2010.

[6] D. Braess, *Finite Elements*. Theory, fast solvers, and application in solid mechanics, 2ed Ed. Cambridge Universityu Press, 2001.

[7] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[8] L. R. Burden and J. D. Faires, *Numerical Analysis*, Fifth Ed. Brook/Cole, CA, 1998.

[9] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-

Holland, New York, 1980.

[10] P. Ciarlet, *Introduction to Matrix Numerical Analysis and Optimization* (in French), Masson, Paris, 1982.

[11] A. Ern and J. -L. Guermond, *Theory and Practice of Finite Elements*, Applied Mathematical Sciences, Springer Verlag, 2004.

[12] K. Eriksson, D. Estep, P. Hansbo and C. Johnson, *Computational Differential Equations*, Studentlitteratur, Lund, 1996.

[13] L. C. Evans, *Partial Differential Equations*, Graduate Studies in Mathematics, 19. Americam Mathematical Society, Providance, RI, 1998.

[14] G. B. Folland, *Fourier Analysis and its Applications*, Waswoth & Cole 1992.

[15] G. B. Folland, *Intorduction to Partial Differential Equations*, Princeton University Press, 1976.

[16] G. Golub and C. V. Loan, *Matrix Computations*, John Hopkins University Press, Maryland, 1983.

[17] K. E. Gustafson, *Partial Differential Equations and Hilbert Space Methods*, John Wiley & Sons, New York, 1980.

[18] T. J. R. Hughes, *The Finite Element Method.* Linear Static and Dynamic Finite Element Analysis, Prentice-Hall, Englewood Cliffs, New Jersy, 1987.

[19] C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Element Method*, Studentlitteratur, Lund, 1991.

[20] S. Larsson and V. Thomee, *Partial differential equations with numerical methods.* Texts in Applied Mathematics, 45. Springer-Verlag, Berlin, 2003.

[21] J. T. Oden, *Finite elements: an introduction.* Handbook of numerical analysis, Vol. II, 3–15, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991.

[22] G. Strang, *Introduction to Applied Mathematics*, Wellesely-Cambridge Press, Cambridge, Mass, 1986.

[23] G. Strang, and G. J. Fix, *An analysis of the finite element method.* Prentice-Hall Series in Automatic Computation. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.

[24] W. Strauss, *Partial Differential Equations*. An Introduction, 2ed Ed. John Wiley & Sons, Ltd, 2008.

[25] M. E. Taylor, *Partial Differential Equations. I. Basic Theory*. Applied Mathematical Sciences, 115. Springer-Verlag, New York, 1996.

[26] V. Thomee, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Mathematics 1054. Springer-Verlag, New York, Tokyo, 1984.

[27] S. Yakowitz and F. Szidarovsky, *An Introduction to Numerical Computations*, 2ed Ed. Macmillan Co. New York, 1989.

[28] O. C. Zienkiewicz, *The Finite Element Method in Structural and Continum Mechanics*. McGraw-Hill, London, 1971.