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

Mohammad Asadzadeh

January 20, 2010

Contents

Chapter 0

Introduction

This note presents an introduction to the Galerkin finite element method (FEM), as a general tool for numerical solution of partial differential equations (PDEs). Iteration procedures and interpolation techniques are also employed to derive basic *a priori* and *a posteriori* error estimates, necessary for, e.g. solution properties, such as stability and convergence. 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 space.

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 of the function u that appears in the equation.

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

$$u_t(x,t) - u_{xx}(x,t) = 0 \quad \text{is a homogeneous PDE of second order,}$$

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. $u(x)$, $u(t,x)$ 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 question in here is: *how close is the approximate solution to the exact solution?* and how and in which environment does one measure this *closeness*? In which extent the approximate solution preserves the physical quality of the exact solution? These are some of the questions that we want to deal with in this notes.

• The linear differential equation of order n in time 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 ordinary time derivative, and $D^k = \frac{d^k}{dt^k}$, $1 \leq k \leq n$. 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

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

The usual three operators involved in differential equations are

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

$$\text{Diffusion operator} \qquad \frac{d}{dt} - \Delta_n, \tag{0.2.2}$$

$$\text{D'Alembert operator} \qquad \square = \frac{d^2}{dt^2} - \Delta_n, \tag{0.2.3}$$

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

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

Classifying the general second order PDEs in two dimensions

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

Second order PDEs 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 1. *Here are the class of the most common equations:*

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

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

Example 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, e.g. the domain of ellipticity is* $y > 0$, *and so on (see the Fig. below)*

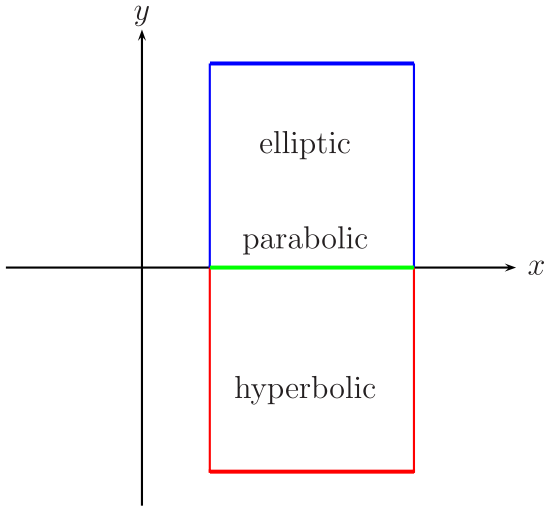


Figure 1: Tricomi: an example of a variable coefficient classification.

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

Mathematical Quantity	**Surname**	**Physical Named**	**Classification Type**
Δ_n	Laplacian	Potential operator	Elliptic
$\dfrac{d}{dt} - \Delta_{n-1}$	Heat	Diffusion operator	Parabolic
$\Box = \frac{d^2}{dt^2} - \Delta_{n-1}$	d'Alembertian	Wave operator	Hyperbolic

The usual three types 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 the *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 3. *The wave equation, on real line, associated with the given initial data:*

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

2. Boundary value problems (BVP)

a. Consider the boundary value problems in $\mathbb{R}$:

Example 4. *The stationary heat equation:*

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

To define a solution $u(x)$ uniquely, the differential equation is complemented by boundary conditions imposed at the boundaries $x = 0$ 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 (BVP) in $\mathbb{R}^n$:

Example 5. *The Laplace equation in $\mathbb{R}^n$, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$:*

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

Remark 1. *In general, in order to obtain a unique solution for a (partial) differential equation, one should supply as many data as the sum of highest order (partial) derivatives involved in the equation. Thus in example 1, to determine a unique solution for the potential equation $u_{xx} + u_{yy}$ we need to give 2 boundary conditions in the x-direction and another 2 in the y-direction, whereas to determine a unique solution for the wave equation $u_{tt} - u_{xx} = 0$, it is necessary to supply 2 initial and 2 boundary conditions.*

3. Eigenvalue problems (EVP)

Let $\mathbf{A}$ be a given matrix. The relation $\mathbf{Av} = \lambda\mathbf{v},\ \mathbf{v} \neq 0$ is a linear equation system where λ is an *eigenvalue* and $\mathbf{v}$ is an *eigenvector*.

Example 6. *In the case of a differential equation; e.g. the equation of a steady state vibrating string*

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

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, and fixed at the end points, is given by

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

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

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

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),

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

2. *Neumann* boundary condition (the derivative of the solution at the direction of outward normal is given)

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

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

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

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

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

Example 7. *For $u = u(x,y)$ we have $\mathbf{n} = (n_1, n_2)$, with $|\mathbf{n}| = \sqrt{n_1^2 + n_2^2} = 1$ and $\mathbf{n} \cdot \nabla u = n_1 u_x + n_2 u_y$.*

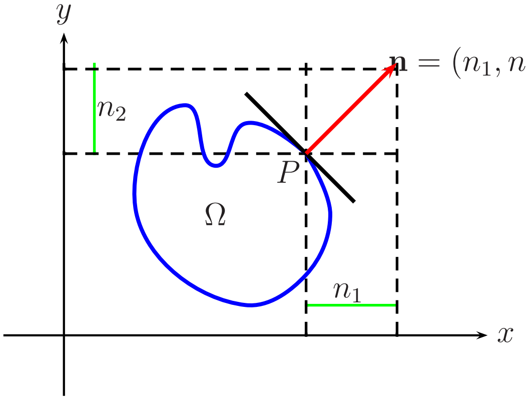


Figure 2: The domain Ω and its outward normal $\mathbf{n}$ at a point $P \in \partial\Omega$.

Example 8. *Let $u(x,y) = x^2 + y^2$. We determine the normal derivative of u in direction $v = (1,1)$. The gradient of u is the vector valued function $\nabla u = 2x \cdot e_1 + 2y \cdot e_2$, where $e_1 = (1,0)$ and $e_2 = (0,1)$ are the unit orthonormal basis in $\mathbb{R}^2$: $e_1 \cdot e_1 = e_2 \cdot e_2 = 1$ and $e_1 \cdot e_2 = e_2 \cdot e_1 = 0$. Note that $v = e_1 + e_2 = (1,1)$ is not a unit vector. The normalized v is obtained viz $v_n = v/|v|$, i.e.*

$$v_n = \frac{e_1 + e_2}{|e_1 + e_2|} = \frac{(1,1)}{\sqrt{1^2 + 1^2}} = \frac{(1,1)}{\sqrt{2}}.$$

Thus with $\nabla u(x,y) = 2x \cdot e_1 + 2y \cdot e_2$, we get

$$v_n \cdot \nabla u(x,y) = \frac{e_1 + e_2}{|e_1 + e_2|}(2x \cdot e_1 + 2y \cdot e_2).$$

which gives

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

Thus

$$v_n \cdot \nabla u(1,1) = \frac{4}{\sqrt{2}} = 2\sqrt{2}.$$

The usual three questions

I. In theory

1. *Existence*, at least one solution u
2. *Uniqueness*, either one solution or no solutions at all
3. *Stability*, continuous dependence of solutions to the data

Remark. 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

1. *Construction*, of the *physical* solution.
2. *Regularity*, how substitutable is the found solution.
3. *Approximation*, when an exact construction of the solution is impossible.

Three general approaches to analyzing differential equations

1. *Transformation to a simpler problem*: The separation of variables technique to reduce the (PDEs) to simpler eigenvalue problems (ODEs). Also called *Fourier method, or solution by eigenfunction expansion* (Fourier analysis).

2. *The multiplier method*: The multiplier method is a strategy for extracting information by multiplying a differential equation by a suitable function and then integrating. This usually is referred as *variational formulation, or energy method* (subject of our study).

3. *Green's Function*, fundamental singularities, or solution by integral equations (an advanced PDE course).

Chapter 1

Polynomial approximation in 1d

Our objective is to present the finite element method 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 finite number of unknowns. The more unknowns we keep, the better will be the accuracy of the approximation and greater the expences.

1.1 Overture

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

Ordinary differential equations (ODEs)
An *initial value problem*, (IVP), in population dynamics can be written as

$$\dot{u}(t) = \lambda u(t), \quad 0 < t < 1 \qquad u(0) = u_0, \tag{1.1.1}$$

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 would blow 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 $\mathbf{u}(t) = [u_1(t), u_2(t), \ldots, u_n(t)]^T$, $\dot{\mathbf{u}}(t) = [u_1'(t), u_2'(t), \ldots, u_n'(t)]^T$ 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 Eucledean 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*,

$$\begin{cases} -\Delta u(\mathbf{x}) + \alpha\mathbf{b} \cdot \nabla u(\mathbf{x}) = f, & \mathbf{x} \in \Omega \subset \mathbb{R}^n, \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega. \end{cases} \tag{1.1.2}$$

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 $\nabla u := \Big(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_n}\Big)$, $\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2}$, 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 following *Heat equation* is an example of a boundary value problem with

- *Neumann boundary condition*

$$\frac{\partial u}{\partial t} = \Delta u, \quad \mathbf{x} \in \Omega \subset \mathbb{R}^k, \qquad \frac{\partial u}{\partial \mathbf{n}} = 0, \mathbf{x} \in \partial\Omega, \tag{1.1.3}$$

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

$$\frac{\partial u}{\partial \mathbf{n}} = \mathbf{n} \cdot \nabla u. \tag{1.1.4}$$

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.

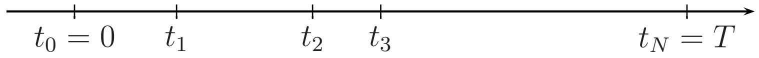
3) $u \in \mathcal{C}^2 :$ all partial derivatives of u of order 2 are continuous.

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. *Note that, we tacitly understand that:* $\mathbf{u}$ *in 1) is a vector-valued function of a single variable as in the example of, general, dynamical system* (1.1.1)*, whereas* u *in 2)-4) is a scalar (real-valued) function of several variables.*

• **Numerical solutions of (IVP)**

Example 9. *A finite difference method.*
We descritize the IVP (1.1.1) *with* explicit (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 $[t_k, t_{k+1}]$ *as* $\dot{u}(t) \approx \frac{u(t_{k+1})-u(t_k)}{t_{k+1}-t_k}$*. Then an approximation of* (1.1.1) *is given by*

$$\frac{u(t_{k+1}) - u(t_k)}{t_{k+1} - t_k} = \lambda \cdot u(t_k), \qquad k = 0, 1, \dots N-1, \qquad \textit{with} \qquad u(0) = u_0, \tag{1.1.5}$$

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

$$u(t_{k+1}) = \Big(1 + \lambda \Delta t_k\Big) u(t_k). \tag{1.1.6}$$

Starting with $k = 0$ *and the data* $u(0) = u_0$*, the solution* $u(t_k)$ *would, iteratively, be produced at the subsequent points:* $t_1,\ t_2, \dots, t_N = T$*.*
For a uniform partition, where all subintervals have the same length Δt*,* (1.1.6) *would be*

$$u(t_{k+1}) = \Big(1 + \lambda \Delta t\Big) u(t_k), \quad k = 0, 1, \dots, N-1. \tag{1.1.7}$$

There are corresponding finite difference methods for PDE's. Our goal, however, is to study the *Galerkin's 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 for finite dimensional linear spaces. Some of these examples are studied in details in the *interpolation chapter*.

I. $P^{(q)}(a,b) :=\{$ The space of polynomials in x of degree $\leq q,\ a\leq x\leq b\}$.

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

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

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

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

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

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 bases functions on an interval are introduced in the *interpolation chapter*: these are Lagrange bases $\{\lambda_i\}_{i=0}^q\in P^{(q)}(a,b)$ associated to a set of $(q+1)$ distinct points $\xi_0<\xi_1<\ldots<\xi_q$ in (a,b) determined by the requirement that

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

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

$$p(x)=p_0\lambda_0(x)+p_1\lambda_1(x)+\ldots+p_q\lambda_q(x). \tag{1.1.8}$$

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

Remark 3. *Our goal is to approximate general functions, rather than polynomials, by piecewise polynomials of Lagrange type. Then, for a given function f the Lagrange coefficients in* (1.1.8) *will be replaced by $p_i = f(\xi_i)$, $1 \leq i \leq q$, and $f(x)$ will be approximated by its* Lagrange *interpolant, viz*

$$f(x) \approx \sum_{i=0}^{q} f(\xi_i)\lambda_i(x) := \pi_q f(x). \tag{1.1.9}$$

We shall illustrate this in the next examples.

Example 10. *The linear Lagrange basis functions, $q = 1$, are given by (see Fig. 1.1.)*

$$\lambda_0(x) = \frac{\xi_1 - x}{\xi_1 - \xi_0} \qquad \text{and} \qquad \lambda_1(x) = \frac{x - \xi_0}{\xi_1 - \xi_0}. \tag{1.1.10}$$

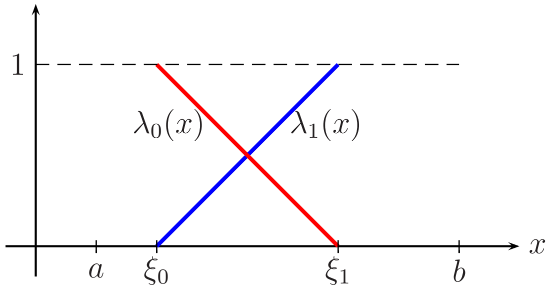


Figure 1.1: Linear Lagrange basis functions for $q = 1$.

Example 11. *A typical application of Lagrange bases is in finding a polynomial interpolant $\pi_q f \in P^q(a, b)$ of a continuous function $f(x)$ on an interval $[a, b]$. The procedure is as follows:*

Choose distinct interpolation nodes $a = \xi_0 < \xi_1 < \ldots < \xi_q = b$ and define $\pi_q f(\xi_i) = f(\xi_i)$. Then $\pi_q f \in P^{(q)}(a, b)$, definied as the sum in (1.1.9)*, interpolates $f(x)$ at the nodes $\{\xi_i\}$, $i = 0, \ldots, q$ and using Lagrange's formula* (1.1.8)*, with $p_i = f(\xi_i)$, $i = 0, 1, \ldots, q$, and $x \in [a, b]$ yields*

$$\pi_q f(x) = f(\xi_0)\lambda_0(x) + f(\xi_1)\lambda_1(x) + \ldots + f(\xi_q)\lambda_q(x).$$

For linear interpolant, i.e. $q = 1$, *we only need 2 nodes and 2 basis functions. Choosing* $\xi_0 = a$ *and* $\xi_1 = b$, *in* (1.1.10) *we get the linear interpolant*

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

where

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

i.e.,

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

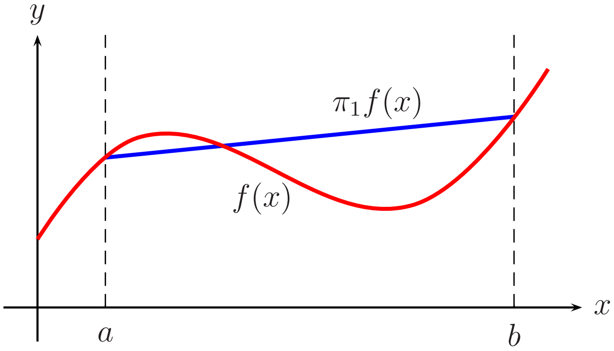


Figure 1.2: The linear interpolant $\pi_1 f(x)$ on a single interval.

V. We shall frequently use the space of *continuous piecewise polynomials* on a partition of an interval into some subintervals. For example 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 T_h. Obviously, $V_h^{(q)} \subset P^{(q)}(0,1)$. Let also

$$\overset{\circ}{V}{}_h^{(q)} = \{v : v \in V_h^{(q)}, \quad v(0) = v(1) = 0\}.$$

Our motivation in introducing these function spaces is due to the fact that these are function spaces, adequate in the numerical study of the

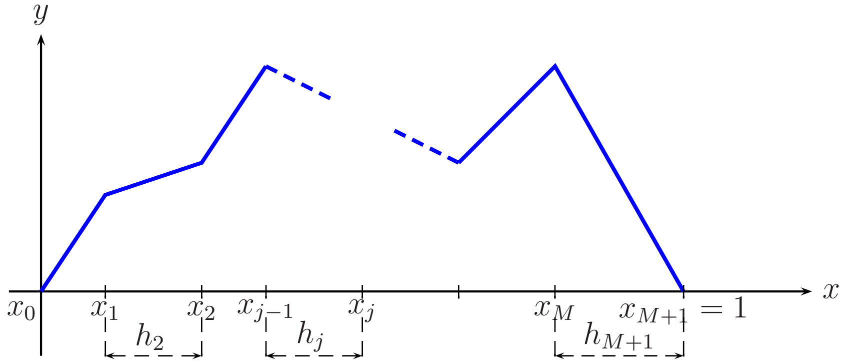


Figure 1.3: Fig shows an example of $\overset{\circ}{V}_h^{(1)}$.

boundary value problems using finite element methods for approximating solutions with piecewise polynomials.

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

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

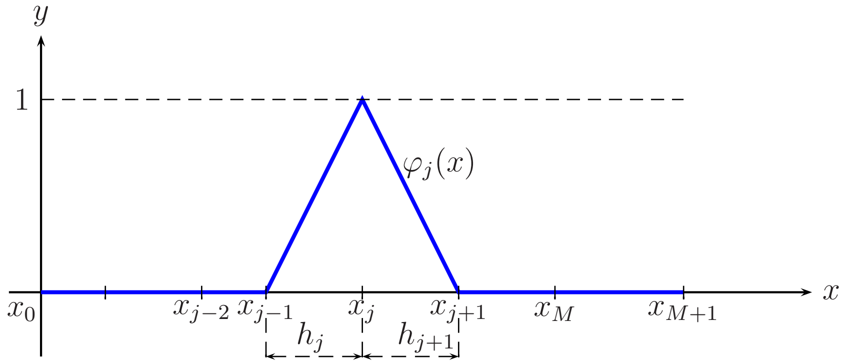


Figure 1.4: Fig shows a general piecewise linear basis function $\varphi_j(x)$.

Vector spaces

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

Definition 1. *A set of functions or vectors V is called a linear space, or a vector space, if for all $u, v \in V$ and all $\alpha \in \mathbb{R}$ (real numbers), we have*

$$\begin{array}{rl} \textit{(i)} & u + \alpha v \in V \\ \textit{(ii)} & u + v = v + u \\ \textit{(iii)} & \forall\, u \in V,\ \exists\,(-u) \in V \qquad \textit{such that} \quad u + (-u) = 0. \end{array} \tag{1.1.11}$$

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

Definition 2 (Scalar product)**.** *A scalar product is a real valued operator on $V \times V$, viz $(u, v) : V \times V \to \mathbb{R}$ such that for all u, v, $w \in V$ and all $\alpha \in \mathbb{R}$,*

$$\begin{array}{rll} \textit{(i)} & \langle u, v\rangle = \langle v, u\rangle, & \textit{(symmetry)} \\ \textit{(ii)} & \langle u + \alpha v, w\rangle = \langle u, w\rangle + \alpha\langle v, w\rangle, & \textit{(bi-linearity).} \end{array} \tag{1.1.12}$$

Definition 3. *A vector space W is called a scalar product space if W is associated with a scalar product $\langle\cdot,\cdot\rangle$, defined on $W \times W$.*

The function spaces $C([0, T])$, $C^k([0, T])$, , P^q, T^q are examples of scalar product spaces associated with usual scalar product defined by

$$\langle u, v\rangle = \int_0^T u(x)v(x)dx, \tag{1.1.13}$$

Definition 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 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*

$$\|u\| = \sqrt{\langle u, u\rangle} = \langle u, u\rangle^{\frac{1}{2}} = \Big(\int_0^T |u(x)|^2 dx\Big)^{\frac{1}{2}}. \tag{1.1.14}$$

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 through out this note: *The Cauchy-Schwarz inequality*,

$$|\langle u, v\rangle| \leq \|u\|\|v\|. \tag{1.1.15}$$

A simple proof of (1.1.15) 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

$$\begin{aligned} 0 \leq \langle u - av, u - av\rangle &= \|u\|^2 - 2a\langle u, v\rangle + a^2\|v\|^2 \\ &= \|u\|^2 - \frac{2\langle u, v\rangle}{\|v\|^4}\|v\|^2. \end{aligned}$$

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

$$\frac{\langle u, v\rangle^2}{\|v\|^2} \leq \|u\|^2,$$

which yields the desired result.

• Galerkin method for (IVP)

For a solution u of the initial value problem (1.1.1) we use test functions v, in a certain vector space V, for which the integrals below are well-defined,

$$\int_0^T \dot{u}(t)v(t)\,dt = \lambda \int_0^T u(t)v(t)\,dt, \qquad \forall v \in V, \tag{1.1.16}$$

or equivalently

$$\int_0^T \Big(\dot{u}(t) - \lambda\, u(t)\Big)v(t)dt = 0, \qquad \forall v(t) \in V, \tag{1.1.17}$$

i.e.

$$\Big(\dot{u}(t) - \lambda\, u(t)\Big) \perp v(t), \quad \forall v(t) \in V. \tag{1.1.18}$$

Definition 6. *If w is an approximation of u in the problem* (1.1.16)*, 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 the uniqueness we would get the exact solution ($w = u$). Our requirement is instead that w satisfies the equation (1.1.1) in average, or in other words we require that w satisfies (1.1.18), i.e,

$$\mathcal{R}\Big(w(t)\Big) \perp v(t), \quad \forall v(t) \in V. \tag{1.1.19}$$

In our case the real solution belongs to $C((0,T))$, or better to

$$H^s(0,T) := \{f : \sum_{k=0}^{s} \int_0^T \Big(\partial^k f/\partial t^k\Big)^2 dt < \infty\}.$$

H^s is a subspace of $C((0,T))$ consisting of all function in $L_2(0,T)$ having also all their derivatives of order $\leq s$ in $L_2(0,T)$. We look for a solution $U(t)$ in a finite dimensional subspace e.g. $V^{(q)}$. More specifically, we want to look at an *approximate solution* $U(t)$, called a trial function for (1.1.1) in the space of piecewise polynomials of degree $\leq q$:

$$V^{(q)} = \{U : U(t) = \xi_0 + \xi_1 t + \xi_2 t^2 + \ldots + \xi_q t^q\}. \tag{1.1.20}$$

Hence, to determine $U(x)$ 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 q linearly independent monomials t, t^2, $\ldots$, t^q. To this approach we define the *test functions space*:

$$\overset{\circ}{V}^{(q)} = \{v \in V^{(q)} : v(0) = 0\}, \tag{1.1.21}$$

in other words v can be written as $v(t) = \xi_1 t + \xi_2 t^2 + \ldots + \xi_q t^q$. Note that

$$\overset{\circ}{V}^{(q)} = span[t,\, t^2,\, \ldots,\, t^q]. \tag{1.1.22}$$

For an approximate solution U we require that its residual $R(U)$ satisfy the orthogonality condition (1.1.19):

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

Thus the *Galerkin method for* (1.1.1) *is formulated* as follows:
Given $u(0) = u_0$, find the approximate solution $U(t) \in V^{(q)}$, for (1.1.1) such that (for simplicity we put $T \equiv 1$)

$$\int_0^1 \mathcal{R}\Big(U(t)\Big)v(t)dt = \int_0^1 (\dot{U}(t) - \lambda\, U(t))v(t)dt = 0,\, \forall v(t) \in \overset{\circ}{V}^{(q)}. \qquad (1.1.23)$$

Formally this can be obtained writing a *wrong!!!* equation by replacing u by $U \in V^{(q)}$ in (1.1.1),

$$\begin{cases} \dot{U}(t) = \lambda\, U(t), & 0 < t < 1 \\ U(0) = u_0, \end{cases} \qquad (1.1.24)$$

then, multiplying (1.1.24) by a function $v(t) \in \overset{\circ}{V}^{(q)}$ from the test function space and integrating over $[0, 1]$.

Now since $U \in V^{(q)}$ we can write $U(t) = u_0 + \sum_{j=1}^{q} \xi_j t^j$, then $\dot{U}(t) = \sum_{j=1}^{q} j\xi_j t^{k-1}$. Further we have $\overset{\circ}{V}^{(q)}$ is spanned by $v_i(t) = t^i, i = 1, 2, \dots, q$. Inserting these representations for U, $\dot{U}$ and $v = v_j$, $j = 1, 2, \dots, q$ in (1.1.22) we get

$$\int_0^1 \Big(\sum_{j=1}^{q} k\xi_j t^{j-1} - \lambda u_0 - \lambda \sum_{j=1}^{q} \xi_j t^j\Big) \cdot t^i dt = 0, \quad i = 1, 2, \dots, q, \qquad (1.1.25)$$

which can be rewritten as

$$\int_0^1 \Big(\sum_{j=1}^{q} (j\xi_j t^{i+j-1} - \lambda\, \xi_j t^{i+j})\Big) dt = \lambda u_0 \int_0^1 t^i dt. \qquad (1.1.26)$$

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

$$\sum_{j=1}^{q} \xi_j \Big[j \cdot \frac{t^{i+j}}{i+j} - \lambda \frac{t^{i+j+1}}{i+j+1}\Big]_{t=0}^{t=1} = \Big[\lambda \cdot u_0 \frac{t^{i+1}}{i+1}\Big]_{t=0}^{t=1}, \qquad (1.1.27)$$

or equivalently

$$\sum_{j=1}^{q} \Big(\frac{j}{i+j} - \frac{\lambda}{i+j+1}\Big)\xi_j = \frac{\lambda}{i+1} \cdot u_0 \quad i = 1, 2, \dots, q, \qquad (1.1.28)$$

which is a linear system of equations with q equations and q unknowns $(\xi_1,\, \xi_2,\, \dots,\, \xi_q)$; given in the matrix form as

$$\mathcal{A}\Xi = \mathbf{b}, \quad \text{with} \quad \mathcal{A} = (a_{ij}), \quad \Xi = (\xi_j)_{j=1}^q, \quad \text{and} \quad \mathbf{b} = (b_i)_{i=1}^q. \qquad (1.1.29)$$

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 two last rows (columns) of $\mathcal{A}$ given by $a_{ij} = \dfrac{j}{i+j} - \dfrac{\lambda}{i+j+1}$, are very close to each others resulting to extreme 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 would yield a diagonal (sparse) coefficient matrix and make the problem well conditioned. This however, is a rather tedious task.

Galerkin's method and orthogonal projection

Let $u = (u_1, u_2, u_3) \in \mathbb{R}^3$ and assume that for some reasons we only have u_1 and u_2 available. Letting $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, the objective, then is to find $U \in \{x : x_3 = 0\}$, such that $(u - U)$ is as small as possible. For orthogonal projection we have $z \cdot n = 0$, for all $z \in \{x : x \cdot \mathbf{n} = 0, x_3 = 0\}$, where $\mathbf{n}$ is the normal vector to the plane $\{(x_1, x_2, 0)\}$. Obviously in this case $U = (u_1, u_2, 0)$ and we have $(u - U) \perp U$.

Note that, if $m < n$, and $\mathbf{u}_m$ is the projection of $\mathbf{u} = (u_1, u_2, \dots, u_{n-1}, u_n)$ on $\mathbb{R}^m$, then $\mathbf{u}_m = (u_1, u_2, \dots, u_m, u_{m+1} = 0, \dots, u_n = 0)$, and the Euclidean distance: $|\mathbf{u} - \mathbf{u}_m| = \sqrt{u_{m+1}^2 + u_{m+2}^2 + \dots + u_n^2} \to 0$ as $m \to n$. This meams the obvious fact that the accuracy of the orthogonal projection will improve by raising the dimension of the projection space.

• Galerkin method for (BVP)

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

$$\begin{cases} -\frac{d}{dx}\Big(a(x) \cdot \frac{d}{dx}u(x)\Big) = f(x), & \quad 0 < x < 1; \\ u(0) = u(1) = 0. \end{cases} \qquad (1.1.30)$$

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

$$-u''(x) = f(x), \quad 0 < x < 1; \qquad u(0) = u(1) = 0. \qquad (1.1.31)$$

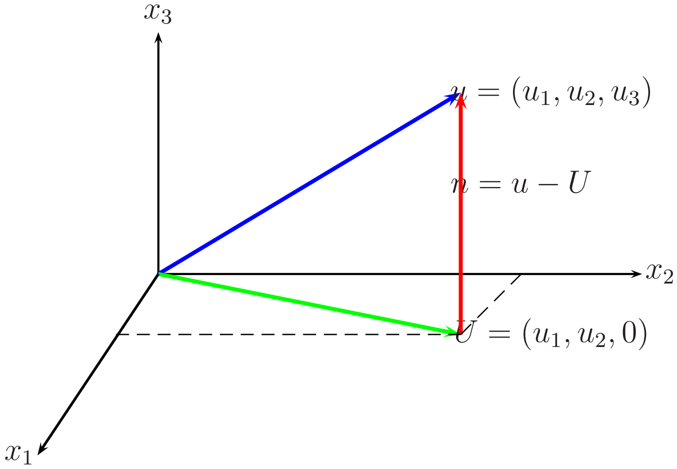


Figure 1.5: Example of a projection on $\mathbb{R}^2$.

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

$$V_h^0 = \{v \in \mathcal{C}(0,1) : v \,\text{is piecewise linear function on}\, T_h, \,\text{and}\, v(0) = v(1) = 0\},$$

with the bases 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 basis functions corresponding to the values at $x_0 = 0$ and $x_{M+1} = 1$.

Remark 4. *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*

$$-u''(x) = f(x), \quad 0 < x < 1; \qquad u(0) = b_0, \quad u'(1) = b_1, \tag{1.1.32}$$

then the corresponding basis 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 at $x_{M+1} = 1$ (dashed in the Fig below).

Now the *Galerkin method* for problem (1.1.31) (which is just the description of the orthogonality condition of the residual $R(U) = -U'' - f$ to the

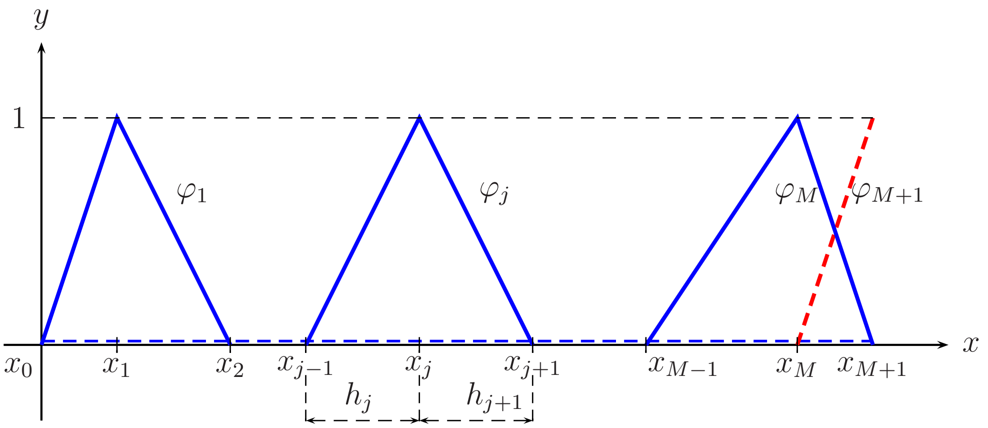


Figure 1.6: General piecewise linear basis functions

test function space V_h^0; i.e., $R(U) \perp V_h^0$) is formulated as follows: Find the approximate solution $U(x) \in V_h^0$ such that

$$\int_0^1 (-U''(x) - f(x))v(x)dx = 0, \qquad \forall v(x) \in V_h^0 \tag{1.1.33}$$

Observe that if $U(x) \in \overset{\circ}{V}_h$, then $U''(x)$ is either equal to zero or is not a well-defined function, in the latter case, the equation (1.1.33) does not make sense, whereas for $U''(x) = 0$ and the data $U(0) = U(1) = 0$ we get the trivial approximation $U(x) \equiv 0$. This however, is relevant only if $f \equiv 0$, but then even $u(x) \equiv 0$ and we have a trivial case. If, however, we perform partial integration then

$$-\int_0^1 U''(x)v(x)dx = \int_0^1 U'(x)v'(x)dx - [U'(x)v(x)]_0^1 \tag{1.1.34}$$

and since $v(x) \in \overset{\circ}{V}_h$; $v(0) = v(1) = 0$, we get

$$-\int_0^1 U''(x)v(x)dx = \int_0^1 U'(x)v'(x)\, dx \tag{1.1.35}$$

Now for $U(x), v(x) \in \overset{\circ}{V}_h$, $U'(x)$ and $v'(x)$ are well-defined (except at the nodes) and the equation (1.1.33) has a meaning.

Hence, *The Galerkin finite element method* (FEM) for the problem (1.1.30) is now reduced to: Find $U(x) \in V_h^0$ such that

$$\int_0^1 U'(x)v'(x)\,dx = \int_0^1 f(x)v(x)dx, \qquad \forall v(x) \in V_h^0. \tag{1.1.36}$$

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

$$U(x) = \sum_{j=1}^{M} \xi_j\,\varphi_j(x) \quad \text{which implies that} \quad U'(x) = \sum_{j=1}^{M} \xi_j \varphi'(x). \tag{1.1.37}$$

Thus we can write (1.1.36) as

$$\sum_{j=1}^{M} \xi_j \int_0^1 \varphi_j'(x)\,v'(x)dx = \int_0^1 f(x)v(x)dx, \qquad \forall v(x) \in V_h^0. \tag{1.1.38}$$

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

$$\sum_{j=1}^{M} \Big(\int_0^1 \varphi_i'(x) \cdot \varphi_j'(x)dx \Big) \xi_j = \int_0^1 f(x)\varphi_i(x)dx, \quad i = 1, 2, \dots, M. \tag{1.1.39}$$

This equation can be written in the equivalent matrix form as

$$\mathbf{A}\xi = \mathbf{b}. \tag{1.1.40}$$

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

$$\mathbf{A} = \{a_{ij}\}_{i,j=1}^{M}, \quad a_{ij} = \int_0^1 \varphi_j'(x) \cdot \varphi_i'(x)dx, \tag{1.1.41}$$

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_M \end{pmatrix}, \text{ with } \quad b_j = \int_0^1 f(x)\varphi_i(x)dx, \text{ and } \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \dots \\ \xi_M \end{pmatrix}. \tag{1.1.42}$$

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

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

Stiffness matrix A:

If $\underline{|i-j| > 1}$, then φ_i and φ_j have disjoint non-overlapping supports, evideltly, we hence

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

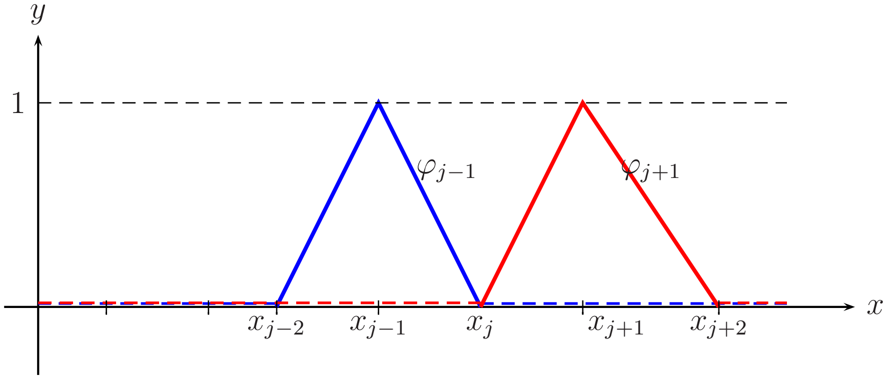


Figure 1.7: φ_{j-1} and φ_{j+1}.

As for $i = j$: we have that

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

It remains to compute a_{ij} for the case $j = i+1$: A straightforward calculation (see the fig below) yields

$$a_{i,i+1} = \int_{x_i}^{x_{i+1}} \Big(-\frac{1}{h_{i+1}}\Big) \cdot \frac{1}{h_{i+1}} dx = -\frac{x_{i+1} - x_i}{h_{i+1}^2} = -\frac{1}{h_{i+1}}. \tag{1.1.43}$$

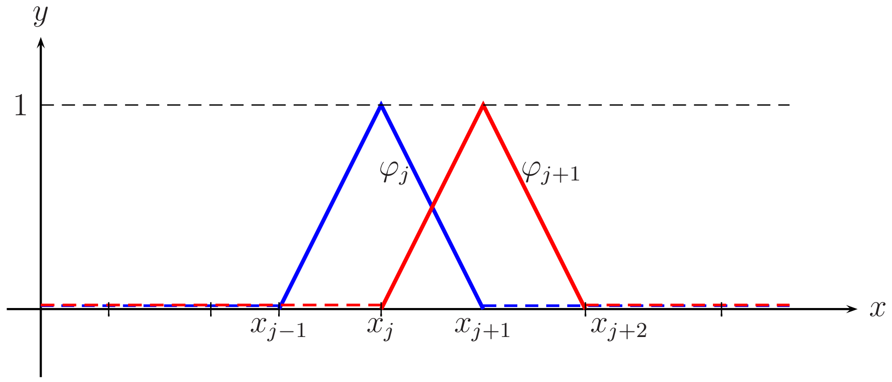


Figure 1.8: φ_j and φ_{j+1}.

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

To *summarize*, we have

$$\begin{cases} a_{ij} = 0, & \text{if } |i-j| > 1, \\ a_{ii} = \frac{1}{h_i} + \frac{1}{h_{i+1}}, & i = 1, 2, \dots, M, \\ a_{i-1,i} = a_{i,i-1} = -\frac{1}{h_i}, & i = 2, 3, \dots, M. \end{cases} \tag{1.1.44}$$

Thus by symmetry we finally have the stiffness matrix for the stationary heat conduction as:

$$\mathbf{A} = \begin{bmatrix} \frac{1}{h_1} + \frac{1}{h_2} & -\frac{1}{h_2} & 0 & \dots & 0 \\ -\frac{1}{h_2} & \frac{1}{h_2} + \frac{1}{h_3} & -\frac{1}{h_3} & 0 & 0 \\ 0 & \dots & \dots & \dots & 0 \\ \dots & 0 & \dots & \dots & -\frac{1}{h_M} \\ 0 & \dots & 0 & -\frac{1}{h_M} & \frac{1}{h_M} + \frac{1}{h_{M+1}} \end{bmatrix} \tag{1.1.45}$$

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

$$\mathbf{A}_{unif} = \frac{1}{h} \cdot \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \dots \\ 0 & -1 & 2 & -1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix} \tag{1.1.46}$$

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 (1.1.31):

$$-u''(x) = f(x), \qquad 0 \le x \le 1; \tag{1.1.47}$$

and motivate for its boundary conditions. The equation (1.1.47) is linear for the unknown function u, with inhomogeneous 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 uncertainity left by these two arbitrary constants C and D will be removed by adding a *boundary condition* at each end point of the interval

$$u(0) = 0, \qquad u(1) = 0. \tag{1.1.48}$$

The result is a *two-point boundary-value problem*, describing not a transient but a steady-state phenomenon–the temperature distribution in a rode, for example with ends fixed at $\overset{\circ}{0}$ 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 it values at equally spaced points $x = h$, $x = 2h, \dots, x = nh$. 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 ends $x = 0$ and $x = 1 = (n+1)h$, 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 a stopping at a finite stepsize, 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 there is just one combination that uses the values at x and $x \pm h$:

$$\frac{d^2u}{d^2x} \approx \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}. \tag{1.1.49}$$

It also has the merit of being symmetric about x. (1.1.49) is obtained using

$$\frac{d^2u}{d^2x} \approx \frac{u'(x) - u'(x-h)}{h}. \tag{1.1.50}$$

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 (1.1.49) we get

$$\begin{aligned}\frac{d^2u}{d^2x} &\approx \frac{(u(x+h) - u(x))/h - (u(x) - u(x-h))/h}{h} \\ &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}.\end{aligned} \tag{1.1.51}$$

To repeat the right side approaches the true value of d^2u/dx^2 as $h \to 0$, but have to stop at a positive h.

At a typical meshpoint $x = jh$, the differential equation $-d^2u/dx^2 = f(x)$ is now replaced by this discrete analogue (1.1.51); after multiplying by h^2,

$$-u_{j+1} + 2u_j - u_{j-1} = h^2 f(jh). \tag{1.1.52}$$

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 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 (1.1.52) 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 loss $h^2 f(jh)$ at the center.

The structure of the n equations (1.1.52) can be better visualized in matrix form $Au = b$ viz

$$\begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \dots \\ 0 & -1 & 2 & -1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} = h^2 \begin{bmatrix} f(h) \\ f(2h) \\ f(3h) \\ \cdot \\ \cdot \\ f(nh) \end{bmatrix}, \qquad (1.1.53)$$

which, once again, gives the structure of our uniform stifness matrix A_{unif} given in (1.1.46).

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

Remark 5. *Unlike the matrix $\mathcal{A}$ for monomial approximation of IVP in* (1.1.28)*, $\mathbf{A}$ has 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 7. *A matrix A is called positive definite if*

$$\forall \eta \in R^M, \ \eta \neq 0, \ \eta^T A \eta > 0, \qquad i.e. \quad \sum_{i,j=1}^M \eta_i a_{ij} \eta_j > 0. \qquad (1.1.54)$$

We shall use the positive definiteness of A to argue that (1.1.40) is uniquely solvable. To this approach we prove the following well-known result:

Proposition 1. *If a square matrix $\mathbf{A}$ is positive definite then $\mathbf{A}$ is invertible and hence $\mathbf{A}\xi = b$ 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 equation system. Finally we illustrate an example of the positive-definiteness argument for $\mathbf{A}_{unif}$.

Example 12. *Assume $M = 2$ and let $U(x,y) = \begin{pmatrix} x \\ y \end{pmatrix}$, then*

$$\begin{aligned} U^T\mathbf{A}_{unif}U &= (x,y)\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = (x,y)\begin{pmatrix} 2x-y \\ -x+2y \end{pmatrix} \\ &= 2x^2 - xy - xy + 2y^2 = x^2 + y^2 + x^2 - 2xy + y^2 \\ &= x^2 + y^2 + (x-y)^2 \geq 0. \end{aligned} \tag{1.1.55}$$

Thus $\mathbf{A}_{unif}$ is positive definite. Since $U^TAU = 0$ only if $x = y = 0$ i.e. $U = 0$.

Remark 6. *For a boundary value problem with, e.g. inhomogeneous Dirichlet boundary data, actually a direct approach would be with the test and trial function spaces having different dimensions; test functions are zero at the boundary: $v \in \overset{\circ}{V}_h$, trial function: $u \in V_h$ (not necessarily zero at the boundary). This would yield a linear equation system $\mathcal{A}\Xi = b$ with a rectangular matrix A instead of a quadratic one. Then to continue we use the least square method and instead solve $A^TA\Xi = A^Tb$. The solution Ξ is approximate and $A\Xi \neq b$. Thus, the corresponding orthogonality condition of the residual to the test function space is now $r := (A\Xi - b) \perp C_j$, where C_j are columns in A.*

Summary: Roughly speaking, a systematic procedure for approximate solution for a differential equations 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 the next chapter.

2. The function $f(x)$ is unknown and the elements of the vector $\mathbf{b}$ as well as the integrals that represent the elements of the coefficient matrix are of involve character, for example when approximating by higher order polynomials and/or solving equations with variable coefficients. Therefore we need to approximate different integrals over subintervals of a partition. This may be done using *Gauss quadrature rules*. In simple case one may use usual or composite *midpoint-*, *trapezoidal-*, or *Simpson's-rules*. In more involved cases one may employ *Gauss quadrature rules*. We shall briefly introduce the idea of Gauss quadrature rule in the next chapter.

3. Finally we end up with linear systems of equations (LSE) of type (1.1.40). To solve LSE efficiently we may use exact *Gauss - elimination* or the iteration procedures as *Gauss-Seidel*, *Gauss-Jacobi* or *Over-relaxation methods*. We discuss these concepts in the chapter of the numerical linear algebra.

1.2 Exercises

Problem 1. *Use the method of least squares to solve the following systems of linear equations.*

$$a.\ \begin{cases} -x_1 + x_2 = 16 \\ 2x_1 + x_2 = -9 \\ x_1 - 2x_2 = -12 \end{cases} \qquad b.\ \begin{cases} x_1 + x_2 = 3 \\ -2x_1 + 3x_2 = 1 \\ 2x_1 - x_2 = 2 \end{cases}$$

$$c.\ \begin{cases} x_1 + 2x_2 = 3 \\ -2x_1 + x_2 = -4 \\ x_1 - 3x_2 = -2 \\ -x_1 + x_2 = -1 \\ 2x_1 + x_2 = 5 \end{cases} \qquad d.\ \begin{cases} x_1 + x_2 + x_3 = 4 \\ -x_1 + x_2 + x_3 = 0 \\ -x_2 + x_3 = 1 \\ x_1 + x_3 = 2 \end{cases}$$

$$e.\ \begin{cases} x_1 + x_2 + x_3 = 7 \\ x_1 + x_2 - x_3 = -1 \\ x_1 - x_2 + x_3 = 1 \\ x_1 - x_2 - x_3 = 3 \end{cases}$$

Problem 2. *Determine the line $y = b + ct$ that fits the following pairs of data (t, y) best.*

a.

t	1	2	3	4	5
y	1	5	2	7	10

b.

t	1	2	3	4	5
y	5	6	10	12	17

c.

t	*1*	*2*	*3*	*4*	*5*
y	*2*	*3*	*1*	*1*	*-2*

Problem 3. *Determine the parameters* a *and* b *such that the parabolic curve* $y = ax^2 + bx + c$ *fits the following values of* x *and* y *best in the least squares sense.*

a.

x	*-2*	*-1*	*0*	*1*	*2*
y	*2*	*1*	*1*	*2*	*3*

b.

x	*-1*	*0*	*1*	*2*
y	*2*	*2*	*1*	*0*

Problem 4. *Let* x *be the solution of the least squares problem* $Ax \approx b$*, where*

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

Let $r - b - Ax$ *be the corresponding residual. Which of the following three vectors is a possible value for* r*?*

$$a. \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad b. \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix} \qquad c. \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix}$$

Problem 5. *Set up and solve the linear least squares system* $Ax \approx b$ *for fitting the model function* $f(t, x) = x_1 t + x_2 e^t$ *to the three data points* $(1, 2)$ $(2, 3)$ *and* $(3, 5)$*.*

Problem 6. *True or false: At the solution to a linear least squares problem* $Ax \approx b$*, the residual vector* $r = b - Ax$ *is orthogonal to the column space of* A*.*

Problem 7. *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 8. *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 9. *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*

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

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

Problem 10. *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 $U(x) = A\sin x + B\sin 2x + C\sin 3x + \frac{2}{\pi^2}x^2$.

Chapter 2

Polynomial Interpolation in 1d

2.1 Preliminaries

We recall the idea of polynomial interpolation. Consider a real-valued function f, defined on an interval $I = [a, b]$, and a partition

$$\mathcal{T}_h : a = x_0 < x_1 < \ldots < x_{M+1} = b,$$

of I into $M + 1$ subintervals $I_j = [x_{j-1}, x_j]$, $j = 1, \ldots, M + 1$.

Definition 8. *An interpolant $\pi_q f$ of f on the partition $\mathcal{T}_h$ is a piecewise polynomial function of degree $\leq q$, having the nodal values at x_j, $j = 1, \ldots, M+1$, coinciding with those of f: $\pi_q f(x_j) = f(x_j)$.*

Here are some simple examples:

Linear interpolation on an interval. We start with the unit interval $I = [0, 1]$, without any partitions, and a function $f : [0, 1] \to \mathbb{R}$, which is Lipschitz continuous. We let $q = 1$ and seek the linear interpolation 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 $\pi_1 f \in \mathcal{P}^1$,

$$\pi_1 f(x) = C_0 + C_1 x, \qquad x \in I, \tag{2.1.1}$$

where

$$\begin{aligned} \pi_1 f(0) = f(0) \quad &\Longrightarrow \quad C_0 = f(0), \\ \pi_1 f(1) = f(1) \quad &\Longrightarrow \quad C_0 + C_1 = f(1) \Longrightarrow C_1 = f(1) - f(0). \end{aligned} \tag{2.1.2}$$

Inserting C_0 and C_1 in (2.1.1) 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 basis:

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

and

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

Note that the functions $\lambda_0(x) = 1-x$ and $\lambda_1(x) = x$ are linearly independent. Since we can easily see that, if

$$0 = \alpha_0(1-x) + \alpha_1 x = \alpha_0 + (\alpha_1 - \alpha_0)x, \qquad x \in I, \tag{2.1.3}$$

then

$$\left.\begin{array}{lll} x = 0 & \Longrightarrow & \alpha_0 = 0 \\ x = 1 & \Longrightarrow & \alpha_1 = 0 \end{array}\right\} \Longrightarrow \alpha_0 = \alpha_1 = 0. \tag{2.1.4}$$

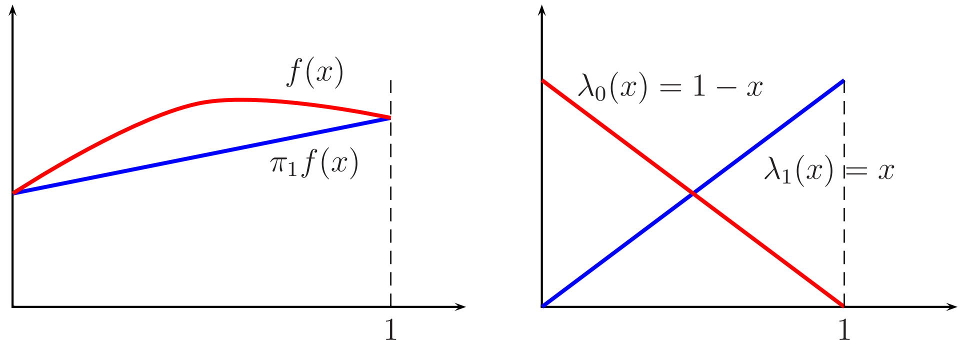


Figure 2.1: Linear interpolation and basis functions for $q = 1$.

Remark 7. *Note that if we define a scalar product on $\mathcal{P}^k(a,b)$ by*

$$(p,q) = \int_a^b p(x)q(x)\,dx, \qquad \forall p,q \in \mathcal{P}^k(a,b), \tag{2.1.5}$$

then neither $\{1,x\}$ nor $\{1-x,x\}$ are orthogonal in the interval $[0,1]$. For example, $(1,x) := \int_0^1 1\cdot x\,dx = [\frac{x^2}{2}] = \frac{1}{2} \neq 0$.

Now it is natural to ask the following question.

Ouestion 1. *What will be the error, if one approximates $f(x)$ by $\pi_1 f(x)$? In other words: what is $f(x) - \pi_1 f(x) =$?*

To answer this question, we need to have a measuring instrument to quantify the difference. Grafically (geometrically), the deviation of $f(x)$ from $\pi_1 f(x)$ (from at being linear) depends on the *curvature* of $f(x)$, i.e. on how *curved* $f(x)$ is. In other words how *large* is $f''(x)$ on say (a, b)? To this end below we introduce some measuring environments for vectors and functions:

Definition 9 (Vector norm)**.** *Let* $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ *be two column vectors. 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 $\mathbf{x}$*:* $\|\mathbf{x}\|$ *as*

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

$L_p(a, b)$**-norm:** *Assume that f is a real valued function such that the integrals as well as the* max *on the right hand sides below are well-defined. Then we define the L_p-norm ($1 \le p \le \infty$) of f as*

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

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

Now, we want to see how far can one answer the question 1 in the L_p-norm environment?

Theorem 1. *(L_∞-error estimates for the 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 $f(x)$ 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 on $[a, b]$ can be written as a linear combination of the basis functions $\lambda_a(x)$ and $\lambda_b(x)$ where

$$\lambda_a(x) = \frac{x-b}{a-b} \quad \text{and} \quad \lambda_b(x) = \frac{x-a}{b-a}. \tag{2.1.6}$$

We point out that linear combinations of $\lambda_a(x)$ and $\lambda_b(x)$ gives the basis functions $\{1, x\}$:

$$\lambda_a(x) + \lambda_b(x) = 1, \qquad a\lambda_a(x) + b\lambda_b(x) = x. \tag{2.1.7}$$

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

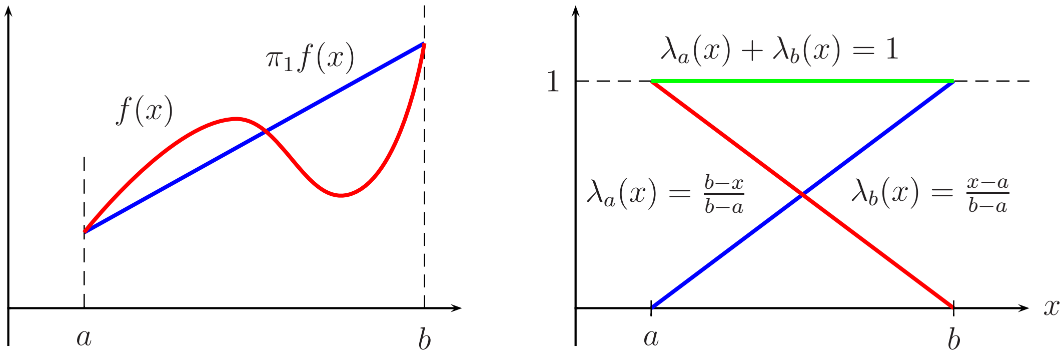

$\pi_1 f(x)$
$f(x)$
a
b
$\lambda_a(x) + \lambda_b(x) = 1$
1
$\lambda_a(x) = \frac{b-x}{b-a}$
$\lambda_b(x) = \frac{x-a}{b-a}$
x
a
b

Figure 2.2: Linear Lagrange basis functions for $q = 1$.

which can be represented by

$$\pi_1 f(x) = f(a)\lambda_a(x) + f(b)\lambda_b(x). \tag{2.1.8}$$

By the Taylor's expansion for $f(a)$ and $f(b)$ about x we can write

$$\begin{aligned} f(a) &= f(x) + (a-x)f'(x) + \frac{1}{2}(a-x)^2 f''(\eta_a), \quad \eta_a \in [a, x] \\ f(b) &= f(x) + (b-x)f'(x) + \frac{1}{2}(b-x)^2 f''(\eta_b), \quad \eta_b \in [x, b]. \end{aligned} \tag{2.1.9}$$

Inserting $f(a)$ and $f(b)$ from (2.1.9) in (2.1.8), 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 and using (2.1.7) (the identity $(a-x)\lambda_a(x)+(b-x)\lambda_b(x)=0$) we get

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

and consequently

$$|\pi_1 f(x)-f(x)| = \Big|\frac{1}{2}(a-x)^2 f''(\eta_a)\lambda_a(x)+\frac{1}{2}(b-x)^2 f''(\eta_b)\lambda_b(x)\Big|. \quad (2.1.10)$$

Now, for $a \le x \le b$ we have $(a-x)^2 \le (a-b)^2$ and $(b-x)^2 \le (a-b)^2$, furthermore $\lambda_a(x) \le 1$, $\lambda_b(x) \le 1$. Finally, 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 (2.1.10) can be estimated as

$$|\pi_1 f(x)-f(x)| \le \frac{1}{2}(a-b)^2\cdot 1\cdot\|f''\|_{L^\infty(a,b)}+\frac{1}{2}(a-b)^2\cdot 1\cdot\|f''\|_{L^\infty(a,b)}, \quad (2.1.11)$$

and hence

$$|\pi_1 f(x)-f(x)| \le (a-b)^2\|f''\|_{L^\infty(a,b)} \quad \text{with} \quad c_i = 1. \quad (2.1.12)$$

The other two estimates are proved similarly. □

Theorem 1 can be proved in a more general setting, for an arbitrary subinterval of $I=(a,b)$, in L_p-norm, $1 \le p \le \infty$. This, general version (concisly stated below as Theorem 2), is the mainly used L_p-interpolation error estimate. The proof is however, just a scaling of the argument used in the proof of Theorem 1.

Remark 8. *For a uniform partition $\mathcal{T}_h : a = x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} = b$ of the interval $[a,b]$ with mesh parameter $h = |x_{j+1}-x_j|$, $j = 0,1,\ldots,n$, it is customary to denote the interpolation function by $\pi_h v(x)$ rather than $\pi_q v(x)$. Here the subindex h refers to the mesh size h, and not to the degree of interpolating polynomial q. The degree q and the meaning of the notation used for the interpolant will be clear from the context.*

Theorem 2. *Let $\pi_h v(x)$ be the piecewise linear interpolant of the function $v(x)$ on the partition $\mathcal{T}_h$. That is $\pi_h v(x_j) = v(x_j)$, for $j = 0, 1, \dots, N+1$. Then, assuming that v is sufficiently regular ($v \in \mathcal{C}^2(a,b)$), there are interpolation constants c_i such that*

$$\begin{aligned} \|\pi_h v - v\|_{L_p} &\le c_i \|h^2 v''\|_{L_p}, \qquad p = 1, 2, \dots, \infty, && (2.1.13)\\ \|(\pi_h v)' - v'\|_{L_p} &\le c_i \|h v''\|_{L_p}, && (2.1.14)\\ \|\pi_h v - v\|_{L_p} &\le c_i \|h v'\|_{L_p}. && (2.1.15) \end{aligned}$$

For $p = \infty$ this is just the previous theorem, applied to each subinterval. Note that for a uniform mesh we have h constant and therefore in this case h can be written outside the norms on the right hand sides..

Proof. This is a straightforward generalization of the proof of the Theorem 1 and left as an excercise. □

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 $\mathcal{T}_h : a = x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} = b$ be a partition of I into subintervals $I_j = (x_{j-1}, x_j)$ of length $h_j = |I_j| := x_j - x_{j-1}$; $j = 1, 2, \dots, N$. Let

$$V_h := \{v | v \text{ is continuous piecewise linear function on} \mathcal{T}_h\}, \qquad (2.1.16)$$

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. It is easy to show that every function in V_h is a linear combination of φ_js:

Lemma 1. *We have that*

$$\forall v \in V_h; \qquad v(x) = \sum_{j=0}^{N} v(x_j)\varphi_j(x), \quad \Longrightarrow \Big(dim V_h = N + 1\Big). \qquad (2.1.17)$$

Proof. Both 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

$$\begin{aligned} RHS =& v(x_0)\varphi_0(x_j) + v(x_1)\varphi_1(x_j) + \ldots + v(x_{j-1})\varphi_{j-1}(x_j)\\ &\quad + v(x_j)\varphi_j(x_j) + v(x_{j+1})\varphi_{j+1}(x_j) + \ldots + v(x_N)\varphi_N(x_j) \qquad (2.1.18)\\ =& v(x_j) = LHS, \end{aligned}$$

where we have used the fact that φ is piecewise linear and $\varphi_i(x_j) = \delta_{ij}$. □

Definition 10. *Assume that f is a Lipschitz continuous function in $[a, b]$. Then the continuous piecewise linear interpolant of f is defined by*

$$\pi_h f(x_j) = f(x_j), \qquad j = 0, 1, \dots, N. \tag{2.1.19}$$

Or, alternatively:

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

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

(i) $\pi_h f(x)$ is linear on I_j, $\Longrightarrow \quad \pi_h f(x) = c_0 + c_1 x$ on I_j.

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

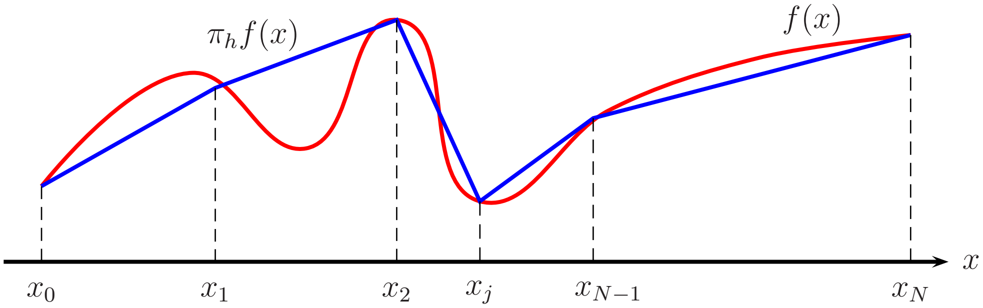


Figure 2.3: Piecewise linear interpolant $\pi_h f(x)$ of $f(x)$.

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

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

Consequently we may write

$$\begin{cases} c_0 = f(x_{j-1}) \frac{x_j}{x_j - x_{j-1}} + f(x_j) \frac{-x_{j-1}}{x_j - x_{j-1}} \\ c_1 x = f(x_{j-1}) \frac{-x}{x_j - x_{j-1}} + f(x_j) \frac{x}{x_j - x_{j-1}}. \end{cases} \tag{2.1.20}$$

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

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

where $\lambda_{j-1}(x)$ and $\lambda_j(x)$ are the piecewise linear basis functions on I_j:

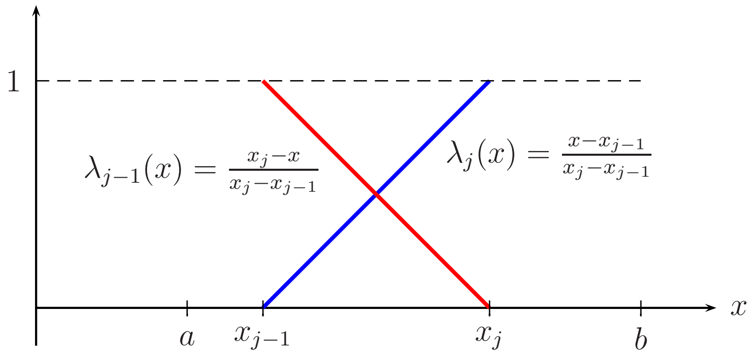

1
$\lambda_{j-1}(x) = \frac{x_j - x}{x_j - x_{j-1}}$
$\lambda_j(x) = \frac{x - x_{j-1}}{x_j - x_{j-1}}$
x
a
x_{j-1}
x_j
b

Figure 2.4: Linear Lagrange basis functions for $q = 1$ on the subinterval I_j.

We generalize the above procedure and introduce Lagrange interpolation bases functions:

2.2 Lagrange interpolation

Consider $P^q(a, b)$; the vector space of all polynomials of degree $\le q$ on the interval (a, b), and the basis functions $1, x, x^2, \dots, x^q$ for P^q.

Definition 11 (Cardinal functions)**.** *Lagrange basis is the set of polynomials $\{\lambda_i\}_{i=0}^q \subset P^q(a, b)$ associated to the $(q+1)$ distinct points, $a = x_0 < x_1 < \dots < x_q = b$, in (a, b) and determined by the requirement that $\lambda_i(x_j) = 1$ for $i = j$, and 0 otherwise, i.e. for $x \in (a, b)$,*

$$\lambda_i(x) = \frac{(x - x_0)(x - x_1)\dots(x - x_{i-1}) \downarrow (x - x_{i+1})\dots(x - x_q)}{(x_i - x_0)(x_i - x_1)\dots(x_i - x_{i-1}) \uparrow (x_i - x_{i+1})\dots(x_i - x_q)}. \quad (2.2.1)$$

By the arrows $\downarrow$, $\uparrow$ *in* (2.2.1) *we want to emphasize that* $\lambda_i(x) = \prod_{j\neq i}\Big(\frac{x-x_j}{x_i-x_j}\Big)$ *does not contain the singular factor* $\frac{x-x_i}{x_i-x_i}$. *Hence*

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

and $\lambda_i(x)$, $i = 1, 2, \dots, q$, *is a polynomial of degree* q *on* (a, b) *with*

$$\lambda_i(x_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{2.2.2}$$

Example 13. *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 $p \in P^q(a, b)$ with the values $p_i = p(x_i)$ at the nodes x_i, $i = 0, 1, \dots, q$, can be expressed in terms of the corresponding Lagrange basis as

$$p(x) = p_0\lambda_0(x) + p_1\lambda_1(x) + \dots + p_q\lambda_q(x). \tag{2.2.3}$$

Using (2.2.2), $p(x_i) = p_0\lambda_0(x_i) + p_1\lambda_1(x_i) + \dots p_i\lambda_i(x_i) + \dots + p_q\lambda_q(x_i) = p_i$. 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 $(\xi_0, \xi_1) \subset (a, b)$.

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

Definition 12 (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*

$$f(x) = T_q f(x_0) + R_q f(x_0), \tag{2.2.4}$$

where

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

is the Taylor polynomial of degree q, approximating f about $x = x_0$ and

$$R_qf(x) = \frac{1}{(q+1)!}f^{(q)}(\xi)(x-x_0)^{q+1}, \tag{2.2.5}$$

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

For a continuous function $f(x)$ on $[a, b]$, we define the *Lagrange interpolation polynomial* $\pi_qf \in P^q(a, b)$, corresponding to the Taylor poynomial $T_qf(x)$.

Definition 13. *Let $a \le \xi_0 < \xi_1 < \ldots < \xi_q \le b$, be $q+1$ distinct interpolation nodes on $[a, b]$. Then $\pi_qf \in P^q(a, b)$ interpolates $f(x)$ at the nodes ξ_i, if*

$$\pi_qf(\xi_i) = f(\xi_i), \qquad i = 0, 1, \ldots, q \tag{2.2.6}$$

and the Lagrange's formula (2.2.3) *for $\pi_qf(x)$ reads as*

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

Example 14. *For $q = 1$, and considering the whole interval we have only the nodes a and b. Recall that $\lambda_a(x) = \dfrac{x-b}{a-b}$ and $\lambda_b(x) = \dfrac{x-a}{b-a}$, thus as we see in the introduction to this chapter*

$$\pi_1f(x) = f(a)\lambda_a(x) + f(b)\lambda_b(x). \tag{2.2.7}$$

Theorem 3. *We have the following interpolation error estimates*

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

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

$$|f(x) - \pi_qf(x)| \le \frac{1}{(q+1)!}\prod_{i=0}^{q}|x-x_i| \cdot \max_{a\le x\le b}|f^{(q+1)}(x)|,$$

for the Lagrange interpolation error which is of order 1 at each node point $x_0, x_1, \ldots, x_q$.

Proof. The *Taylor part* is well known from elementary calculus. As for the *Lagrange* interpolation error we note that at the node points x_i we have that $f(x_i) - \pi_q f(x_i) = 0$, for $i = 0, 1, \dots, q$. Since $f(x_i) - \pi_q f(x_i)$ has $q+1$ zeros in $[a, b]$, hence there is a function $g(x)$ defined on $[a, b]$ such that

$$f(x) - \pi_q f(x) = (x - x_0)(x - x_1)\dots(x - x_q)g(x). \tag{2.2.8}$$

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

$$\varphi(t) := f(t) - \pi_q f(t) - (t - x_0)(t - x_1)\dots(t - x_q)g(x). \tag{2.2.9}$$

Note that $g(x)$ is independent of t. Further, the function $\varphi(t) = 0$ 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

$$\varphi^{(q+1)}(t) = f^{(q+1)}(t) - 0 - (q+1)!g(x), \tag{2.2.10}$$

where we use the fact that $\deg(\pi_q f(x)) = q$, $(t - x_0)(t - x_1)\dots(t - x_q) = t^{q+1} + \alpha t^q + \dots$, (for some constant α), and $g(x)$ is independent of t. Thus

$$0 = \varphi^{(q+1)}(\xi) = f^{(q+1)}(\xi) - (q+1)!g(x), \tag{2.2.11}$$

and we have

$$g(x) = \frac{f^{(q+1)}(\xi)}{(q+1)!}. \tag{2.2.12}$$

Hence, inserting g from (2.2.12) in (2.2.8), we get the error in Lagrange interpolation as

$$E(x) = f(x) - \pi_q f(x) = \frac{f^{(q+1)}(\xi)}{(q+1)!} \prod_{i=0}^{q} (x - x_i), \tag{2.2.13}$$

and the proof is complete. □

Theorem 4 (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 approximation procedure of solving differential equations we encountered matrices with entries being the integrals of products of basis functions. 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. Below we briefly review some of these numerical integration techniques.

2.3 Numerical integration, Quadrature rule

We want to approximate the integral $I = \int_a^b f(x)dx$ where, on each subinterval, we approximate f using piecewise polynomials of degree d. We denote the approximate value by I_d. To this end we assume, without loss of generality, that $f(x) > 0$ is continuous on the interval $[a, b]$, then the integral $I = \int_a^b f(x)dx$ can be interpreted as the area of the domain under the curve $y = f(x)$; limited by 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.

Simple midpoint rule. Uses midpoint $\frac{a+b}{2}$ of $[a, b]$, i.e. only $f\Big(\frac{a+b}{2}\Big)$. 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

$$I = \int_a^b f(x)dx \approx (b-a)f\Big(\frac{a+b}{2}\Big). \tag{2.3.1}$$

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

$$I \sim I_0 = h \cdot f(a + \frac{h}{2}). \tag{2.3.2}$$

Simple trapezoidal rule. We use 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 points $\Big(a, f(a)\Big)$ and $\Big(b, f(b)\Big)$ and the area under the curve $y = f(x)$ by

$$I = \int_a^b f(x)dx \approx (b-a)\frac{f(a) + f(b)}{2}. \tag{2.3.3}$$

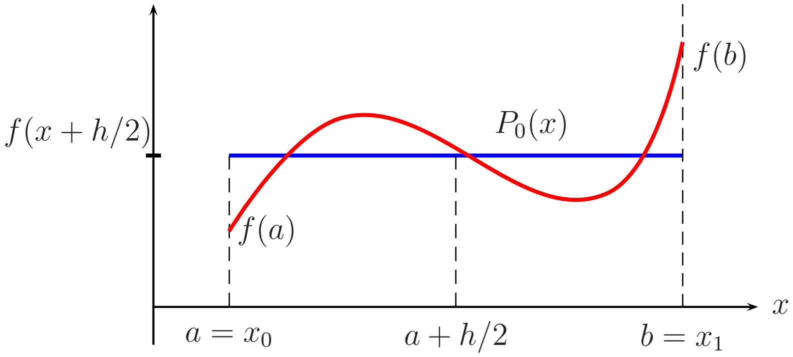


Figure 2.5: Midpoint approximationI_0 of the integral $I = \int_{x_0}^{x_1} f(x)dx$.

This is the area of the trapezoidal between the lines $y = 0$, $x = a$ and $x = b$ and under $P_1(x)$, and therefore is refereed as the *simple trapezoidal rule*. Once again, for the purpose of generalization, we let $x_0 = a$, $x_1 = b$ and assume that the length of interval is h, then (2.3.3) can be written as

$$I \sim I_1 = h \cdot f(a) + \frac{h[f(a+h) - f(a)]}{2} = h\frac{f(a) + f(a+h)}{2}. \qquad (2.3.4)$$

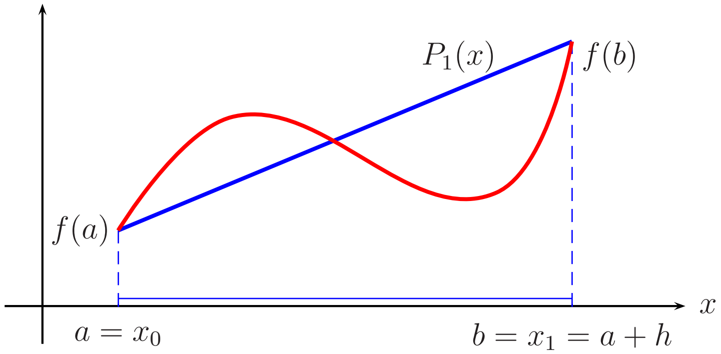


Figure 2.6: Trapezoidal approximation I_1 of the integral $I = \int_{x_0}^{x_1} f(x)dx$.

Simple Simpson's rule. Here we use two endpoints a and b, and the midpoint

$\frac{a+b}{2}$, i.e. $f(a)$, $f(b)$, and $f\Big(\frac{a+b}{2}\Big)$. 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 $P_2(a) = f(a)$, $P_2\Big(\frac{a+b}{2}\Big) = f\Big(\frac{a+b}{2}\Big)$, and $P_2(b) = f(b)$. To determine $P_2(x)$ we may use Lagrange interpolation functions for $q = 2$: let $x_0 = a$, $x_1 = (a+b)/2$ and $x_2 = b$, then

$$P_2(x) = f(x_0)\lambda_0(x) + f(x_1)\lambda_1(x) + f(x_2)\lambda_2(x), \tag{2.3.5}$$

where

$$\begin{cases} \lambda_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, \\ \lambda_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}, \\ \lambda_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}. \end{cases} \tag{2.3.6}$$

Thus

$$I = \int_a^b f(x)dx \approx \int_a^b f(x)P_2(x)\,dx = \sum_{i=0}^{2} f(x_i)\int_a^b \lambda_i(x)\,dx. \tag{2.3.7}$$

Now we can easily compute the integrals

$$\int_a^b \lambda_0(x)\,dx = \int_a^b \lambda_2(x)\,dx = \frac{b-a}{6}, \quad \int_a^b \lambda_1(x)\,dx = \frac{4(b-a)}{6}. \tag{2.3.8}$$

Hence

$$I = \int_a^b f(x)dx \approx \frac{b-a}{6}[f(x_0) + 4f(x_1) + f(x_2)]. \tag{2.3.9}$$

This is the basic idea behind the *simple Simpson's rule*. For the generalization purpose, due to the fact that in this approximation we are using 2 intervals $(a, \frac{a+b}{2})$ and $(\frac{a+b}{2}, b)$, it is more appropriate to assume an interval of length $2h$. Then, (2.3.9) can be written as

$$I = \int_a^b f(x)dx \approx \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)]. \tag{2.3.10}$$

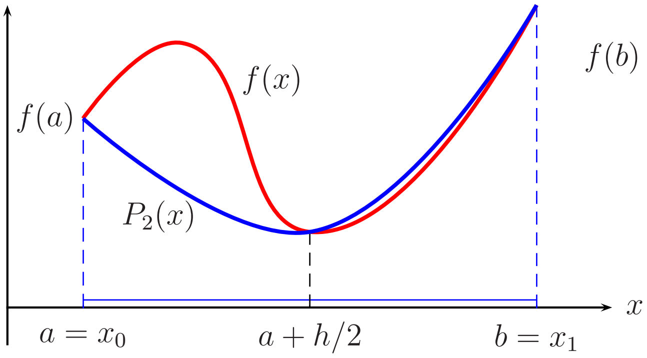


Figure 2.7: Simpson's rule I_2 approximating the integral $I = \int_{x_0}^{x_1} f(x)dx$.

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

General algorithm. To approximate the integral

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

we use the following steps

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

$$a = z_0 < z_1 < z_2 < \ldots < z_{N-1} < z_N = b. \tag{2.3.11}$$

(ii) Write the integral as

$$\int_a^b f(x)dx = \int_{z_0}^{z_1} f(x)\,dx + \ldots + \int_{z_{N-1}}^{z_N} f(x)\,dx = \sum_{k=1}^{N} \int_{z_{k-1}}^{z_k} f(x)\,dx. \tag{2.3.12}$$

(ii) For each subinterval $I_k := [z_{k-1}, z_k]$, $k = 1, 2, \ldots, N$, apply the *very same* integration rule. Then we have the following generalizations.

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

$$h = |I_N| = \frac{b-a}{N}, \qquad \bar{x}_k = \frac{z_{k-1}+z_k}{2}, \; k = 1, 2, \dots, N,$$

then using the simple midpoint rule for the interval $I_k := [z_{k-1}, z_k]$, we have

$$\int_{z_{k-1}}^{z_k} f(x)\, dx \approx \int_{z_{k-1}}^{z_k} f(\bar{x}_k)\, dx = hf(\bar{x}_k). \tag{2.3.13}$$

Summing over $k = 1, 2, \dots, N$, we get

$$\int_a^b f(x)dx \approx \sum_{k=1}^{N} hf(\bar{x}_k) = h[f(\bar{x}_1) + \ldots + f(\bar{x}_N)] := m_N. \tag{2.3.14}$$

(T) *Composite trapezoidal rule*: Simple trapezoidal rule on I_k yields

$$\int_{z_{k-1}}^{z_k} f(x)\, dx \approx \int_{z_{k-1}}^{z_k} f(\bar{x}_k)\, dx = \frac{h}{2}[f(z_{k-1}) + f(z_k)]. \tag{2.3.15}$$

Summing over $k = 1, 2, \dots, N$, we have

$$\begin{aligned} \int_a^b f(x)dx &\approx \sum_{k=1}^{N} \frac{h}{2}[f(z_{k-1}) + f(z_k)] \\ &= \frac{h}{2}[f(z_0) + 2f(z_1) + \ldots + 2f(z_{k-1}) + f(z_k)] := t_N. \end{aligned} \tag{2.3.16}$$

(S) *Composite Simpson's rule*: (Quadratic approximation on each subinterval). Recall that this corresponds to a quadrature rule based on piecewise quadratic interpolation using the endpoints and midpoint of each subinterval. Thus, this time we use the simple Simpson's rule on each subinterval $I_k = [z_{k-1}, z_k]$ using the points z_{k-1}, $\frac{z_{k-1}+z_k}{2}$ and z_k:

$$\int_{z_{k-1}}^{z_k} f(x)\, dx \approx \frac{h}{3}\Big[f(z_{k-1}) + 4f\Big(\frac{z_{k-1}+z_k}{2}\Big) + f(z_k)\Big]. \tag{2.3.17}$$

To proceed we use the following identification on each subinterval I_k, $k = 1, \dots, k$,

$$x_{2k-2} = z_{k-1}, \quad x_{2k-1} = \frac{z_{k-1}+z_k}{2} := \bar{z}_k, \quad x_{2k} = z_k. \tag{2.3.18}$$

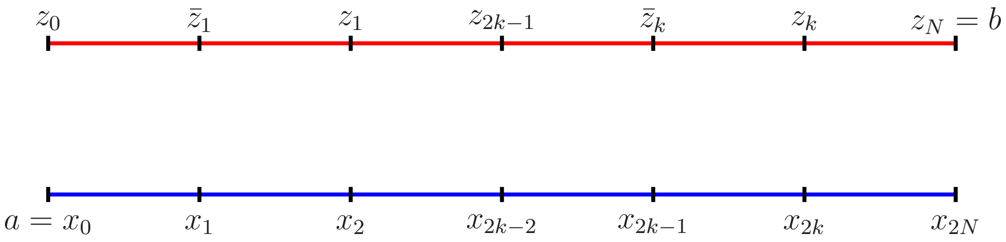


Figure 2.8: Identification of subintervals for composite Simpson's rule

Thus summing (2.3.17) over k and using the above identification, we finally obtain the composite Simpson's rule viz,

$$\begin{aligned}\int_a^b f(x)dx \approx &\sum_{k=1}^{N} \frac{h}{3}\Big[f(z_{k-1}) + 4f\Big(\frac{z_{k-1}+z_k}{2}\Big) + f(z_k)\Big]\\ =&\sum_{k=1}^{N} \frac{h}{3}\Big[f(x_{2k-2}) + 4f(x_{2k-1}) + f(x_{2k})\Big]\\ =&\frac{h}{3}\Big[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4)\\ &+ \ldots + 2f(x_{2N-2}) + 4f(x_{2N-1}) + f(x_{2N})\Big] := S_N.\end{aligned} \tag{2.3.19}$$

Here is the starting procedure where the numbers in the brackets indicate the actual coefficient on each subinterval. For instance the end of interval 1 :$z_1 = x_2$ coincides with the start of interval 2, yielding to the add-up $[1] + [1] = 2$ as the coefficient of $f(x_2)$. A *resonance* which is repeated for each interior node z_k. $k = 1, \ldots, N-1$.

Remark 9. *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:*

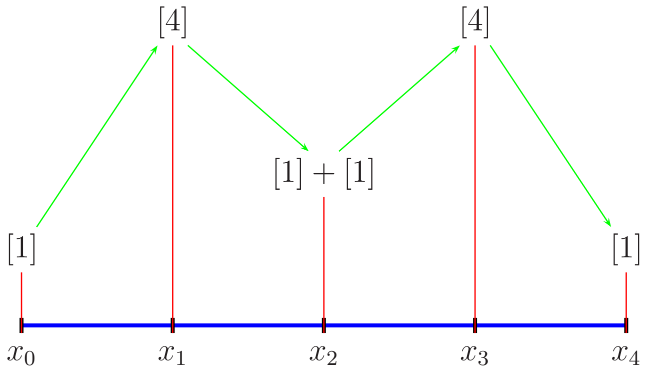


Figure 2.9: Identification of subintervals for composite Simpson's rule

2.3.1 Gauss quadrature rule

This is to choose the points of evolution in an *optimal* manner, not at equally spaced points. We demonstrate this rule through an example viz:

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

$$\int_a^b f(x)dx - \sum_{i=1}^n c_i f(x_i) \text{ for an } \textit{arbitrary} \text{ function } f(x). \tag{2.3.20}$$

Solution. There are $2n$ unknowns in (2.3.20) consisting of n nodes x_i and n coefficients c_i. Therefore we need $2n$ equations. Thus if f were a polynomial, optimal choice of our parameters yields a quadrature rule (2.3.20) which is *exact* for *polynomials* of *degree* $\le 2n - 1$.

Example 15. *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*

$$\int_{-1}^1 f(x)dx \approx c_1 f(x_1) + c_2 f(x_2), \tag{2.3.21}$$

is indeed exact for $f(x)$ replaced by polynomials of degree ≤ 3. Thus we replace f by a polynomial of the form $f(x) = Ax^3 + Bx^2 + Cx + D$ and *require equality in* (2.3.21)*. Thus to determine the coefficients* c_1*,* c_2 *and the nodes* x_1*,* x_2*, in an optimal way, it suffices to use the above approximation as equality when* f *is replaced by the basis functions for polynomials of degree 3:* $1, x, x^2$ *and* x^3*. Consequently we get the equation system*

$$\begin{aligned}
&\int_{-1}^{1} 1dx = c_1 + c_2 \textit{ and we get } [x]_{-1}^{1} = 2 = c_1 + c_2 \\
&\int_{-1}^{1} xdx = c_1 \cdot x_1 + c_2 \cdot x_2 \textit{ and } \Big[\frac{x^2}{2}\Big]_{-1}^{1} = 0 = c_1 \cdot x_1 + c_2 \cdot x_2 \\
&\int_{-1}^{1} x^2dx = c_1 \cdot x_1^2 + c_2 \cdot x_2^2 \textit{ and } \Big[\frac{x^3}{3}\Big]_{-1}^{1} = \frac{2}{3} = c_1 \cdot x_1^2 + c_2 \cdot x_2^2 \\
&\int_{-1}^{1} x^3dx = c_1 \cdot x_1^3 + c_2 \cdot x_2^3 \textit{ and } \Big[\frac{x^4}{4}\Big]_{-1}^{1} = 0 = c_1 \cdot x_1^3 + c_2 \cdot x_2^3,
\end{aligned} \tag{2.3.22}$$

which, although nonlinear, has the solution presented below:

$$\begin{cases} c_1 + c_2 = 2 \\ c_1x_1 + c_2x_2 = 0 \\ c_1x_1^2 + c_2x_2^2 = \frac{2}{3} \\ c_1x_1^3 + c_2x_2^3 = 0 \end{cases} \implies \begin{cases} c_1 = 1 \\ c_2 = 1 \\ x_1 = -\frac{\sqrt{3}}{3} \\ x_2 = \frac{\sqrt{3}}{3} \end{cases} \tag{2.3.23}$$

Thus the approximation

$$\int_{-1}^{1} f(x)dx \approx c_1f(x_1) + c_2f(x_2) = f\Big(-\frac{\sqrt{3}}{3}\Big) + f\Big(\frac{\sqrt{3}}{3}\Big), \tag{2.3.24}$$

is exact for polynomials of degree ≤ 3*.*

Example 16. *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.*

Generalized Gauss quadrature. To generalize Gauss quadrature rule Legendre polynomials are used: 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 $m < n \Longleftrightarrow \int_{-1}^{1} P_n(x)P_m(x)dx = 0$

The *Legendre polynomial* can be obtained through formula:

$$P_k(x) = (-1)^k \frac{d^k}{dx^k}(x^k(1-x)^k), \qquad \text{or} \qquad P_n(x) = \frac{2}{2^n n! dx^n}(x^2-1)^n,$$

Here are a 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* and the *correct choices* as *quadrature* points, i.e. they are giving the points $x_i, 1 \le i \le n$, as the roots of the n-th Legendre polynomial. ($p_0 = 1$ is an exception).

Example 17. *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 5. *Suppose that $x_i, i = 1, 2, \ldots, n$, are roots of n-th Legendre polynomial P_n and that*

$$c_i = \int_{-1}^{1} \prod_{\substack{j=1 \\ j\neq i}}^{n} \Big(\frac{x - x_j}{x_i - x_j}\Big) dx, \ \textit{where} \ \prod_{\substack{j=1 \\ j\neq i}}^{n} \Big(\frac{x - x_j}{x_i - x_j}\Big) \ \textit{is the Lagrange basis.}$$

If $f(x)$ is a polynomial of $\underline{\text{degree} < 2n}$, then $\displaystyle\int_{-1}^{1} f(x)dx \equiv \sum_{i=1}^{n} c_i f(x_i)$.

Proof. Consider a polynomial $R(x)$ of degree $< n$. Rewrite $R(x)$ as $(n-1)$ Lagrange polynomials with nodes at the roots of the n-th Legendre polynomial P_n. This representation of $R(x)$ is exact, since the error is

$$E(x) = \frac{1}{n!}(x-x_1)(x-x_2)\ldots(x-x_n)R^{(n)}(\xi), \quad \text{where } R^{(n)}(\xi) \equiv 0. \quad (2.3.25)$$

Further we have $R(x) = \sum_{i=1}^{n} \prod_{\substack{j=1\\ j\neq i}}^{n} \Big(\frac{x-x_j}{x_i-x_j}\Big) R(x_i)$, so that

$$\begin{aligned}\int_{-1}^{1} R(x)dx &= \int_{-1}^{1} \Big[\sum_{i=1}^{n} \prod_{\substack{j=1\\ j\neq i}}^{n} \Big(\frac{x-x_j}{x_i-x_j}\Big) R(x_i)\Big] dx \\ &= \sum_{i=1} \Big[\int_{-1}^{1} \prod_{\substack{j=1\\ j\neq i}}^{n} \Big(\frac{x-x_j}{x_i-x_j}\Big) dx\Big] R(x_i).\end{aligned} \tag{2.3.26}$$

Moreover

$$\int_{-1}^{1} R(x)dx = \sum_{i=1}^{n} c_i R(x_i) \tag{2.3.27}$$

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

$$P(x) = Q(x) \times P_n(x) + R(x), \qquad \text{deg} Q(x) < n, \quad \text{deg} R(x) < n, \tag{2.3.28}$$

and

$$\int_{-1}^{1} P(x)dx = \int_{-1}^{1} Q(x)P_n(x)dx + \int_{-1}^{1} R(x)dx. \tag{2.3.29}$$

Since $Q(x) \perp P_n(x)$, $\forall Q(x)$ with degree$< n$, thus using (2.3.28) it follows that

$$\int_{-1}^{1} Q(x)P_n(x)dx = 0 \Longrightarrow \int_{-1}^{1} P(x)dx = \int_{-1}^{1} R(x)dx. \tag{2.3.30}$$

Then x_i's are roots of $P_n(x)$, thus $P_n(x_i) = 0$ and we can use (2.3.28) to write

$$P(x_i) = Q(x_i)P_n(x_i) + R(x_i) = R(x_i). \tag{2.3.31}$$

Now using (2.3.27) we obtain that

$$\int_{-1}^{1} P(x)dx = \int_{-1}^{1} R(x)dx = \sum_{i=1}^{n} c_i R(x_i) = \sum_{i=1}^{n} c_i P(x_i). \tag{2.3.32}$$

Summing up:

$$\int_{-1}^{1} P(x)dx = \sum_{i=1}^{n} c_i P(x_i), \tag{2.3.33}$$

and the proof is complete. □

2.4 Exercises

Problem 11. *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 12. *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 13. *Determine the linear interpolation of the function*

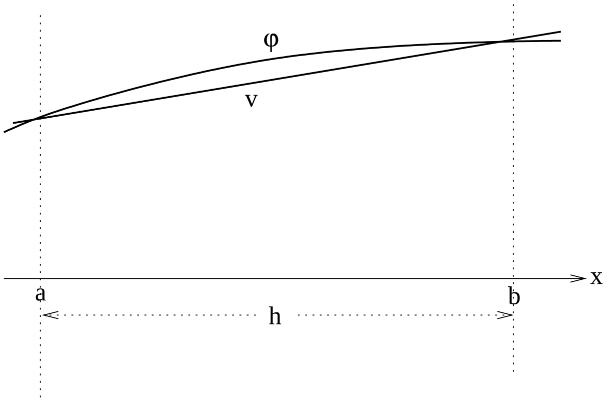
$$f(x) = \frac{1}{\pi^2}(x-\pi)^2 - \cos^2(x - \frac{\pi}{2}), \qquad -\pi \le x \le \pi.$$

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

Problem 14. *Assume that* $w' \in L_1(I)$. *Let* $x, \bar{x} \in I = [a,b]$ *and* $w(\bar{x}) = 0$. *Show that*

$$|w(x)| \le \int_I |w'| dx. \tag{2.4.1}$$

Problem 15. *Assume that* v *interpolates* φ, *at the points* a *and* b.



Show, using (2.4.1) *that*

$$\begin{aligned}
&\text{(i)} && |(\varphi - v)(x)| \le \int_I |(\varphi - v)'|\,dx,\\
&\text{(ii)} && |(\varphi - v)'(x)| \le \int_I |(\varphi - v)''|\,dx = \int_I |\varphi''|\,dx,\\
&\text{(iii)} && \max_I |\varphi - v| \le \max_I |h^2\varphi''|,\\
&\text{(iv)} && \int_I |\varphi - v|\,dx \le \int_I |h^2\varphi''|\,dx,\\
&\text{(v)} && \|\varphi - v\|_I \le \|h^2\varphi''\|_I \quad \textit{and} \quad \|h^{-2}(\varphi - v)\|_I \le \|\varphi''\|_I,\\
&\textit{where} && \|w\|_I = \Big(\int_I w^2\,dx\Big)^{1/2} \quad \textit{is the } L_2(I)\textit{-norm.}
\end{aligned}$$

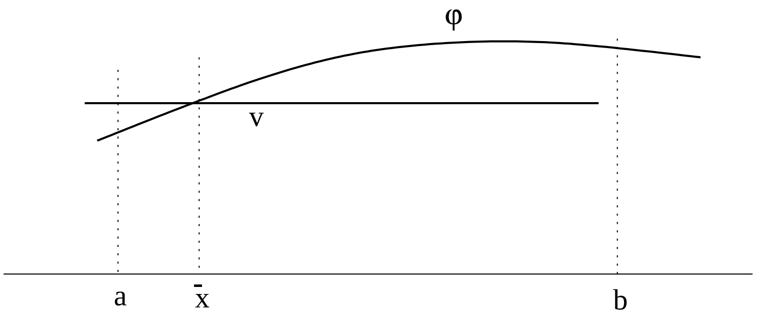
Problem 16. *Use, in the above problem*

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

and show that

$$\begin{aligned}
&\text{(vi)} && |(\varphi - v)(x)| \le 2\int_I |\varphi'|\,dx,\\
&\text{(vii)} && \int_I h^{-1}|\varphi - v|\,dx \le 2\int_I |\varphi'|\,dx \quad \textit{and} \quad \|h^{-1}(\varphi - v)\| \le 2\|\varphi'\|_I.
\end{aligned}$$

Problem 17. *Let now $v(t)$ be the constant interpolant of φ on I.*



Show that

$$\int_I h^{-1}|\varphi - v|\,dx \le \int_I |\varphi'|\,dx. \tag{2.4.2}$$

Problem 18. *Show that*

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

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 19. *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 20. *Prove the following interpolation error estimate:*

$$||\Pi_1 f - f||_{L_\infty(a,b)} \le \frac{1}{8}(b-a)^2||f''||_{L_\infty(a,b)}.$$

Hint: Use Theorem 5.1 from PDE Lecture Notes.

Problem 21. *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 22. *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 23. *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 24. *Prove that*

$$\Big|\int_{x_0}^{x_1} f(x)\,dx - f\Big(\frac{x_1+x_0}{2}\Big)(x-x_0)\Big|$$

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

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

Chapter 3

Linear System of Equations

This chapter is devoted to numerical solution of the linear system of equations of type $Ax = b \Leftrightarrow x = A^{-1}b$. To this approach we shall review the well-known direct method of Gauss elimination and then continue with some more efficient iterative methods. Numerical linear algebra is undoubtedly the most applied tool in the computational aspects of almost all disciplines.

3.1 Direct methods

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

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

We introduce the enlarged $n \times (n+1)$ coefficient matrix $\mathcal{A}$ by

$$\mathcal{A} := \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{pmatrix}, \tag{3.1.1}$$

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

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ \qquad a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \qquad \ldots \quad \ldots \quad \ldots \\ \qquad \ldots \quad \ldots \quad \ldots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ \qquad a_{nn}x_n = b_n, \end{cases} \tag{3.1.2}$$

yields

$$\begin{cases} x_1 = \dfrac{1}{a_{11}}[b_1 - a_{12}x_2 - \ldots - a_{1n}x_n] \\ \qquad \ldots \quad \ldots \quad \ldots \\ \qquad \ldots \quad \ldots \quad \ldots \\ x_{n-1} = \dfrac{1}{a_{n-1,n-1}}[b_{n-1} - a_{n-1,n}x_n] \\ \\ \qquad x_n = \dfrac{b_n}{a_{nn}}. \end{cases} \tag{3.1.3}$$

• **Number of operations.** 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 x_n from (3.1.3) are zero and the number of divisions is one.
• To solve x_{n-1} we need one multiplication and one division.
• To solve x_1 we need $(n-1)$ multiplication and one division.
Thus to solve the linear system of equations given by (3.1.2) 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.** The Gauss elimination method is based on the following obvious facts expressing that: a linear system 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.

Before continuing with the Gauss elimination procedure we recall the simple 3×3 dimensional *uper triangular matrix* U, *lower triangular matrix* L and *diagonal matrix* D.

$$U = \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}, \qquad L = \begin{pmatrix} a & 0 & 0 \\ g & d & 0 \\ h & i & f \end{pmatrix}, \qquad D = \begin{pmatrix} a & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & f \end{pmatrix}.$$

The Gauss *elimination procedure* relay on the elementary row operations and converts the coefficient matrix of the linear equation system to an *upper* triangular matrix. To this end, we start from the first row of the coefficient matrix of the equation system and using elementary row operations eliminate the elements $a_{i1}, i > 1$, under a_{11} (make $a_{i1} = 0$).

The equation system corresponding to this newly obtained matrix $\tilde{A}$ with elements $\tilde{a}_{ij}, \tilde{a}_{i1} = 0, i > 1$, has the same solution as the original one. We repeat the same procedure of the elementary row operations to eliminate the elements $a_{i2}, i > 2$, from the matrix $\tilde{A}$. Continuing in this way, we thus obtain an upper triangular matrix U with corresponding equation system equivalent to the original system (has the same solution). Below we shall illustrate this procedure through an *example*:

Example 18. *Solve the equation system:*

$$\begin{cases} 2x_1 + \ \ x_2 + \ \ x_3 &= 2 \\ 4x_1 - \ \ x_2 + 3x_3 &= 0 \\ 2x_1 + 6x_2 - 2x_3 &= 10. \end{cases} \tag{3.1.4}$$

In the coefficient matrix:

$$\mathcal{A} = \begin{pmatrix} 2 & 1 & 1 & | & 2 \\ 4 & -1 & 3 & | & 0 \\ 2 & 6 & -2 & | & 10 \end{pmatrix}, \tag{3.1.5}$$

we have that $a_{11} = 2,\ a_{21} = 4,$ *and* $a_{31} = 2$. *We introduce the multipliers* $m_{i1}, i > 1$ *by letting*

$$m_{21} = \frac{a_{21}}{a_{11}} = \frac{4}{2} = 2 \qquad m_{31} = \frac{a_{31}}{a_{11}} = \frac{2}{2} = 1. \tag{3.1.6}$$

Now we multiply the first row by m_{21} *and then subtract it from row 2 and replace the result in row 2:*

$$\begin{pmatrix} 2 & 1 & 1 & | & 2 \\ 4 & -1 & 3 & | & 0 \\ 2 & 6 & -2 & | & 10 \end{pmatrix} \begin{matrix} \cdot(-2) \\ \\ \\ \end{matrix} \Longrightarrow \begin{pmatrix} 2 & 1 & 1 & | & 2 \\ 0 & -3 & 1 & | & -4 \\ 2 & 6 & -2 & | & 10 \end{pmatrix} \tag{3.1.7}$$

Similarly we multiply the first row by $m_{31} = 1$ *and subtract it from row 3 to get*

$$\begin{pmatrix} 2 & 1 & 1 & | & 2 \\ 0 & -3 & 1 & | & -4 \\ 0 & 5 & -3 & | & 8 \end{pmatrix}. \tag{3.1.8}$$

In this setting we have $\tilde{a}_{22} = -3$ *and* $\tilde{a}_{32} = 5,$ *and*

$$\tilde{A} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 5 & -3 \end{pmatrix}. \tag{3.1.9}$$

Now let $m_{32} = \tilde{a}_{32}/\tilde{a}_{22} = -5/3$, *then multiplying the second row in* $\tilde{A}$ *by* m_{32} *and subtracting the result from row 3 yields*

$$\begin{pmatrix} 2 & 1 & 1 & | & 2 \\ 0 & -3 & 1 & | & -4 \\ 0 & 0 & -\frac{4}{3} & | & \frac{4}{3} \end{pmatrix}, \tag{3.1.10}$$

where we have obtained the upper triangular matrix

$$U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -\dfrac{4}{3} \end{pmatrix}. \tag{3.1.11}$$

The new equivalent equation system is now

$$\begin{cases} 2x_1 + x_2 + x_3 = & 2 \\ \quad -3x_2 + x_3 = & -4 \\ \qquad -\dfrac{4}{3}x_3 = & \frac{4}{3} \end{cases} \tag{3.1.12}$$

with the obvious 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* (3.1.4)

Definition 14. *We define the lower triangular matrices:*

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{pmatrix}, L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{pmatrix} \quad \text{and } L = \begin{pmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{pmatrix}.$$

The matrices L_1, L_2 *and* L_3 *are* unite *(ones on the diagonal) lower triangular* 3×3*-matrices with the property that*

$$L = (L_2L_1)^{-1} = L_1^{-1}L_2^{-1}, \qquad \text{and} \quad A = LU. \tag{3.1.13}$$

• LU factorization of the matrix A

We generalize the above procedure for the 3×3 linear system of equations to $n \times n$. We have then $A = LU$, where L is a unite lower triangular matrix and U is an upper triangular matrix obtained from A by Gauss elimination. To solve the system $Ax = b$ we let now $y = Ux$, and first solve $Ly = b$ by forward substitution (from the first row to the last) and obtain the vector y, then using y as the known right hand side finally we solve $Ux = y$ by backward substitution (from the last row to the first) and get the solution x.

Example 19. *Following our previous example* $m_{21} = 2, m_{31} = 1$ *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 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.

In general case we have the following result:

Proposition 2. *The* $n \times n$ *unit lower triangular* 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* $(L_2 L_1)^{-1} = L$*, 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.*

Thus $Ax = b \Longleftrightarrow (LU)x = b \Longleftrightarrow L(Ux) = b$. As we outlined above we let $y = Ux$ and solve $Ly = b$ to obtain y. Then with such obtained y we solve x from $Ux = y$. We illustrate this procedure through our *example*:

Example 20. *In our example we have that*

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{5}{3} & 1 \end{pmatrix} \quad \textit{and} \quad b = \begin{pmatrix} 2 \\ 0 \\ 10 \end{pmatrix}.$$

- $Ly = b$ *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*

$$U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -\frac{4}{3} \end{pmatrix} \quad \textit{and} \quad y = \begin{pmatrix} 2 \\ -4 \\ \frac{4}{3} \end{pmatrix},$$

- $Ux = y$ *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 viz $x_1 = 1,\ x_2 = 1,\ x_3 = -1$.

Theorem 6 (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 $k = 1, 2, \dots, n$, where*

$$\Delta_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \dots & \dots & \dots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}.$$

(iv) *$A = LL^T$ where L is lower triangular and has positive diagonal elements. (Cholesky's factorization)*

We do not give a proof of this theorem. The interested reader is referred to literature in algebra and matrix theory.

3.2 Iterative method

Instead of solving $Ax = b$ directly, we consider iterative solution methods based on computing a sequence of approximations $x^{(k)}, k = 1, 2, \dots$ such that

$$\lim_{k\to\infty} x^{(k)} = x \quad \text{or} \quad \lim_{k\to\infty} \|x^{(k)} - x\| = 0, \quad \text{for some norm.}$$

Thus consider the general $n \times n$ linear system of equations $Ax = b$ where both the coefficient matrix A and the vector b have real entries,

$$Ax = b \Longleftrightarrow \begin{cases} a_{11}x_1 + & a_{12}x_2 & \dots & +a_{1n}x_n & = b_1 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n-1,1}x_1 + & \dots & \dots & +a_{n-1,n}x_n & = b_{n-1} \\ a_{n1}x_1 + & \dots & \dots & +a_{nn}x_n & = b_n. \end{cases} \tag{3.2.1}$$

For the system (3.2.1) we shall introduce the two main iterative methods. • **Jacobi iteration:** Assume that $a_{ii} \neq 0$, then

$$\begin{cases} x_1 = -\dfrac{1}{a_{11}}[a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n - b_1] \\ x_{n-1} = -\dfrac{1}{a_{n-1,n-1}}[a_{n-1,1}x_1 + a_{n-1,2}x_2 + \ldots + a_{n-1,n}x_n - b_{n-1}] \\ x_n = -\dfrac{1}{a_{nn}}[a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{n,n}x_n - b_n]. \end{cases}$$

Given an initial approximation for the solution:

$$x^{(0)} = (x_1^{(0)} = c_1,\, x_2^{(0)} = c_2,\, \ldots,\, x_n^{(0)} = c_n),$$

the iteration steps are given by

$$\begin{cases} x_1^{(k+1)} = -\dfrac{1}{a_{11}}[a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} - b_1] \\ \\ x_2^{(k+1)} = -\dfrac{1}{a_{22}}[a_{21}x_1^{(k)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2] \\ \qquad\qquad\qquad\qquad \ldots \\ x_n^{(k+1)} = -\dfrac{1}{a_{nn}}[a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \ldots + a_{n,n-1}x_{n-1}^{(k)} - b_n] \end{cases}$$

Or in compact form in *Jacobi coordinates:*

$$\begin{cases} \sum_{j=1}^n a_{ij}x_j = b_i \Longleftrightarrow a_{ii}x_i = -\sum_{\substack{j=1\\ j\neq i}}^n a_{ij}x_j + b_i, \\ a_{ii}x_i^{(k+1)} = -\sum_{\substack{j=1\\ j\neq i}}^n a_{ij}x_j^{(k)} + b_i. \end{cases} \tag{3.2.2}$$

Convergence criterion

Jacobi gives convergence to the exact solution if A is *diagonally dominant*.

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^n |a_{ij}| \quad i = 1, 2, \ldots, n.$$

Problem 25. *Show that* $A = \begin{pmatrix} 4 & 2 & 1 \\ 1 & 5 & 1 \\ 0 & 1 & 3 \end{pmatrix}$ *is diagonally dominant.*

Note, the Jacobi method needs less operations than Gauss elimination.

Example 21. *Solve* $Ax = b$ *where* $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ *and* $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$*.*
A is diagonally dominant and the matrix equation $Ax = b$ *is equivalent to the linear equation system*

$$\begin{cases} 2x_1 - x_2 = 1 \\ -x_1 + 2x_2 = 1. \end{cases} \tag{3.2.3}$$

We choose zero initial values for x_1 *and* x_2*, i.e.* $x_1^{(0)} = 0$ *and* $x_2^{(0)} = 0$ *and build the Jacobi iteration system*

$$\begin{cases} 2x_1^{(k+1)} = x_2^{(k)} + 1 \\ 2x_2^{(k+1)} = x_1^{(k)} + 1, \end{cases} \tag{3.2.4}$$

where k *is the iteration step. Then we have*

$$\begin{cases} 2x_1^{(1)} = x_2^{(0)} + 1 \\ 2x_2^{(1)} = x_1^{(0)} + 1 \end{cases} \quad \text{with the solution} \quad \begin{cases} x_1^{(1)} = \dfrac{1}{2} \\ x_2^{(1)} = \dfrac{1}{2}. \end{cases} \tag{3.2.5}$$

In the next iteration step:

$$\begin{cases} 2x_1^{(2)} = x_2^{(1)} + 1 \\ 2x_2^{(1)} = x_1^{(1)} + 1 \end{cases} \Rightarrow \begin{cases} 2x_1^{(2)} = \dfrac{1}{2} + 1 \\ 2x_2^{(2)} = \frac{1}{2} \end{cases} \Rightarrow \begin{cases} x_1^{(2)} = \dfrac{3}{4} \\ x_2^{(2)} = \frac{3}{4} \end{cases} \tag{3.2.6}$$

Continuing we have obviously $\lim_{k \to \infty} x_i^{(k)} = x_i$*,* $i = 1, 2$*, where* $x_1 = x_2 = 1$*. Below we have a few first 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: $\|e_k\|_\infty := \max\limits_{i=1,2} |x_i^{(k)} - x_i|$*, then*

$$\|e_0\|_\infty = \max(|x_1^{(0)} - x_1|, |x_2^{(0)} - x_2|) = \max\left(\Big|0-1\Big|, \Big|0-1\Big|\right) = 1$$

$$\|e_1\|_\infty = \max(|x_1^{(1)} - x_1|, |x_2^{(1)} - x_2|) = \max\left(\Big|\frac{1}{2}-1\Big|, \Big|\frac{1}{2}-1\Big|\right) = \frac{1}{2}$$

$$\|e_2\|_\infty = \max(|x_1^{(2)} - x_1|, |x_2^{(2)} - x_2|) = \max\left(\Big|\frac{3}{4}-1\Big|, \Big|\frac{3}{4}-1\Big|\right) = \frac{1}{4}$$

$$\|e_3\|_\infty = \max(|x_1^{(3)} - x_1|, x_2^{(3)} - x_2|) = \max\left(\Big|\frac{7}{8}-1\Big|, \Big|\frac{7}{8}-1\Big|\right) = \frac{1}{8}$$

In this way $\|e_{k+1}\|_\infty = \dfrac{1}{2}\|e_k\|_\infty$*, where* e_k *is the error for step* $k, k \geq 0$*. Iterating we get that for the k-the Jacobi iteration the convergence rate is* $\left(\dfrac{1}{2}\right)^k$*:*

$$\|e_k\|_\infty = \frac{1}{2}\|e_{k-1}\|_\infty = \left(\frac{1}{2}\right)^2 \|e_{k-2}\|_\infty = \ldots = \left(\frac{1}{2}\right)^k \|e_0\|_\infty = \left(\frac{1}{2}\right)^k.$$

• **Gauss-Seidel Method**

Give an initial approximation of the solution

$$x = \Big(x_1^{(0)},\, x_2^{(0)},\, \ldots,\, x_n^{(0)}\Big),$$

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 we have already computed values for $x_1^{(k+1)}, \ldots, x_i^{(k+1)}$ on the right hand sides of the first i rows. The idea with the Gauss-Seidel method is that, in the same iteration step, simultaneously use this 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*.

$$Ax = b \Longleftrightarrow \sum_{j=1}^{n} a_{ij}x_j = b_i \Longleftrightarrow \sum_{j=1}^{i} a_{ij}x_j + \sum_{j=1+1}^{n} a_{ij}x_j = b_i. \tag{3.2.7}$$

Therefore the iterative forms for the Gauss-Seidel is given by

$$\begin{cases} \sum_{j=1}^{i} a_{ij}x_j^{(k+1)} = -\sum_{j=i+1}^{n} x_j^{(k)} + b_i \Longleftrightarrow \\ a_{ii}x_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i. \end{cases} \tag{3.2.8}$$

Example 22. *We consider the same example as above: $Ax = b$ with*

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Recall the Jacobi iteration system

$$\begin{cases} 2x_1^{(k+1)} = x_2^{(k)} + 1 \\ 2x_2^{(k+1)} = x_1^{(k)} + 1. \end{cases} \tag{3.2.9}$$

The corresponding Gauss-seidel *iteration system reads as follows*

$$\begin{cases} 2x_1^{(k+1)} = x_2^{(k)} + 1 \\ 2x_2^{(k+1)} = x_1^{(k+1)} + 1 \end{cases} \tag{3.2.10}$$

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 (3.2.10)*:*

$$2x_1^{(1)} = x_2^{(0)} + 1 \Longrightarrow x_1^{(1)} = \frac{1}{2}.$$

Inserting this value of $x_1^{(1)} = \frac{1}{2}$ in the second equation in (3.2.10) *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 a few first 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 $\lim_{k\to\infty} x_1^{(k)} = \lim_{k\to\infty} x_2^{(k)} = 1$. Now with $\|e_k\|_\infty = \max_{i=1,2} |x_i^{(k)} - x_i|$, we get the successive iteration errors:

$$\|e_1\|_\infty = \max(|x_1^{(1)} - x_1|, |x_2^{(1)} - x_2|) = \max\left(\left|\frac{1}{2} - 1\right|, \left|\frac{3}{4} - 1\right|\right) = \frac{1}{2}$$

$$\|e_2\|_\infty = \max\left(\left|\frac{7}{8} - 1\right|, \left|\frac{15}{16} - 1\right|\right) == \frac{1}{8}, \quad \|e_3\|_\infty = \max\left(\frac{1}{32}, \frac{1}{64}\right) = \frac{1}{32}.$$

Thus for the Gauss-Seidel iteration $\|e_{k+1}\|_\infty = \frac{1}{4}\|e_k\|_\infty$, *where* e_k *is the error for step* k, *and hence we can conclude that the Gauss-Seidel method converges faster than the Jacobi method:*

$$\|e_k\|_\infty = \frac{1}{4}\|e_{k-1}\|_\infty = \left(\frac{1}{4}\right)^2\|e_{k-2}\|_\infty = \cdots = \left(\frac{1}{4}\right)^k\|e_0\|_\infty = \left(\frac{1}{4}\right)^k.$$

• **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

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\Big[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big] \qquad (3.2.11)$$

For $\omega > 1$ the method is called an *over-relaxation method* and if $0 < \omega < 1$, it is referred as an *under-relaxation method*.
S.O.R. coordinates:

$$a_{ii}x_i^{(k+1)} = a_{ii}x_i^{(k)} - \omega\Big(\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} - b_i\Big) \qquad (3.2.12)$$

• **Abstraction of iterative methods**
In our procedures we have considered $Ax = b$ and $x = Bx + C$ as equivalent linear equation systems, where B is the iteration matrix and $x_{k+1} = Bx_k + C$.

Potential advantages of iteration methods over direct methods: they are
(i) Faster (depends on B, accuracy is required)
(ii) Less memory is required (Sparsity of A can be preserved.)

Questions:
(Q1) For a given A, what is a good choice for B?
(Q2) When does $x_k \to x$?
(Q3) What is the rate of convergence?
The error at step k is $e_k = x_k - x$ and that of step $(k+1)$ is $e_{k+1} = x_{k+1} - x$. Then we have $e_{k+1} = x_{k+1} - x = (Bx_k + C) - (Bx - C) = B \cdot \underbrace{(x_k - x)}_{e_k} = Be_k$.

Iterating, we have

$$e_k = Be_{k-1} = B \cdot Be_{k-2} = B \cdot B \cdot Be_{k-3} = B^4 e_{k-4} = \ldots = B^k e_{k-k} = B^k\ e_0.$$

Thus we have shown that $e_k = B^k e_0$. 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 $A = L + D + U$, which is a *splitting* of A. Now we can rewrite $Ax = b$ as $(D + D + U)x = b$ then $Dx = -(L + U)x + b$, and we may reformulate the iterative methods as follows:

Jacobi's method

$$Dx_{k+1} = -(L + U)x_k + b \Rightarrow B_J = -D^{-1}(L + U),$$

where B_J is the *Jacobi's iteration matrix.*

Example 23. *Write the linear system in the matrix form $x = B_J x + C$*

$$\begin{cases} 2x_1 - x_2 = 1 \\ -x_1 + 2x_2 = 1 \end{cases} \Rightarrow \begin{cases} x_1 = \dfrac{1}{2}x_2 + \dfrac{1}{2} \\ \\ x_2 = \dfrac{1}{2}x_1 + \dfrac{1}{2} \end{cases} \quad \textit{which in the matrix form is}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \ \textit{where}$$

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, B_J = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} \ \textit{and } C = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Example 24. *Determine the same 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

$$D^{-1} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix},$$

and thus

$$B_J = -D^{-1}(L+U) = \begin{pmatrix} -\dfrac{1}{2} & 0 \\ 0 & -\dfrac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \dfrac{1}{2} \\ \dfrac{1}{2} & 0 \end{pmatrix}.$$

Gauss-Seidel's method

As above we write $Ax = b$ as $(L+D+U)x = b$ but now we choose $(L+D)x = -Ux+b$. Similar to the previous procedure we have $(L+D)x_{k+1} = -Ux_k+b$, and then $B_{GS} = -(L+D)^{-1}U$, where B_{GS} is *Gauss-Seidel's iteration matrix.*

Relaxation

Gauss-Seidel gives that $(L+D)x = -Ux+b$, thus the iteration procedure is:

$$Dx_{k+1} = Dx_k - [Lx_{k+1} + (D+U)x_k - b].$$

where ω is the Relaxation parameter, $\omega = 1$ gives the Gauss-seidel iteration. Now we have that

$$(\omega L + D)x_{k+1} = [(1-\omega)D - \omega U]x_k + \omega b,$$

thus the *Relaxation iteration matrix* is:

$$B_\omega = (\omega L + D)^{-1}[(1-\omega)D - \omega U].$$

3.3 Exercises

Problem 26. *Illustrate the LU factorization for the matrix*

$$A = \begin{bmatrix} 1 & 3 & 2 \\ -2 & -6 & 1 \\ 2 & 5 & 7 \end{bmatrix}.$$

Problem 27. *Solve* $A^4 x = b$ *for*

$$A = \begin{bmatrix} -1 & 2 \\ 2 & -3 \end{bmatrix} \qquad b = \begin{bmatrix} 144 \\ -233 \end{bmatrix}$$

Problem 28. *Find the unique the LDU factorization for the matrix*

$$A = \begin{bmatrix} 1 & 1 & -3 \\ 0 & 1 & 1 \\ 3 & -1 & 1 \end{bmatrix}.$$

Problem 29. *Show that every orthogonal* 2×2 *matrix is of the form*

$$A_1 = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \qquad \text{or} \qquad A_2 = \begin{bmatrix} c & s \\ s & -c \end{bmatrix},$$

where $c^2 + s^2 = 1$

Problem 30. *Solve the following system*

$$\begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

using 3 iterations of the following methods using 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).*

Chapter 4

Two-points BVPs

In this chapter we focus on finite element approximation procedure for the two point *boundary value problems* (BVPs) of Dirichlet, Neumann and mixed type. For each PDE we formulate a corresponding *variational formulation*(VF) and a *minimization problem* (MP) and prove that to solve the boundary value problem is equivalent to the (VF) which in turn is equivalent to solve a minimization problem (MP), i.e,

$$(BVP) \Longleftrightarrow (VF) \Longleftrightarrow (MP).$$

We also prove *a priori* and *a posteriori* error estimates for BVPs.

4.1 A Dirichlet problem

Assume a horizontal elastic bar that occupies the interval $I := [0, 1]$, is fixed at the end-points. Let $u(x)$ denote the displacement at a point $x \in I$, and $a(x)$ be the *modulus of elasticity* and $f(x)$ a *load function*, then one can show that u satisfies the following boundary value problem for the Poisson's equation

$$(BVP)_1 \qquad \begin{cases} -\Big(a(x)u'(x)\Big)' = f(x), \quad 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{4.1.1}$$

We assume that $a(x)$ is piecewise continuous in $(0, 1)$, bounded for $0 \le x \le 1$ and $a(x) > 0$ for $0 \le x \le 1$.

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 set

$$H_0^1 = \Big\{ v(x) : \int_0^1 [v(x)^2 + v'(x)^2]dx < \infty, \quad v(0) = v(1) = 0 \Big\}. \tag{4.1.2}$$

The *variational formulation* (VF) for $(\mathrm{BVP})_1$ is obtained by multiplying the equation by a function $v(x) \in H_0^1(0,1)$ and integrating over $(0,1)$:

$$-\int_0^1 [a(x)u'(x)]'v(x)dx = \int_0^1 f(x)v(x)dx. \tag{4.1.3}$$

By partial integration we get

$$-\Big[a(x)u'(x)v(x)\Big]_0^1 + \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx. \tag{4.1.4}$$

Now since $v(0) = v(1) = 0$ we have

$$\int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx. \tag{4.1.5}$$

Thus the *variational formulation* for the problem (4.1.1) is as follows:

Find $u(x) \in H_0^1$ such that

$$(\mathrm{VF})_1 \qquad \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \quad \forall v(x) \in H_0^1. \tag{4.1.6}$$

In other words we have that if u satisfies $(\mathrm{BVP})_1$ (4.1.1), then u also satisfies the $(\mathrm{VF})_1$, (4.1.5) above. We write this as $(\mathrm{BVP})_1 \Longrightarrow (\mathrm{VF})_1$. Now the question is whether the reverse implication is through, i.e. if which conditions can we deduce the reverse implication $(\mathrm{VF})_1 \Longrightarrow (\mathrm{BVP})_1$? It appears that this question has an affirmative answer and the two problems are indeed equivalent. We prove this observation in the following theorem.

Theorem 7. *u satisfies* $(\mathrm{BVP})_1 \Longleftrightarrow u$ *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 (4.1.5) and using $v(0) = v(1) = 0$ we come back to

$$-\int_0^1 [a(x)u'(x)]'v(x)dx = \int_0^1 f(x)v(x)dx, \qquad \forall v(x) \in H_0^1 \tag{4.1.7}$$

which can also be written as

$$\int_0^1 \Big[- \Big(a(x)u'(x)\Big)' - f(x)\Big] v(x)dx = 0, \qquad \forall v(x) \in H_0^1. \tag{4.1.8}$$

We *claim* that (4.1.8) implies

$$-\Big(a(x)u'(x)\Big)' - f(x) \equiv 0, \qquad \forall x \in (0,1). \tag{4.1.9}$$

If our claim is *not true!*, then there exists at least one $\xi \in (0,1)$, such that

$$-\Big(a(\xi)u'(\xi)\Big)' - f(\xi) \neq 0, \tag{4.1.10}$$

where we may assume, without loss of generality, that

$$-\Big(a(\xi)u'(\xi)\Big)' - f(\xi) > 0 \quad (\text{or } < 0). \tag{4.1.11}$$

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,

$$g(x) := -\Big(a(x)u'(x)\Big)' - f(x) > 0, \qquad \text{for} \quad x \in (\xi - \delta, \xi + \delta). \tag{4.1.12}$$

Now we take $v(x)$ in (4.1.8) as a hat function, $v^*(x) > 0$ on $(\xi-\delta, \xi+\delta)$ and

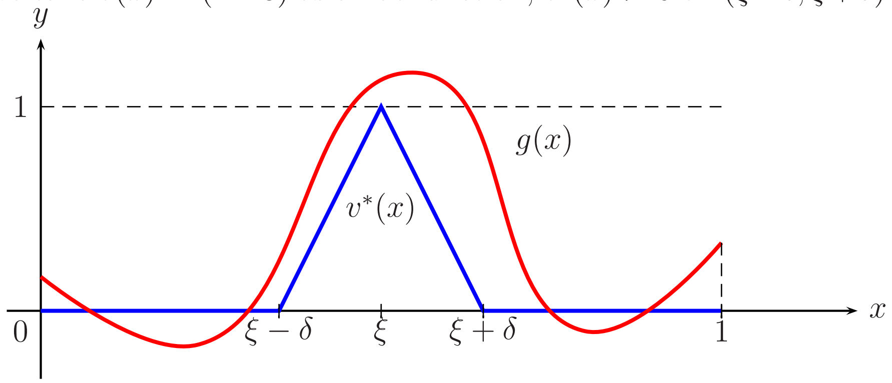


Figure 4.1: The hat function $v^*(x)$ over the interval $(\xi - \delta, \xi + \delta)$.

we have $v^*(x) \in H_0^1$ and $\displaystyle\int_0^1 \underbrace{\Big[- \Big(a(x)u'(x)\Big)' - f(x)\Big]}_{>0} \underbrace{v^*(x)}_{>0}\, dx > 0$, which contradicts (4.1.8), thus our claim is true and the proof is complete. □

Corollary 1. *(i) If both $f(x)$ and $a(x)$ are continuous and $a(x)$ is 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 meaning. Therefore (VF) covers a larger set of data than (BVP).

4.2 Minimization problem

For the problem (4.1.1), i.e. our $(\mathrm{BVP})_1$

$$\begin{cases} -\Big(a(x)u'(x)\Big)' = f(x), \quad 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{4.2.1}$$

we formulate a *minimization problem* (MP) as follows:

Problem 31. *Find $u \in H_0^1$ such that $F(u) \le F(w)$, $\forall w \in H_0^1$, where $F(w)$ is the total energy of $w(x)$ given by*

$$F(w) = \underset{\textit{Internal energy}}{\frac{1}{2}\int_0^1 a(w')^2 dx} \quad - \underset{\textit{Load potential}}{\int_0^1 fw dx} \tag{4.2.2}$$

Theorem 8. *The minimization problem above is equivalent to variational formulation $(VF)_1$,*

$$(MP) \Longleftrightarrow (VF) \qquad i.e.$$

$$F(u) \le F(w), \forall w \in H_0^1 \Longleftrightarrow \int_0^1 au'v'dx = \int_0^1 fvdx, \quad \forall v \in H_0^1. \tag{4.2.3}$$

Proof. ($\Longleftarrow$) For $w \in H_0^1$ we let $v = w - u$, then $v \in H_0^1$ and

$$\begin{aligned} F(w) = F(u+v) &= \frac{1}{2}\int_0^1 a\Big((u+v)'\Big)^2 dx - \int_0^1 f(u+v)dx = \\ &= \underbrace{\frac{1}{2}\int_0^1 2au'v'dx}_{(i)} + \underbrace{\frac{1}{2}\int_0^1 a(u')^2 dx}_{(ii)} + \frac{1}{2}\int_0^1 a(v')^2 dx \\ &\quad - \underbrace{\int_0^1 fudx}_{(iii)} - \underbrace{\int_0^1 fvdx.}_{(iv)} \end{aligned}$$

but $(i)+(iv)=0$, since by $(\mathrm{VF})_1$ $\displaystyle\int_0^1 au'v'dx = \int_0^1 fvdx$. Further by definition of the functional F we have $(ii)+(iii)=F(u)$. Thus

$$F(w)=F(u)+\frac{1}{2}\int_0^1 a(x)(v'(x))^2dx, \tag{4.2.4}$$

and since $a(x)>0$ we have $F(w)>F(u)$.
($\Longrightarrow$) Let now $F(u)\le F(w)$ and set $g(\varepsilon,w)=F(u+\varepsilon v)$, then g has a *minimum* at $\varepsilon=0$. In other words $g'(\varepsilon,w)\Big|_{\varepsilon=0}=0$. We have

$$\begin{aligned} g(\varepsilon,w)=F(u+\varepsilon v)&=\frac{1}{2}\int_0^1 a\Big((u+\varepsilon v)'\Big)^2dx-\int_0^1 f(u+\varepsilon v)dx=\\ &=\frac{1}{2}\int_0^1\{a(u')^2+a\varepsilon^2(v')^2+2a\varepsilon u'v'\}dx-\int_0^1 fudx-\varepsilon\int_0^1 fvdx.\end{aligned}$$

Now we compute the derivative $g'_\varepsilon(\varepsilon,w)$.

$$g'_\varepsilon(\varepsilon,w)=\frac{1}{2}\{2a\varepsilon(v')^2+2au'v'\}dx-\int_0^1 fvdx \tag{4.2.5}$$

and $g'_\varepsilon\Big|_{(\varepsilon=0)}=0$, yields

$$\int_0^1 au'v'dx-\int_0^1 fvdx=0, \tag{4.2.6}$$

which is the variational formulation. Thus we conclude that $F(u)\le F(w)\Longrightarrow$ $(\mathrm{VF})_1$ and the proof is complete. □

We summarize the two theorems in short as

Corollary 2.

$$(BVP)_1 \Longleftrightarrow (VF)_1 \Longleftrightarrow (MP).$$

4.3 A mixed Boundary Value Problem

Obviously changing the boundary conditions would require changes in the variational formulation. This can be, e.g. seen in formulating the (VF) corresponding to the following mixed boundary value problem

$$(\mathrm{BVP})_2 \qquad \begin{cases} -\Big(a(x)u'(x)\Big)' = f(x), \quad 0 < x < 1 \\ u(0) = 0, \qquad a(1)u'(1) = g_1 \neq 0. \end{cases} \tag{4.3.1}$$

We multiply the equation by a suitable function $v(x)$ with $v(0) = 0$ and integrate over the interval $(0, 1)$ to obtain

$$-\int_0^1 [a(x)u'(x)]'v(x)dx = \int_0^1 f(x)v(x)dx. \tag{4.3.2}$$

By partial integration we get, as before, that

$$-[a(x)u'(x)v(x)]_0^1 + \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx \tag{4.3.3}$$

Using the boundary data $v(0) = 0$ and $a(1)u'(1)v(1) = g_1v(1)$ we get

$$\int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx + g_1v(1), \quad \forall v \in \tilde{H}_0^1, \tag{4.3.4}$$

where

$$\tilde{H}_0^1 = \{v(x) : \int_0^1 [v(x)^2 + v'(x)^2]dx < \infty, \text{ such that } v(0) = 0\}. \tag{4.3.5}$$

Then (4.3.4) 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 we want to show that

Theorem 9. $(\mathrm{BVP})_2 \Longleftrightarrow (\mathrm{VF})_2$

Proof. ($\Longrightarrow$) This part is trivial and already proved along the above lines.

($\Longleftarrow$) To prove 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 show

(i) the solution satisfies the differential equation

(ii) the solution satisfies the boundary conditions

We start with $(VF)_2$ and perform a reversed order partial integration to get

$$\int_0^1 a(x)u'(x)v'(x)dx = [a(x)u'(x)v(x)]_0^1 - \int_0^1 [a(x)u'(x)]'v(x)\ dx. \tag{4.3.6}$$

Since $v(0) = 0$, we get

$$\int_0^1 a(x)u'(x)v'(x)dx = a(1)u'(1)v(1) - \int_0^1 [a(x)u'(x)]'vdx \tag{4.3.7}$$

Thus the variational formulation $(\mathrm{VF})_2$ can be written as

$$-\int_0^1 [a(x)u'(x)]'vdx + a(1)u'(1)v(1) = \int_0^1 f(x)v(x)dx + g_1v(1). \tag{4.3.8}$$

The equation (4.3.8) is valid for every $v(x) \in \tilde{H}_0^1(0,1)$, including a test function $v(x)$ 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 $v(1) = 0$ (4.3.8) is reduced to

$$-\int_0^1 [a(x)u'(x)]'vdx = \int_0^1 f(x)v(x)dx, \qquad \forall v(x) \in H_0^1 \tag{4.3.9}$$

Now as in the case of the Dirichlet problem (4.3.9) gives the differential equation in (4.3.1) and hence claim (i) is through.

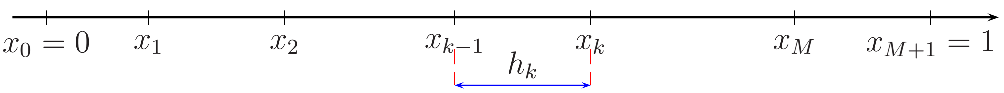
On the other hand (4.3.9) is just the equation in (4.3.1) multiplied by a test function v and integrated over $(0,1)$, so (4.3.9) is equally valid for $v \in \tilde{H}_0^1(0,1)$. Now inserting (4.3.9) in (4.3.8) we also get $g_1v(1) = a(1)u'(1)v(1)$, which choosing $v(1) \neq 0$, e.g. $v(1) = 1$, gives that $g_1 = a(1)u'(1)$ and the proof is complete. □

Remark 10. *i) The Dirichlet boundary conditions is called the essential boundary condition and is strongly imposed in the test function space: Enforced explicitly to the trial and test functions in (VF).*

ii) The Neumann and Robin Boundary conditions are called the natural boundary conditions and are automatically satisfied in (VF), therefore are weakly imposed.

4.4 The finite element method. (FEM)

Let $\mathcal{T}_h = \{0 = x_0 < x_1 < \ldots < x_M < x_{M+1} = 1\}$ be a partition of $0 \le x \le 1$ into subintervals $I_k = [x_{k-1}, x_k]$ and $h_k = x_k - x_{k-1}$ Define the piecewise



constant function $h(x) = x_k - x_{k-1} = h_k$ for $x \in I_k$. Let now $\mathcal{C}\Big(I, P_1(I_k)\Big)$ denote the set of all continuous piecewise linear functions on $\mathcal{T}_h$ (continuous in whole I, linear on each subinterval I_k), and define

$$V_h^{(0)} = \{v : v \in \mathcal{C}\Big(I, P_1(I_k)\Big), \quad v(0) = v(1) = 0\} \tag{4.4.1}$$

Note that $V_h^{(0)}$ is a subspace of

$$H_0^1 = \{v(x) : \int_0^1 [v(x)^2 + v'(x)^2]dx < \infty, \quad \text{and} \quad v(0) = v(1) = 0\}. \tag{4.4.2}$$

A *cG(1) (continuous Galerkin of degree 1) finite element formulation* for our Dirichlet boundary value problem (BVP) is given by: Find $u_h \in V_h^{(0)}$ such that

$$\text{(FEM)} \qquad \int_0^1 a(x)u_h'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \quad \forall v \in V_h^{(0)}. \tag{4.4.3}$$

Now the purpose is to make *estimate of error* arising in approximating the solution for (BVP) by the functions in $V_h^{(0)}$. To this end we need to introduce some measuring environment for the error. 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 *elastic energy* for an elastic string modeled for our Dirichlet boundary value problem (BVP).

4.5 Error estimates in the energy norm

There are two types of error estimates: An *a priori error estimate* depends on the *exact solution* $u(x)$ and not on the approximate solution $u_h(x)$. In such estimates the error analysis are performed theoretically and before computations. An *a posteriori error estimate* where the error depends on the *residual*,i.e, the difference between the left and right hand side in the equation when the exact solution $u(x)$ is replaced by the approximate solution $u_h(x)$. A posteriori error estimates can be derived after that the approximate solution is computed.

Below first we shall prove a general theorem which shows that the finite element solution is the best approximate solution for either of our Dirichlet problem in the energy norm.

Theorem 10. *Let $u(x)$ be a solution of the Dirichlet boundary value problem*

$$BVP \qquad \begin{cases} -\Big(a(x)u'(x)\Big)' = f(x), & 0 < x < 1 \\ u(0) = 0 \quad u(1) = 0. \end{cases} \tag{4.5.1}$$

and $u_h(x)$ its finite element element approximation given by (4.4.3). *Then we have*

$$\|u - u_h\|_E \leq \|u - v\|_E, \forall v(x) \in V_h^{(0)}. \tag{4.5.2}$$

This means that the finite element solution $u_h \in V_h^{(0)}$ is the best approximation of the solution u by functions in $V_h^{(0)}$.

Proof. Recall the variational formulation associated to the problem (4.4.1):

$$\text{(VF)} \quad \int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \qquad \forall v \in H_0^1. \tag{4.5.3}$$

We take an arbitrary $v \in V_h^{(0)}$, then by the definition of the energy norm

$$\begin{aligned}
\|u-u_h\|_E^2 &= \int_0^1 a(x)(u'-u_h')^2(x)dx \\
&= \int_0^1 a(x)\Big(u'(x)-u_h'(x)\Big)\Big(u'(x)\underbrace{-v'(x)+v'(x)}_{=0}-u_h'(x)\Big)dx \\
&= \int_0^1 a(x)\Big(u'(x)-u_h'(x)\Big)\Big(u'(x)-v'(x)\Big)dx \\
&\quad+\int_0^1 a(x)\Big(u'(x)-u_h'(x)\Big)\Big(v'(x)-u_h'(x)\Big)\,dx
\end{aligned} \tag{4.5.4}$$

Now since $v-u_h \in V_n^{(0)} \subset H_0^1$, we have by the variational formulation (4.5.3)

$$\int_0^1 a(x)u'(x)\Big(v'(x)-u_h'(x)\Big)dx = \int_0^1 f\Big(v(x)-u_h(x)\Big), \tag{4.5.5}$$

with its finite element counterpart, see (4.4.3),

$$\int_0^1 a(x)u_h'(x)\Big(v'(x)-u_h'(x)\Big)dx = \int_0^1 f\Big(v(x)-u_h(x)\Big). \tag{4.5.6}$$

Subtracting these two relations the last line of the estimate (4.5.4) above vanishes, so we end up with

$$\begin{aligned}
\|u-u_h\|_E^2 &= \int_0^1 a(x)[u'(x)-u_h'(x)][u'(x)-v'(x)]dx \\
&= \int_0^1 a(x)^{\frac12}[u'(x)-u_h'(x)]a(x)^{\frac12}[u'(x)-v'(x)]dx \\
&\le \Big(\int_0^1 a(x)[u'(x)-u_h'(x)]^2dx\Big)^{\frac12}\Big(\int_0^1 a(x)[u'(x)-v'(x)]^2dx\Big)^{\frac12} \\
&= \|u-u_h\|_E\cdot\|u-v\|_E,
\end{aligned} \tag{4.5.7}$$

where, in the last estimate, we used Cauchy-Schwartz inequality. Thus

$$\|u-u_h\|_E \le \|u-v\|_E, \tag{4.5.8}$$

and the proof is complete. □

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 shall study is $v(x) = \pi_h u(x)$: the *piecewise linear interpolant of* $u(x)$, introduced in chapter 2. Recall the interpolation error estimate in L_p-norms:

Theorem 11. *(i) Let* $0 = x_0 < x_1 < x_2 < \dots < x_M < x_{M+1} = 1$ *be a partition of* $[0, 1]$ *and* $h = (x_{j+1} - x_j), j = 0, 1, \dots, M$

(ii) Let $\pi_h v(x)$ *be the piecewise linear interpolant of* $v(x)$. *Then there is an interpolation constant* c_i *such that*

$$\|\pi_h v - v\|_{L_p} \le c_i \|h^2 v''\|_{L_p} \quad 1 \le p \le \infty \tag{4.5.9}$$

$$\|(\pi_h v)' - v'\|_{L_p} \le c_i \|h v''\|_{L_p} \tag{4.5.10}$$

$$\|\pi_h v - v\|_{L_p} \le c_i \|h v'\|_{L_p}. \tag{4.5.11}$$

Theorem 12 (An apriori 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*

$$\|u - u_h\|_E \le C_i \|h u''\|_a. \tag{4.5.12}$$

Proof. According to our general above we have

$$\|u - u_h\|_E \le \|u - v\|_E, \qquad \forall v \in V_h^{(0)}. \tag{4.5.13}$$

Now since $\pi_h u(x) \in V_h^{(0)}$, we may take $v = \pi_h u(x)$ in (4.5.13) and use, e.g. the second estimate in the interpolation theorem to get

$$\begin{aligned} \|u - u_h\|_E &\le \|u - \pi_h u\|_E = \|u' - (\pi_h u)'\|_a \\ &\le C_i \|h u''\|_a = C_i \Big(\int_0^1 a(x) h^2(x) u''(x)^2 \, dx \Big)^{1/2}, \end{aligned} \tag{4.5.14}$$

which is the desired result. □

Remark 11. *Now if the objective is to divide (0,1) into a fixed, finite, number of subintervals, then one can use the proof of theorem 8.3: to obtain an optimal (a best possible) partition of (0,1); in the sense that: whenever* $a(x)u''(x)$ *gets large we compensate by making* $h(x)$ *smaller. This, however, requires that the exact solution* $u(x)$ *is known. Now we want to study a posteriori error analysis, where instead of the unknown value of* $u(x)$*, we use the known computed values of the approximate solution.*

Theorem 13 (a posteriori error estimate)**.** *There is an interpolation constant c_i depending only on $a(x)$ such that the error in finite element approximation of the Dirichlet boundary value problem (BVP)* (4.5.1) *satisfies*

$$\|e(x)\|_E \le c_i\Big(\int_0^1 \frac{1}{a(x)} h^2(x) R^2[u_h(x)]dx\Big)^{\frac{1}{2}}, \tag{4.5.15}$$

where $e(x) = u(x) - u_h(x)$, note that $e \in H_0^1$.

Proof. By the definition of the *energy norm* we have

$$\begin{aligned}\|e(x)\|_E^2 &= \int_0^1 a(x)[e'(x)]^2 dx = \int_0^1 a(x)[u'(x) - u_h'(x)]e'(x)dx \\ &= \int_0^1 a(x)u'(x)e'(x)dx - \int_0^1 a(x)u_h'(x)e'(x)dx\end{aligned} \tag{4.5.16}$$

Since $e \in H_0^1$ the variational formulation (VF) gives that

$$\int_0^1 a(x)u'(x)e'(x)dx = \int_0^1 f(x)e(x)dx. \tag{4.5.17}$$

Thus we have

$$\|e(x)\|_E^2 = \int_0^1 f(x)e(x)dx - \int_0^1 a(x)u_h'(x)e'(x)dx. \tag{4.5.18}$$

Adding and subtracting the interpolant $\pi_h e(x)$ and $\pi_h e'(x)$ 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 a solution of the (FEM) (4.4.3) and $\pi_h e(x) \in V_h^{(0)}$ we have $-(ii) + (i) = 0$. Hence

$$\begin{aligned}\|e(x)\|_E^2 &= \int_0^1 f(x)[e(x) - \pi_h e(x)]dx - \int_0^1 a(x)u_h'(x)[e'(x) - \pi_h e'(x)]dx \\ &= \int_0^1 f(x)[e(x) - \pi_h e(x)]dx - \sum_{k=1}^{M} \int_{x_{k-1}}^{x_k} a(x)u_h'(x)[e'(x) - (\pi_h e'(x)]dx.\end{aligned}$$

Now, for the integrals in the sum above, we integrate by parts over each subinterval (x_{k-1}, x_k):

$$-\int_{x_{k-1}}^{x_k} \underbrace{a(x)u_h'(x)}_{g(x)} \underbrace{(e'(x) - \pi_h e'(x))}_{F'(x)} dx = [\text{P.I.}] =$$

$$= -\Big[\underbrace{a(x)u_h'(x)}_{g(x)} \underbrace{(e(x) - \pi_h e(x))}_{F(x)} \Big]_{x_{k-1}}^{x_k} + \int_{x_{k-1}}^{x_k} \underbrace{(a(x)u_h'(x))'}_{g'(x)} \underbrace{(e(x) - \pi_h e(x))}_{F(x)} dx$$

Since $e(x_k) = \pi_h e(x_k)$, $k = 0, 1 \ldots, M$, where x_k:s are the interpolation nodes we have $F(x_k) = F(x_{k-1}) = 0$, and thus

$$-\int_{x_{k-1}}^{x_k} a(x)u_h'(x)(e'(x) - \pi_h e'(x))dx = \int_{x_{k-1}}^{x_k} \Big(a(x)u_h'(x)\Big)'(e(x) - \pi_h e(x))dx.$$

Hence summing over k, we get

$$-\int_0^1 a(x)u_h'(x)[e'(x) - \pi_h e'(x)]dx = \int_0^1 [a(x)u_h'(x)]'(e(x) - \pi_h e(x))dx,$$

and therefore

$$\begin{aligned} \|e(x)\|_E^2 &= \int_0^1 f(x)[e(x) - \pi_h e(x)]dx + \int_0^1 [a(x)u_h'(x)]'(e(x) - \pi_h e(x))dx \\ &= \int_0^1 \{f(x) + [a(x)u_h'(x)]'\}(e(x) - \pi_h e(x))dx, \end{aligned}$$

Let now $R(u_h(x)) = f(x) + (a(x)u_h'(x))'$, i.e. $R(u_h(x))$ is the *residual error*, which is a well-defined function except in the set $\{x_k\}$, since $(a(x_k)u_x'(x_k))'$ are not defined. Thus we can get the following estimate

$$\begin{aligned} \|e(x)\|_E^2 &= \int_0^1 R(u_h(x))(e(x) - \pi_h e(x))dx = \\ &= \int_0^1 \frac{1}{\sqrt{a(x)}} h(x)R(u_h(x)) \cdot \sqrt{a(x)}\Big(\frac{e(x) - \pi_h e(x)}{h(x)}\Big) dx \\ &\le \Big(\int_0^1 \frac{1}{a(x)} h^2(x)R^2(u_h(x))dx\Big)^{\frac{1}{2}} \Big(\int_0^1 a(x)\Big(\frac{e(x) - \pi_h e(x)}{h(x)}\Big)^2 dx\Big)^{\frac{1}{2}}, \end{aligned}$$

where we have used Cauchy Schwarz inequality. Now recalling the definition of the weighted L_2-norm we have,

$$\Big\|\frac{e(x) - \pi_h e(x)}{h(x)}\Big\|_a = \Big(\int_0^1 a(x)\Big(\frac{e(x) - \pi_h e(x)}{h(x)}\Big)^2 dx\Big)^{\frac{1}{2}}. \qquad (4.5.19)$$

To estimate (4.5.19) we use the third interpolation estimate for $e(x)$ in a subinterval and get

$$\Big\| \frac{e(x)-\pi_h e(x)}{h(x)} \Big\|_a \leq c_i \|e'(x)\|_a = c_i \|e(x)\|_E. \tag{4.5.20}$$

Thus

$$\|e(x)\|_E^2 \leq \Big(\int_0^1 \frac{1}{a(x)} h^2(x) R^2(u_h(x)) dx \Big)^{\frac{1}{2}} \cdot c_i \|e(x)\|_E, \tag{4.5.21}$$

and the proof is complete. □

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 this predicts local mesh refinement, i.e. indicates changing the length of the interval $h(x)$ in the regions (subintervals) which is necessary. More concretely the idea is as follows: Assume that one seeks an error bound less that a given error tolerance TOL:

$$\|e(x)\|_E \leq \text{TOL}. \tag{4.5.22}$$

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 for which $\frac{1}{a(x)} R^2(u_h(x))$ is large and perform the steps (ii) and (iii) again.

4.6 Exercises

Problem 32. *Consider the two-point boundary value problem*

$$-u'' = f, \quad 0 < x < 1; \qquad u(0) = u(1) = 0. \tag{4.6.1}$$

Let $V = \{v : \|v\| + \|v'\| < \infty, \quad v(0) = v(1) = 0\}$.

a. Use V *to derive a variational formulation of* (4.6.1).

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 33. *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*

$$F(v) = \frac{1}{2}\int_0^1 (v')^2\,dx - \int_0^1 fv\,dx. \tag{4.6.2}$$

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 34. *Consider the two-point boundary value problem*

$$-u'' = 1, \quad 0 < x < 1; \qquad u(0) = u(1) = 0. \tag{4.6.3}$$

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* (4.6.3).

b. Prove that $U \in V_h$ *is unique.*

Problem 35. *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

$$\|(u-\pi_h u)'\| \leq C\|hu''\|, \tag{4.6.4}$$

where C is a constant and $\pi_h u$ a piecewise linear interpolant to u.

Problem 36. *Consider the two-point boundary value problem*

$$\begin{aligned} -(au')' &= f, & 0 < x < 1, \\ u(0) &= 0, & a(1)u'(1) = g_1, \end{aligned} \tag{4.6.5}$$

where $a > 0$ is a positive function and g_1 is a constant.

a. Derive the variational formulation of (4.6.5).

b. Discuss how the boundary conditions are implemented.

Problem 37. *Consider the two-point boundary value problem*

$$-u'' = 0, \quad 0 < x < 1; \qquad u(0) = 0, \quad u'(1) = 7. \tag{4.6.6}$$

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 (4.6.6).

b. Calculate by hand the finite element approximation $U \in V_h$ to (4.6.6). Study how the boundary condition at $x = 1$ is approximated.

Problem 38. *Consider the two-point boundary value problem*

$$-u'' = 0, \quad 0 < x < 1; \qquad u'(0) = 5, \quad u(1) = 0. \tag{4.6.7}$$

Let $\mathcal{T}_h : x_j = jh$, $j = 0, 1, \dots, N$, $h = 1/N$ 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 (4.6.7)*.*

b. Compute the finite element approximation $U \in V_h$ assuming $N = 3$.

Problem 39. *Consider the problem of finding a solution approximation to*

$$-u'' = 1, \quad 0 < x < 1; \qquad u'(0) = u'(1) = 0. \tag{4.6.8}$$

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 (4.6.8) *by integrating twice.*

b.Compute a finite element approximation $U \in V_h$ to u if possible.

Problem 40. *Consider the two-point boundary value problem*

$$-((1 + x)u')' = 0, \quad 0 < x < 1; \qquad u(0) = 0, \quad u'(1) = 1. \tag{4.6.9}$$

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 (4.6.9)*.*

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 41. *Consider the two-point boundary value problem*

$$-u'' = 0, \quad 0 < x < 1; \qquad u(0) = 0, \quad u'(1) = k(u(1) - 1). \tag{4.6.10}$$

Let $\mathcal{T}_h : 0 = x_0 < x_1 < x_2 < x_3 = 1$, *where* $x_1 = \frac{1}{3}$ *and* $x_1 = \frac{2}{3}$ *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* (4.6.10) *assuming* $k = 1$.

b. Discuss how the parameter k *influence the boundary condition at* $x = 1$.

Problem 42. *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

$$\frac{1}{h}\begin{bmatrix} -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} \alpha \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \beta \end{bmatrix}. \tag{4.6.11}$$

b. If $\alpha = 2$ *and* $\beta = 3$ *sgow that* (4.6.11) *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 43. *Compute a finite element solution approximation to*

$$-u'' + u = 1; \qquad 0 \le x \le 1, \qquad u(0) = u(1) = 0, \tag{4.6.12}$$

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 44. *Consider the following eigenvalue problem*

$$-au'' + bu = 0; \qquad 0 \le x \le 1, \qquad u(0) = u'(1) = 0, \tag{4.6.13}$$

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.*

Chapter 5

Scalar Initial Value Problem

Consider the following ordinary differential equation (ODE)

$$\begin{array}{ll} \text{(DE)} & \left\{ \begin{array}{r} \dot{u}(t) + a(t)u(t) = f(t), \quad 0 < t \leq T \\ u(0) = u_0 \end{array} \right. \\ \text{(IV)} & \end{array} \tag{5.0.1}$$

where $f(t)$ is the source term and $\dot{u}(t) = \dfrac{du}{dt}$. Here $a(t)$ is a bounded function. For $a(t) \geq 0$ (5.0.1) is called a *parabolic problem*, while $a(t) > 0$ yields a *dissipative problem*. Below first we give a few analytic aspects

5.1 Fundamental solution and stability

Theorem 14 (Fundamental solution)**.** *The solution for the ODE* (5.0.1) *is given by*

$$u(t) = u_0 \cdot e^{-A(t)} + \int_0^t e^{-(A(t)-A(s))} f(s)ds, \tag{5.1.1}$$

where $A(t) = \int_0^t a(s)ds$ *is the integrating factor.*

Proof. Multiplying the (DE) by the integrating factor $e^{A(t)}$ we get

$$\dot{u}(t)e^{A(t)} + \dot{A}(t)e^{A(t)}u(t) = e^{A(t)}f(t), \quad \text{i.e.} \quad \frac{d}{dt}[u(t)e^{A(t)}] = e^{A(t)}f(t),$$

where we used that $a(t) = \dot{A}(t)$. Integrating over $(0, t)$ yields

$$\int_0^t \frac{d}{ds}[u(s)e^{A(s)}]ds = \int_0^t e^{A(s)}f(s)ds \Longleftrightarrow u(t)e^{A(t)} - u(0)e^{A(0)} = \int_0^t e^{A(s)}f(s)ds.$$

Now since $A(0) = 0$ and $u(0) = u_0$ we get the desired result

$$u(t) = u_0 \cdot e^{-A(t)} + \int_0^t e^{-(A(t)-A(s))} f(s)ds. \tag{5.1.2}$$

□

Theorem 15 (Stability estimates)**.** *Using the fundamental solution we can derive the following stability estimates:*

(i) If $a(t) \geq \alpha > 0$, then $|u(t)| \leq e^{-\alpha t}|u_0| + \frac{1}{\alpha}(1 - e^{-\alpha t}) \max_{0 \leq s \leq t} |f(s)|$

(ii) If $a(t) \geq 0$ (i.e. $\alpha = 0$ the parabolic case), then

$$|u(t)| \leq |u_0| + \int_0^t |f(s)|ds \text{ or } |u(t)| \leq |u_0| + \|f\|_{L_1} \tag{5.1.3}$$

Proof. (i) For $a(t) \geq 0$, $\forall t > 0$, we have that $A(t) = \int_0^t a(s)ds$ is non-decreasing and $A(t) - A(s) \geq 0$, $\forall t > s$. For $a(t) \geq \alpha > 0$ we have $A(t) = \int_0^t a(s)ds \geq \int_0^t \alpha \cdot ds = \alpha t$. Further

$$A(t) - A(s) = \int_s^t a(r)\, dr \geq \alpha(t - s). \tag{5.1.4}$$

Thus $e^{-A(t)} \leq e^{-\alpha t}$ and $e^{-(A(t)-A(s))} \leq e^{-\alpha(t-s)}$. Hence using (5.1.2) we get

$$u(t) \leq u_0 \cdot e^{-\alpha t} + \int_0^t e^{-\alpha(t-s)} \max_{0 \leq s \leq t} |f(s)|ds, \tag{5.1.5}$$

which after integration gives that

$$|u(t)| \leq e^{-\alpha t}|u_0| + \max_{0 \leq s \leq t} |f(s)| \left[\frac{1}{\alpha} e^{-\alpha(t-s)}\right]_{s=0}^{s=t}$$

$$|u(t)| \leq e^{-\alpha t}|u_0| + \frac{1}{\alpha}(1 - e^{-\alpha t}) \max_{0 \leq s \leq t} |f(s)|.$$

(ii) Let $\alpha = 0$ in (5.1.5) then $|u(t)| \leq |u_0| + \int_0^t |f(s)|ds$, and the proof is complete. □

Remark 12. *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, as the test space.*

5.2 Galerkin finite element methods (FEM) for IVP

To start, *for discretization in time* we shall introduce some general class of piecewise polynomial test and trial functions. However, in most of our studies in this notes we shall restrict ourselves to two simple cases:
• **cG(1)**, *continuous Galerkin of degree 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 are indifferent spaces.
• **dG(0)**, *Discontinuous Galerkin of degree 0*: Here both the trial and test functions are piecewise constant and discontinuous, i.e. *like the cG(1) for BVP they are in the same space of functions, however, they are of one lower degree (piecewise constant) and discontinuous.*
Generally we have
• **gG(q)**, *Global Galerkin of degree q:* Formulated for our initial value problem (5.0.1) as follows: Find $U \in \mathcal{P}^q(0,T)$ with $U(0) = u_0$ such that

$$\int_0^T (\dot{U} + aU)v dt = \int_0^T f v\, dt, \quad \forall v \in \mathcal{P}^q(0,T), \text{ with } \quad v(0) = 0, \tag{5.2.1}$$

where $v := \{t, t^2, \dots, t^q\} := span[t, t^2, \dots, t^q]$.
• **cG(q)**, *Continuous Galerkin of degree q:* Find $U \in \mathcal{P}^q(0,T)$ with $U(0) = u_0$ such that

$$\int_0^T (\dot{U} + aU)v dt = \int_0^T f v dt, \quad \forall v \in \mathcal{P}^{q-1}(0,T), \tag{5.2.2}$$

where now $v := \{1, t, t^2, \dots, t^{q-1}\}$.
Note the difference between the two test function spaces above.

Example 25. *Consider cG(q) with $q = 1$ then $t^{q-1} = t^0 = 1$ and $v \equiv 1$, thus*

$$\int_0^T (\dot{U} + aU)v dt = \int_0^T (\dot{U} + aU) dt = U(T) - U(0) + \int_0^T aU(t) dt \tag{5.2.3}$$

But $U(t)$ is a linear function through $U(0)$ and the unknown quantity $U(T)$, thus

$$U(t) = U(T)\frac{t}{T} + U(0)\frac{T-t}{T}, \tag{5.2.4}$$

inserting $U(t)$ *in* (5.2.3) *we get*

$$U(T) - U(0) + \int_0^T a\Big(U(T)\frac{t}{T} + U(0)\frac{T-t}{T}\Big)\, dt = \int_0^T f\, dt. \qquad (5.2.5)$$

which gives us $U(T)$ *and consequently, through (5.2.4) and a given* $U(0)$, $U(t)$. *Using this idea we can formulate:*

• **The cG(1) Algorithm** *for the partition* $\mathcal{T}_k$ *of* $[0,T]$ *to subintervals* $I_k = (t_{k-1}, t_k]$.

(1) Given $U(0) = U_0$, *apply (5.2.5) to* $(0, t_1]$ *and compute* $U(t_1)$. *Then using* (5.2.4) *one gets automatically* $U(t), \forall t \in [0, t_1]$.

(2) Assume that U *is computed in all the successive intervals* $(t_{k-1}, t_k]$, $k = 0, 1, n-1$.

(3) Compute $U(t)$ *for* $t \in (t_{n-1}, t_n]$.

This is done through applying (5.2.5) *to the interval* $(t_{n-1}, t_n]$, *instead of* $(0,T]$: *i.e. with* $U_n := U(t_n)$ *and* $U_{n-1} := U(t_{n-1})$,

$$U_n - U_{n-1} + \int_{t_{n-1}}^{t_n} a\Big(\frac{t-t_{n-1}}{t_n - t_{n-1}}U_n + \frac{t_n - t}{t_n - t_{n-1}}U_{n-1}\Big)dt = \int_{t_{n-1}}^{t_n} f dt.$$

Now since U_{n-1} *is known we can calculate* U_n *and then* $U(t), t \in (t_{n-1}, t_n]$ *is determined by the nth-version of the relation formula* (5.2.4)*:*

$$U(t) = U_n \frac{t}{t_n} + U_{n-1}\frac{t_n - t}{t_n}.$$

Global forms

•**Continuous Galerkin cG(q):** Find $U(t) \in V_k^{(q)}$, such that $U(0) = U_0$ and

$$\int_0^{t_n} (\dot{U} + aU) w dt = \int_0^{t_n} f w dt, \qquad \forall w \in W_k^{(q-1)}, \qquad (5.2.6)$$

$$V_k^{(q)} = \{v : v \text{ continuous piecewise polynomials of degree } q \text{ on } \mathcal{T}_k\},$$

$$W_k^{(q-1)} = \{w : w \text{ discontinuous piecewise polynomials of degree } q{-}1 \text{ on } \mathcal{T}_k\}.$$

•**Discontinuous Galerkin dG(q):** Find $U(t) \in \mathcal{P}^q(0,T)$ such that

$$\int_0^T (\dot{U} + aU)v dt + a(U(0) - u(0))v(0) = \int_0^T f v dt, \quad \forall v \in \mathcal{P}^q(0,T). \quad (5.2.7)$$

This approach gives up the requirement that $U(t)$ satisfies the initial condition. Instead, the initial condition is represented by $U(0) - u(0) \neq 0$.

In the sequel we shall use the following notation:
Let $v^{\pm n} = \lim_{s \to 0^+} v(t_n \pm s)$ and $[v_n] = v_n^+ - v_n^-$ is the *jump* in $v(t)$ at time t.

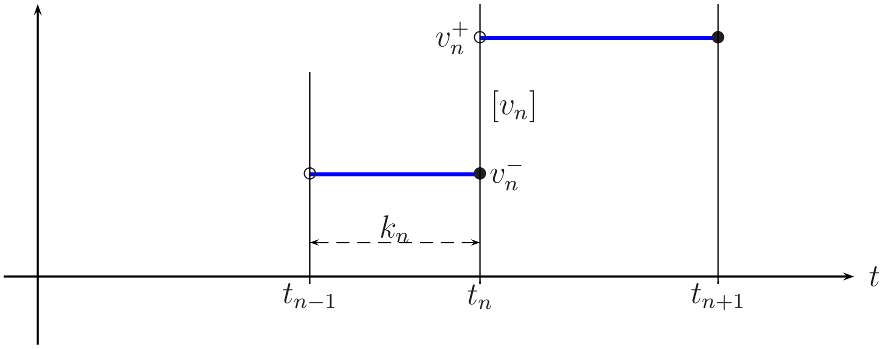


Figure 5.1: The jump $[v_n]$ and the right and left limits $v^\pm$

Then **dG(q)** reads as follows: For $n = 1, \ldots, N$ find $U(t) \in \mathcal{P}^q(t_{n-1}, t_n)$ such that

$$\int_{t_{n-1}}^{t_n} (\dot{U} + aU)v dt + U_{n-1}^+ v_{n-1}^+ = \int_{t_{n-1}}^{t_n} f v dt + U_{n-1}^- v_{n-1}^+, \quad \forall v \in \mathcal{P}^q(t_{n-1}, t_n). \quad (5.2.8)$$

Let $q = 0$, then $v \equiv 1$ is the only base function and 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 $q = 0$ (5.2.8) gives the dG(0) formulation: For $n = 1, \ldots, N$ find piecewise constants U_n such that

$$\int_{t_{n-1}}^{t_n} aU_n dt + U_n = \int_{t_{n-1}}^{t_n} f dt + U_{n-1}. \quad (5.2.9)$$

Finally summing over n in (5.2.8), we get the global dG(q) formulation: Find $U(t) \in W_k^{(q)}$, with $U_0^- = u_0$ such that

$$\sum_{n=1}^N \int_{t_{n-1}}^{t_n} (\dot{U} + aU)w dt + \sum_{n=1}^N [U_{n-1}]w_{n-1}^+ = \int_0^{t_N} f w dt, \quad \forall w \in W_k^{(q)}. \quad (5.2.10)$$

5.3 An a posteriori error estimate for cG(1)

•**The continuous problem** Recall the initial value problem

$$\dot{u}(t) + a(t)u(t) = f(t), \quad \forall t \in (0, T), \qquad u(0) = u_0. \tag{5.3.1}$$

Let us rewrite (5.3.1) in a general *variational form*

$$\int_0^T (\dot{u} + au)v dt = \int_0^T f v dt,$$

for all test functions v. Integrating by parts we get the equivalent equation

$$u(T)v(T) - u(0)v(0) + \int_0^T u(t)\Big(- \dot{v}(t) + av(t)\Big) dt = \int_0^T f v dt. \tag{5.3.2}$$

If we now choose v to be the solution of the *dual problem*:

$$-\dot{v} + av = 0, \quad \text{in } (0, T), \tag{5.3.3}$$

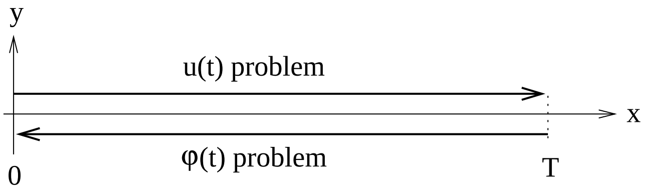
then (5.3.2) is simplified to

$$u(T)v(T) = u(0)v(0) + \int_0^T f v dt, \quad \forall v(t) \in P^q(0, T). \tag{5.3.4}$$

In other words choosing v to be the solution of the dual problem (5.3.3) 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, e.g. a posteriori error analysis as in the proof of the next theorem.
The *Dual problem* for (5.3.1) is formulated as follows: Find $\varphi(t)$ such that

$$\begin{cases} -\dot{\varphi}(t) + a(t)\varphi(t) = 0, & t_N > t \geq 0 \\ \varphi(t_N) = e_N, & e_N = u_N - U_N = u(t_N) - U(t_N). \end{cases} \tag{5.3.5}$$

Note that (5.3.5) runs *backward in time* starting at time $t = t_N$.



Theorem 16. *For $N = 1, 2, \dots$ the cG(1) solution $U(t)$ satisfies*

$$|e_N| \leq S(t_N) \cdot \max_{[0,t_N]} |k\, r(U)|, \tag{5.3.6}$$

where $k = k_n = |I_n|$ for $t \in I_n = (t_{n-1}, t_n)$ is the time step and $r(U) = \dot{U} + aU - f$ is the residual error. Further $S(t_N)$, specified below, is the stability factor satisfying the quantitative bound

$$S(t_N) := \frac{\int_0^{t_N} |\dot{\varphi}| dt}{e_N} \leq \begin{cases} e^{\lambda t_N}, & \textit{if} \quad |a(t)| \leq \lambda, \quad \forall t \\ 1, & \textit{if} \quad a(t) \geq 0, \quad \forall t \end{cases} \tag{5.3.7}$$

Proof. Let $e(t) = u(t) - U(t)$. Using the dual problem $-\dot{\varphi}(t) + a(t)\varphi(t) = 0$ we can write

$$e_N^2 = e_N^2 + 0 = e_N^2 + \int_0^{t_N} e(-\dot{\varphi} + a\varphi)\, dt, \tag{5.3.8}$$

and by partial integration we get

$$\int_0^{t_N} e(-\dot{\varphi} + a(t)\varphi) dt = [-e(t)\varphi(t)]_0^{t_N} + \int_0^{t_N} \dot{e}\varphi\, dt + \int_0^{t_N} ea\varphi\, dt$$

$$= -\underbrace{e(t_N)}_{=e_N} \underbrace{\varphi(t_N)}_{=e_N} + \int_0^{t_N} (\dot{e} + ae)\varphi\, dt = -e_N^2 + \int_0^{t_N} (\dot{e} + ae)\varphi\, dt,$$

where evaluating the boundary term we used $e(0) = 0$. Note that

$$\dot{e}(t) + a(t)e(t) = \dot{u}(t) - \dot{U}(t) + a(t)u(t) - a(t)U(t),$$

and since $f(t) = \dot{u}(t) + a(t)u(t)$ we observe that

$$\dot{e}(t) + a(t)e(t) = f(t) - \dot{U}(t) - a(t)U(t) := -r(U), \tag{5.3.9}$$

where the last equality is just the definition of the residual: $r(U) = \dot{U} + aU - f$. Consequently we get the *error representation formula*:

$$e_N^2 = -\int_0^{t_N} r(U(t))\varphi(t) dt. \tag{5.3.10}$$

To continue we use the interpolant $\pi_k\varphi = \frac{1}{k_n}\int_{I_n}\varphi(s)ds$ of φ and write

$$e_N^2 = -\int_0^{t_N} r(U)(\varphi(t) - \pi_k\varphi(t))dt + \int_0^{t_N} r(U)\pi_k\varphi(t)dt. \qquad (5.3.11)$$

Now from the discrete variational formulation:

$$\int_0^{t_N}(\dot{U} + aU)\pi_k\varphi(t)dt = \int_0^{t_N} f\pi_k\varphi(t)dt \qquad (5.3.12)$$

we have the *Galerkin orthogonality* relation

$$\int_0^{t_N} r(U)\pi_k\varphi(t)dt = 0. \qquad (5.3.13)$$

Thus the final form of the error representation formula is

$$e_N^2 = -\int_0^{t_N} r(U)(\varphi(t) - \pi_k\varphi(t))dt. \qquad (5.3.14)$$

Now applying the *interpolation error* to the function φ in the interval I_n, $|I_n| = k_n$ we have

$$\int_{I_n}|\varphi - \pi_k\varphi|dt \leq k_n\int_{I_n}|\dot{\varphi}|dt. \qquad (5.3.15)$$

This would yield the estimate

$$\int_0^{t_N}|\varphi - \pi_k\varphi|dt = \sum_{n=1}^{N}\int_{I_n}|\varphi - \pi_k\varphi|dt \leq \sum_{n=1}^{N} k_n\int_{I_n}|\dot{\varphi}|dt \qquad (5.3.16)$$

Let now $|v|_J = \max_{t\in J}|v(t)|$, then using (5.3.16) and the final form of the error representation formula (5.3.14) we have that

$$|e_N|^2 \leq \sum_{n=1}^{N}|r(U)|_{I_n}\cdot k_n\int_{I_n}|\dot{\varphi}|dt \leq \max_{1\leq n\leq N}(k_n|r(U)|_{I_n})\int_0^{t_N}|\dot{\varphi}|dt.$$

Now since $\int_0^{t_N}|\varphi|dt = |e_N|\cdot S(t_N)$, (see the definition of $S(t_N)$), we finally get

$$|e_N|^2 \leq |e_N|S(t_N)\max_{[0,t_N]}(k|r(U)|). \qquad (5.3.17)$$

This completes the proof of the first assertion of the theorem.

To prove the second assertion, we claim that:

$$\begin{aligned} &|a(t)| \le \lambda, \quad 0 \le t \le t_N \Longrightarrow |\varphi(t)| \le e^{\lambda t_N}|e_N|, \quad 0 \le t \le t_N && (5.3.18)\\ &|a(t)| \ge 0, \quad \forall t \qquad\qquad \Longrightarrow |\varphi(t)| \le |e_N|, \qquad \forall t \in [0, t_N]. && (5.3.19)\end{aligned}$$

To prove this claim let $s = t_N - t,\ (t = t_N - s)$ and define $\psi(s) = \varphi(t_N - s)$, then using the chain rule

$$\frac{d\psi}{ds} = \frac{d\psi}{dt} \cdot \frac{dt}{ds} = -\dot{\varphi}(t_N - s). \tag{5.3.20}$$

The dual problem is now reformulated as find $\varphi(t)$ such that

$$-\dot{\varphi}(t_N - s) + a(t_N - s)\varphi(t_N - s) = 0. \tag{5.3.21}$$

The corresponding problem for $\psi(s)$:

$$\begin{cases} \dfrac{d\psi(s)}{ds} + a(t_N - s)\psi(s) = 0, & t_N > s \ge 0 \\ \psi(0) = \varphi(t_N) = e_N, & e_N = u_N - U_N = u(t_N) - U(t_N), \end{cases}$$

has the fundamental solution $\psi(s) = Ce^{A(t_N - s)}$, where $\psi(0) = e_N$ implies that $C = e^{-A(t_N)}e_N$ and thus $\psi(s) = e_N\, e^{-A(t_N)}e^{A(t_N-s)} = e_N\, e^{A(t)-A(t_N)}$. Now inserting back in the relation $\psi(s) = \varphi(t),\ t_N - s = t$, we get

$$\varphi(t) = e_N \cdot e^{A(t)-A(t_N)}, \quad \text{and} \quad \dot{\varphi}(t) = e_N \cdot a(t)e^{A(t)-A(t_N)}. \tag{5.3.22}$$

Now the proof of both assertion in the claims are easily followed:

(a) For $|a(t)| \le \lambda$, we have

$$|\varphi(t)| = |e_N|e^{\int_{t_N}^{t} a(s)ds} \le |e_N|e^{\max_t |a(t)|(t_N - t)} \le |e_N|e^{\lambda \cdot t_N} \tag{5.3.23}$$

(b) For $|a(t)| \ge 0$, we have

$$|\varphi(t)| = |e_N|e^{\int_0^{t_N} a(s)ds} \le |e_N|e^{\min_t a(t)(t - t_N)} \tag{5.3.24}$$

and since $(t - t_N) < 0$ we get that $|\varphi(t)| \le |e_N|$.

Now we return to the estimates for $S(t_N)$. Note that for $a(t) \geq 0$ we have using the second relation in (5.3.22) that

$$\begin{aligned}\int_0^{t_N} |\dot{\varphi(t)}|dt &= |e_N| \int_0^{t_N} a(t) e^{A(t)-A(t_N)} dt = |e_N| \cdot [e^{A(t)-A(t_N)}]_0^{t_N} \\ &= |e_N| \cdot \Big(1 - e^{A(0)-A(t_N)}\Big) \leq 1,\end{aligned}$$

which gives that $S(t_N) = \dfrac{\int_0^{t_N} |\dot{\varphi(t)}|dt}{|e_N|} \leq 1$.

As for the case $|a(t)| \leq \lambda$, we use again (5.3.22): $\dot{\varphi}(t) = a(t) e_N \cdot e^{A(t)-A(t_N)}$ and write

$$|\dot{\varphi}(t)| \leq \lambda |e_N| e^{A(t)-A(t_N)} = \lambda |e_N| e^{\int_{t_N}^t a(s)ds} \leq \lambda |e_N| e^{\lambda(t_N - t)}. \qquad (5.3.25)$$

Integrating over $(0, t_N)$ we get

$$\int_0^{t_N} |\dot{\varphi}(t)|dt \leq |e_N| \int_0^{t_N} \lambda e^{\lambda(t_N-t)} dt = |e_N| \Big[-e^{\lambda(t_N-t)} \Big]_0^{t_N} = |e_N|(-1 + e^{\lambda t_N}),$$

which gives that $S(t_N) \leq (-1 + e^{\lambda \cdot t_N}) \leq e^{\lambda \cdot t_N}$, and completes the proof of the second assertion. $\square$

Theorem 17 (Convergence order $\mathcal{O}(k^2)$). *For $N = 1, 2, \ldots$ and with S_N as in the previous theorem, the error for cG(1) solution $U(t)$ satisfies*

$$|e_N| \leq S(t_N) \max_{[0,t_N]} \Big| k^2 \left(aU - f\right) \Big|. \qquad (5.3.26)$$

Proof. Using the orthogonality $[g(t) - \pi_k g(t)] \perp$ (constants) $\forall g(t)$, and since $\dot{U}(t)$ is constant on I_N we have that $\int_0^{t_N} \dot{U}(\varphi - \pi_k \varphi)dt = 0$. Thus using error representation formula (5.3.14) yields

$$\begin{aligned} e_N^2 &= -\int_0^{t_N} r(U)[\varphi(t) - \pi_k \varphi(t)]dt = \int_0^{t_N} (f - aU - \dot{U})(\varphi - \pi_k \varphi)dt \\ &= \int_0^{t_N} (f - aU)(\varphi - \pi_k \varphi)dt - \int_0^{t_N} \dot{U}(\varphi - \pi_k \varphi)dt \\ &= -\int_0^{t_N} (aU - f)(\varphi - \pi_k \varphi)dt. \end{aligned}$$

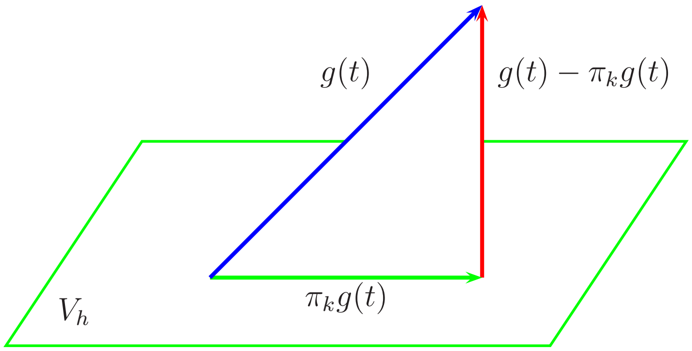


Figure 5.2: Orthogonality: $(g(t) - \pi_k g(t)) \perp$ (constants) $\forall g(t)$.

Similarly using the fact that $\pi_k(aU - f)$ is a constant we get

$$\int_0^{t_N} \pi_k(aU - f)(\varphi - \pi_k\varphi)dt = 0. \tag{5.3.27}$$

Consequently we can write

$$e_N^2 = -\int_0^{t_N} \Big((aU - f) - \pi_k(aU - f)\Big)(\varphi - \pi_k\varphi)dt. \tag{5.3.28}$$

Now using the above theorem and the interpolation error estimate we get

$$\begin{aligned} |e_N| &\leq S(t_N) \cdot \Big|k|(aU - f) - \pi_k(aU - f)|\Big|_{[0,t_N]} \\ &\leq S(t_N) \cdot \Big|k^2 \frac{d}{dt}(aU - f)\Big|_{[0,t_N]}. \end{aligned} \tag{5.3.29}$$

□

5.4 A dG(0) a posteriori error estimate

Theorem 18. *For $N = 1, 2, \dots$, the dG(0) solution $U(t)$ satisfies*

$$|u(t_N) - U_N| \leq S(t_N)|kR(U)|_{[0,t_N]}, \quad U_N = U(t_N) \tag{5.4.1}$$

where

$$R(U) = \frac{|U_N - U_{N-1}|}{k_n} + |f - aU| \quad \textit{for} \quad t_{N-1} < t \leq t_N. \tag{5.4.2}$$

Proof. The proof uses similar techniques as in the cG(1) case. Note that here the residual error includes jump terms and since dual problem satisfies $-\dot{\varphi}(t) + a(t)\varphi(t) = 0$, we can write

$$\begin{aligned} e_N^2 &= e_N^2 + \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} e[-\dot{\varphi}(t) + a(t)\varphi(t)]dt = [PI] = \\ &= e_N^2 + \sum_{n=1}^{N} \Big(\int_{t_{n-1}}^{t_n} (\dot{e} + ae)\varphi(t)dt - [e\varphi]_{t_{n-1}}^{t_n} \Big) \\ &= e_N^2 + \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (f - aU)\varphi dt - \sum_{n=1}^{N} [e\varphi]_{t_{n-1}}^{t_n}, \end{aligned} \tag{5.4.3}$$

where in the last relation we use $\dot{e} + ae = \dot{u} - \dot{U} + au - aU = f - aU$ and also the fact that $U =$ constant $\dot{U} = 0$. We rewrite the last sum as follows

$$\begin{aligned} \sum_{n=1}^{N} (e\varphi)_{t_{n-1}}^{t_n} &= \sum_{n=1}^{N} \Big(e(t_n^-)\varphi(t_n^-) - e(t_{n-1}^+)\varphi(t_{n-1}^+) \Big) \\ &= \{\text{for a given functiong; } g(t_n^-) = g_n^-, g(t_{n-1}^+) = g_{n-1}^+\} \\ &= \sum_{n=1}^{N} (e_n^-\varphi_n^- - e_{n-1}^+\varphi_{n-1}^+) = (e_1^-\varphi_1^- - e_0^+\varphi_0^+) + (e_2^-\varphi_2^- - e_1^+e_1^+) \\ &+ \dots + (e_{N-1}^-\varphi_{N-1}^- - e_{N-2}^+\varphi_{N-2}^+) + (e_N^-\varphi_N^- - e_{N-1}^+\varphi_{N-1}^+). \end{aligned}$$

To continue for $i = 1, \dots N - 1$, we write $\varphi_i^- = (\varphi_i^- - \varphi_i^+ + \varphi_i^+)$, then

$$\begin{aligned} -\sum_{n=1}^{N} (e\varphi)_{t_{n-1}}^{t_n} = -e_N^-\varphi_N^- + e_0^+\varphi_0^+ &- e_1^-(\varphi_1^- - \varphi_1^+ + \varphi_1^+) + e_1^+\varphi_1^+ \\ &- e_2^-(\varphi_2^- - +\varphi_2^+ + \varphi_2^+) + e_2^+\varphi_2^+ \dots \\ &- e_{N-1}^-(\varphi_{N-1}^- - \varphi_{N-1}^+ + \varphi_{N-1}^+) + e_{N-1}^+\varphi_{N-1}^+, \end{aligned}$$

where a general i-th term can be rewritten as

$$\begin{aligned}
&-e_i^-(\varphi_i^- - \varphi_i^+ + \varphi_i^+) + e_i^+\varphi_i^+ = -e_i^-\varphi_i^- + -e_i^-\varphi_i^+ - e_i^-\varphi_i^+ + e_i^+\varphi_i^+\\
&\qquad = e_i^-(\varphi_i^+ - \varphi_i^-) + \varphi_i^+(e_i^+ - e_i^-) = e_i^-[\varphi_i] + \varphi_i^+[e_i],
\end{aligned}$$

with $[g] = g^+ - g^-$ representing the jump. Hence we have

$$-\sum_{n=1}^{N}(e\varphi)|_{t_{n-1}}^{t_n} = -e_N^2 + e_0^+\varphi_0^+ + \sum_{n=1}^{N-1}[e_n]\varphi_n^+ + \sum_{n=1}^{N-1}e_n^-[\varphi_n]. \tag{5.4.4}$$

Inserting in (5.4.3) we get that

$$\begin{aligned}
e_N^2 &= e_N^2 + \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(f - aU)\varphi dt - \sum_{n=1}^{N}[e\varphi]_{t_{n-1}}^{t_n}\\
&= e_N^2 + \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(f - aU)\varphi dt - e_N^2 + e_0^+\varphi_0^+ + \sum_{n=1}^{N-1}[e_n]\varphi_n^+ + \sum_{n=1}^{N-1}[\varphi_n]e_n^- =\\
&= \{\varphi_n,\ u_n \text{ smooth } \Rightarrow [\varphi_n] = 0,\ [u_n] = 0\}\\
&= e_0^+\varphi_0^+ + \sum_{n-1}^{N}\int_{t_{n-1}}^{t_n}(f - aU)\varphi dt + \sum_{n=1}^{N-1}[e_n]\varphi_n^+ = \{[u_n] = 0 \Rightarrow [e_n] = [-U_n]\}\\
&= \sum_{n=1}^{N}\Big(\int_{t_{n-1}}^{t_n}(f - aU)\varphi dt - [U_{n-1}]\varphi_{n-1}^+\Big) =\\
&= \{\text{Galerkin}\} = \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\{(f - aU)(\varphi - \pi_k\varphi) - [U_{n-1}](\varphi - \pi_k\varphi)_{n-1}^+\}dt.
\end{aligned}$$

Now to continue we just follow the previous theorem. □

•Adaptivity for dG(0)

To guarantee that the dG(0) approximation $U(t)$ satisfies

$$|e_N| = |u(t_n) - U(t_n)| \le TOL, \qquad \text{(TOL is a given tolerance)} \tag{5.4.5}$$

we seek to determine the time step k_n so that

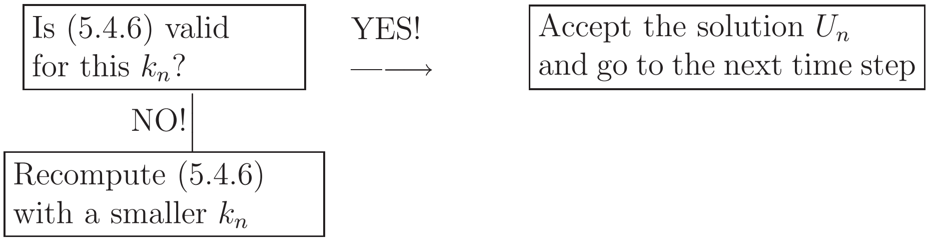
$$S(t_N)\max_{t\in I_n}|k_n R(U)| = TOL, \qquad n = 1, 2, \dots, N. \tag{5.4.6}$$

•An adaptivity algorithm

(i) Compute U_n from U_{n-1} using a predicted step k_n, for example

$$\int_{t_{n-1}}^{t_n} aU_n dt + U_n = \int_{t_{n-1}}^{t_n} f dt + U_{n-1}. \tag{5.4.7}$$

(ii) Compute $|kR(U)|_{I_n} := \max_n |k_n R(U)|$ and follow the chart:


Is (5.4.6) valid for this k_n?
YES!
Accept the solution U_n and go to the next time step
NO!
Recompute (5.4.6) with a smaller k_n

5.5 A priori error analysis

•The discontinuous Galerkin method dG(0).
The dG(0) method for $\dot{u} + au = f$, a=constant, is formulated as follows: Find $U = U(t)$, $t \in I_n$, such that

$$\int_{t_{n-1}}^{t_n} \dot{U} dt + a \int_{t_{n-1}}^{t_n} U dt = \int_{I_n} f dt. \tag{5.5.1}$$

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

$$\int_{t_{n-1}}^{t_n} \dot{U} dt + a \int_{t_{n-1}}^{t_n} U dt = U(t_n) - U(t_{n-1}) + ak_n U_n = U_n - U_{n-1} + ak_n U_n.$$

Hence with a given initial data $u(0) = u_0$, the equation (5.5.1) is written as

$$U_n - U_{n-1} + ak_n U_n = \int_{I_n} f dt \quad n = 1, 2, \dots \qquad U_0 = u_0. \tag{5.5.2}$$

For the *exact* solution $u(t)$ of $\dot{u} + au = f$, the same procedure yields

$$u(t_n) - u(t_{n-1}) + k_n a u_n(t) = \int_{I_n} f dt + k_n a u_n(t) - a \int_{t_{n-1}}^{t_n} u(t) dt, \tag{5.5.3}$$

where we have moved the term $a\int_{t_{n-1}}^{t_n} u(t)dt$ to the right hand side and add $k_n a u_n(t)$ to both sides. Thus from (5.5.2) and (5.5.3) we have that

$$\begin{aligned} (1+k_n a)U_n(t) &= U_{n-1}(t) + \int_{I_n} f dt, \\ (1+k_n a)u_n(t) &= u_{n-1}(t) + \int_{I_n} f dt + k_n a u_n(t) - a\int_{t_{n-1}}^{t_n} u(t)dt. \end{aligned} \tag{5.5.4}$$

Let now $e_n = u_n - U_n$ and $e_{n-1} = u_{n-1} - U_{n-1}$ then (5.5.3) – (5.5.2) yields

$$e_n = (1+k_n a)^{-1}(e_{n-1} + \rho_n) \tag{5.5.5}$$

where $\rho_n := k_n a u_n(t) - a\displaystyle\int_{t_{n-1}}^{t_n} u(t)dt$. Thus in order to estimate the error e_n we need an iteration procedure and an estimate of ρ_n.

Lemma 2. *We have that*

$$|\rho_n| \le \frac{1}{2}|a||k_n|^2 \max_{I_n} |\dot{u}(t)| \tag{5.5.6}$$

Proof. Recalling the definition we have $\rho_n = k_n a u_n(t) - a\int_{t_{n-1}}^{t_n} u(t)dt$. Thus

$$|\rho_n| \le |a||k_n|\Big|u_n - \frac{1}{|k_n|}\int_{I_n} u dt\Big|. \tag{5.5.7}$$

Using a Taylor expansion of the integrand $u(t)$ about t_n, viz

$$u(t) = u_n + \dot{u}(\xi)(t - t_n), \qquad \text{for some} \quad \xi, \quad t_{n-1} < \xi < t_n \tag{5.5.8}$$

we get that

$$\begin{aligned} |\rho_n| &\le |a||k_n|\Big|u_n - \frac{1}{k_n}\int_{I_n}[u_n + \dot{u}(\xi)(t-t_n)]dt\Big| \\ &\le |a||k_n|\Big|u_n - \frac{1}{k_n}k_n u_n - \frac{1}{k_n}\dot{u}(\xi)\Big[\frac{(t-t_n)^2}{2}\Big]_{t_{n-1}}^{t_n}\Big| \\ &= |a||k_n|\Big| - \frac{1}{k_n}\dot{u}(\xi)\Big[0 - \frac{k_n^2}{2}\Big]\Big| = |a||k_n|\Big| - \frac{1}{k_n}\dot{u}(\xi)\frac{k_n^2}{2}\Big| = |a||k_n|^2\frac{1}{2}|\dot{u}(\xi)|. \end{aligned}$$

Thus we have the following final estimate for ρ_n,

$$|\rho_n| \le \frac{1}{2}|a||k_n|^2 \max_{I_n} |\dot{u}(t)| \tag{5.5.9}$$

□

To simplify the estimate for e_n we split, and gather, the proof of technical details in the following lemma:

Lemma 3. *For $k_n|a| \leq 1/2,\ n \geq 1$ we have that*

(i) $(1 - k_n|a|)^{-1} \leq e^{2k_n|a|}$.

(ii) Let $\tau_n = t_N - t_{n-1}$ *then* $|e_N| \leq \frac{1}{2}\sum_{n=1}^{N}(e^{2|a|\tau_n}|a|k_n) \max_{1\leq n\leq N} k_n|\dot{u}|I_n$.

(iii) $\sum_{n=1}^{N} e^{2|a|\tau_n}|a|k_n \leq e\int_0^{t_N} |a|e^{2|a|\tau}d\tau$.

We postpone the proof of this lemma and first show that using these results we can obtain a bound for the error e_N (our main result) viz,

Theorem 19. *If $k_n|a| \leq \frac{1}{2}, n \geq 1$ then the error of the $dG(0)$ approximation U satisfies*

$$|u(t_N) - U(t_N)| = |e_N| \leq \frac{e}{4}\Big(e^{2|a|t_N} - 1\Big) \max_{1\leq n\leq N} k_n|\dot{u}(t)|_{I_n}. \tag{5.5.10}$$

Proof. Using the estimates (ii) and (iii) of the above lemma we have that

$$\begin{aligned}|e_N| &\leq \frac{1}{2}\sum_{n=1}^{N}(e^{2|a|\tau_n}|a|k_n) \max_{1\leq n\leq N} k_n|\dot{u}|_{I_n} \leq \frac{1}{2}\Big(e\int_0^{t_N} |a|e^{2|a|\tau}d\tau\Big) \max_{1\leq n\leq N} k_n|\dot{u}|_{I_n} \\ &= \frac{1}{2}e\Big[\frac{e^{2|a|\tau}}{2}\Big]_0^{t_N} \cdot \max_{1\leq n\leq N} k_n|\dot{u}(t)|_{I_n} = \frac{e}{4}\Big(e^{2|a|t_N} - 1\Big) \max_{1\leq n\leq N} k_n|\dot{u}(t)|_{I_n}.\end{aligned}$$

□

Note that the stability constant $\frac{e}{4}\Big(e^{2|a|t_N} - 1\Big)$ may grow depending on $|a|$ and t_N, and then this result may not be satisfactory at all.

Now we return to the proof of our technical results:

Proof of Lemma 3. (i) For $0 \leq x := k_n|a| \leq 1/2$, we have that $1/2 \leq 1 - x < 1$ and $0 < 1 - 2x \leq 1$. We can now multiply both side of the first claim: $\frac{1}{1-x} < e^{2x}$ by $1 - x \geq 1/2 > 0$ to obtain the equivalent relation

$$f(x) := (1 - x)e^{2x} > 1. \tag{5.5.11}$$

Note that since $f(0) = 1$ and $f'(x) = (1-2x)e^{2x} > 0$ the relation (5.5.11) is valid.
(ii) We recall that $e_n = (1+k_n a)^{-1}(e_{n-1}+\rho_n)$. To deal with the coefficient $(1+k_n a)^{-1}$ first we note that $(1+k_n a)^{-1} \leq (1-k_n a)^{-1}$ if $a \geq 0$. Thus $(1+k_n|a|)^{-1} \leq (1-k_n|a|)^{-1}$, $a \in \mathbb{R}$. Further the assumption $k_n|a| \leq \dfrac{1}{2}$ for $n \geq 1$, combined with (i), implies that $(1-k_n|a|)^{-1} \leq e^{2k_n|a|}$, $n \geq 1$. Thus

$$|e_N| \leq \frac{1}{1-k_N|a|}|e_{N-1}| + \frac{1}{1-k_N|a|}|\rho_N| \leq |e_{N-1}| \cdot e^{2k_N|a|} + |\rho_N| \cdot e^{2k_N|a|}. \tag{5.5.12}$$

Relabeling, e.g. N to $N-1$ we get

$$|e_{N-1}| \leq |e_{N-2}| \cdot e^{2k_{N-1}|a|} + |\rho_{N-1}| \cdot e^{2k_{N-1}|a|} = e^{2k_{N-1}|a|}\Big(|e_{N-2}| + |\rho_{N-1}|\Big),$$

which, inserting in (5.5.12) gives that

$$|e_N| \leq e^{2k_N|a|}e^{2k_{N-1}|a|}\Big(|e_{N-2}| + |\rho_{N-1}|\Big) + |\rho_N| \cdot e^{2k_N|a|}. \tag{5.5.13}$$

Similarly we have $|e_{N-2}| \leq e^{2k_{N-2}|a|}\Big(|e_{N-3}| + |\rho_{N-2}|\Big)$. Now iterating (5.5.13) and using the fact that $e_0 = 0$ we get,

$$\begin{aligned}|e_N| \leq & e^{2k_N|a|}e^{2k_{N-1}|a|}e^{2k_{N-2}|a|}|e_{N-3}| + e^{2k_N|a|}e^{2k_{N-1}|a|}e^{2k_{N-2}|a|}|\rho_{N-2}| \\ & + e^{2k_N|a|}e^{2k_{N-1}|a|}|\rho_{N-1}| + |\rho_N| \cdot e^{2k_N|a|} \leq \cdots \leq \\ \leq & e^{2|a|\sum_{n=1}^N k_n}|e_0| + \sum_{n=1}^N e^{2|a|\sum_{m=n}^N k_m}|\rho_n| = \sum_{n=1}^N e^{2|a|\sum_{m=n}^N k_m}|\rho_n|.\end{aligned}$$

Recalling (5.5.6) (Lemma 2): $|\rho_n| \leq \dfrac{1}{2}|a||k_n|^2 \max_{I_n} |\dot{u}(t)|$. Thus

$$|e_N| \leq \sum_{n=1}^N e^{2|a|\sum_{m=n}^N k_m} \frac{1}{2}|a||k_n|^2 \max_{I_n} |\dot{u}(t)|. \tag{5.5.14}$$

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 we have shown the assertion (ii) of the lemma, i.e.

$$|e_N| \le \sum_{n=1}^{N} e^{2|a|(t_N - t_{n-1})}\frac{1}{2}|a||k_n|^2 \max_{I_n}|\dot{u}(t)| = \frac{1}{2}\sum_{n=1}^{N}(e^{2|a|\tau_n}|a|k_n) \max_{1\le n\le N} k_n|\dot{u}|_{I_n}.$$

(iii) To prove this part we note that

$$\tau_n = t_N - t_{n-1} = (t_N - t_n) + (t_n - t_{n-1}) = \tau_{n+1} + k_n, \tag{5.5.15}$$

and since $|a|k_n \le 1/2$ we have $2|a|\tau_n = 2|a|\tau_{n+1} + 2|a|k_n \le 2|a|\tau_{n+1} + 1$. Further for $\tau_{n+1} \le \tau \le \tau_n$, we can write

$$\begin{aligned} e^{2|a|\tau_n}\cdot k_n &= \int_{\tau_{n+1}}^{\tau_n} e^{2|a|\tau_n} d\tau \le \int_{\tau_{n+1}}^{\tau_n} e^{(2|a|\tau_{n+1}+1)} d\tau \\ &= \int_{\tau_{n+1}}^{\tau_n} e^1 \cdot e^{2|a|\tau_{n+1}} d\tau \le e\int_{\tau_{n+1}}^{\tau_n} e^{2|a|\tau} d\tau. \end{aligned} \tag{5.5.16}$$

Multiplying (5.5.16) by $|a|$ and summing over n we get

$$\begin{aligned} \sum_{n=1}^{N} e^{2|a|\tau_n}|a|k_n &\le e\Big(\sum_{n=1}^{N}\int_{\tau_{n+1}}^{\tau_n} e^{2|a|\tau} d\tau\Big)|a| \\ &= e\int_{\tau_{N+1}}^{\tau_1} e^{2|a|\tau}|a| d\tau = e\int_0^{t_N} |a|e^{2|a|\tau} d\tau, \end{aligned} \tag{5.5.17}$$

which is the desired result and the proof is complete. □

□

5.6 The parabolic case $(a(t) \ge 0)$

We state and proof the basic estimate of this case

Theorem 20. *Consider the $dG(0)$ approximation U for $\dot{u} + au = f$, with $a(t) \ge 0$. Assume that $k_j|a|_{I_j} \le \frac{1}{2}$, $\forall j$, then we have the error estimates*

$$|u(t_N) - U_N| \le \begin{cases} 3e^{2\lambda t_N} \max\limits_{0\le t\le t_N} |k\dot{u}| & \text{if } |a(t)| \le \lambda \\ 3\max\limits_{0\le t\le t_N} |k\dot{u}| & \text{if } a(t) \ge 0. \end{cases} \tag{5.6.1}$$

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 interpolation error with $\pi_k u$ being the L_2-projection into $W_k^{(0)}$. To estimate $\bar{e}$, we shall use the following *discrete dual problem (DDP)*:
Find $\Phi \in W_k^{(0)}$, such that for $n = N, N-1, \dots, 1$.

$$(DDP) \qquad \begin{cases} \displaystyle\int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi) v dt - [\Phi_n] v_n = 0, \quad \forall v \in W_k^{(0)} \\ \Phi_N^+ = \Phi_{N+1} = (\pi_k u - U)_N :\equiv \bar{e}_N. \end{cases} \tag{5.6.2}$$

Let now $v = e$, then

$$|\bar{e}_N|^2 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi) \bar{e} dt - \sum_{n=1}^{N-1} [\Phi_n] \bar{e}_n + \Phi_N \bar{e}_N. \tag{5.6.3}$$

We now use $\bar{e} = (\pi_k u - U) = (\pi_k u - u + u - U)$ and write (5.6.3) as

$$\begin{aligned} |e_N|^2 = & \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} [-\dot{\Phi} + a(t)\Phi](\pi_k u - u + u - U) UT \\ & - \sum_{n=1}^{N-1} [\Phi_n](\pi_k u - u + u - U)_n + \Phi_N (\pi_k u - u + u - U)_N. \end{aligned}$$

Using Galerkin orthogonality we replace u by U. Therefore the total contribution from the terms with the factor $u - U$ is identical to zero. Thus due to the fact that $\dot{\Phi} = 0$ on each subinterval, we have the error representation formula:

$$\begin{aligned} |e_N|^2 = & \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi)(\pi_k u - u) dt - \sum_{n=1}^{N-1} [\Phi_n](\pi_k u - u)_n + \Phi_N (\pi_k u - u)_N \\ = & \int_0^{t_N} (a(t)\Phi)(u - \pi_k u)\, dt + \sum_{n=1}^{N-1} [\Phi_n](u - \pi_k u)_n - \Phi_N (u - \pi_k u)_N. \end{aligned}$$

To continue we shall need the following results: □

Lemma 4. *If $|a(t)| \leq \lambda, \forall t \in (0, t_N)$ and $k_j |a|_{I_j} \leq \frac{1}{2}, j = 1, 2, \dots, N$, then the solution of the discrete dual problem satisfies*

(i) $|\Phi_n| \leq e^{2\lambda(t_N - t_{n-1})} |\bar{e}_N|$.

(ii) $\displaystyle\sum_{n=1}^{N-1}|[\Phi_n]|\le e^{2\lambda t_N}|\bar{e}_N|.$

(iii) $\displaystyle\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}a(t)|\Phi_n|dt\le e^{2\lambda t_N}|\bar{e}_N|.$

(iv) If $a(t)\ge 0$ *then*

$$Max\Big(|\Phi_n|,\,\sum_{n=1}^{N-1}|[\Phi_n]|,\,\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}a(t)|\Phi_n|\,dt\Big)\le|\bar{e}_N|.$$

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 dual problem with $v\equiv 1$:

$$(DDP)\qquad\begin{cases}\displaystyle\int_{t_{n-1}}^{t_n}(-\dot{\Phi}+a(t)\Phi)dt-[\Phi_n]=0,\\ \Phi_{N+1}=(\pi_k u-U)_N:\equiv\bar{e}_N.\end{cases}\tag{5.6.4}$$

For dG(0) this becomes

$$(DDP)\quad\begin{cases}-\Phi_{n+1}+\Phi_n+\Phi_n\int_{t_{n-1}}^{t_n}a(t)=0,\quad n=N,N-1,\dots,1\\ \Phi_{N+1}=\bar{e}_N,\qquad\Phi_n=\Phi|_{I_n}.\end{cases}\tag{5.6.5}$$

By iterating we get

$$\Phi_n=\prod_{j=n}^{N}\Big(1+\int_{I_j}a(t)dt\Big)^{-1}\Phi_{N+1}\tag{5.6.6}$$

For $a(t)\ge 0$ we have $\Big(1+\int_{I_j}a(t)dt\Big)^{-1}\le 1$, thus (5.6.6) implies that

$$|\Phi_n|\le\Phi_{N+1}=|\bar{e}_N|.\tag{5.6.7}$$

Further we have using (5.6.6) 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

$$[\Phi_n] = \Phi_n^+ - \Phi_n^- = \Phi_{n+1} - \Phi_n \geq 0. \tag{5.6.8}$$

Thus

$$\begin{aligned} \sum_{n=1}^{N} |[\Phi_n]| &= \Phi_{N+1} - \Phi_N + \Phi_N - \Phi_{N-1} + \ldots + \Phi_2 - \Phi_1 \\ &= \Phi_{N+1} - \Phi_1 \leq \Phi_{N+1} \leq |\bar{e}_N|. \end{aligned} \tag{5.6.9}$$

Finally in the discrete equation:

$$\int_{t_{n-1}}^{t_n} (-\dot{\Phi} + a(t)\Phi) v dt - [\Phi_n] v_n = 0, \qquad \forall v \in W_k^{(0)} \tag{5.6.10}$$

we have $v \equiv 1$ and $\dot{\Phi} \equiv 0$ for the dG(0). Hence (5.6.10) can be rewritten as

$$\int_{t_{n-1}}^{t_n} a(t)\Phi_n dt = [\Phi_n]. \tag{5.6.11}$$

Summing over n, this gives that

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} a(t)\Phi_n dt \leq \sum_{n=1}^{N} [\Phi_n] \leq |\bar{e}_n|. \tag{5.6.12}$$

Combining (5.6.7), (5.6.9), and (5.6.12) the proof of (iv) is now complete. □

•**Quadrature rule for** f: Assume that a=constant. Then the error representation formula, combining $dG(0)$, with the quadrature role for f is as follows:

$$\begin{aligned} e_N^2 = &\sum_{n=1}^{N} \Big(\int_{t_{n-1}}^{t_n} (f - aU)(\varphi - \pi_k\varphi) dt - [U_{n-1}](\varphi - \pi_k\varphi)_{n-1}^+ \\ &+ \underbrace{\int_{t_{n-1}}^{t_n} f\pi_k\varphi dt - (\overline{f\pi_k\varphi})_n k_n \Big)}_{\text{quadrature error}} \end{aligned} \tag{5.6.13}$$

where 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 define the *weak stability factor* $\tilde{S}(t_N) := \dfrac{\int_0^{t_N} |\varphi| dt}{|e_N|}$, where φ is the solution of the dual problem

$$-\dot{\varphi} + a\varphi = 0, \qquad \text{for} \quad t_N > t \geq 0 \qquad \varphi(t_N) = e_N.$$

Note that $\pi_k\varphi$ is piecewise constant and

$$\int_{I_n} |\pi_k\varphi(t)| dt \leq \int_{I_n} |\varphi(t)| dt.$$

We can prove the following relations between the two stability factors:

$$\tilde{S}(t_N) \leq t_N(1 + S(t_N)).$$

Note that if $a > 0$ is sufficiently small, then $\tilde{S}(t_N) >> S(t_N)$.

Theorem 21 (The modified a posteriori estimate for dG(0))**.** *The dG*(0) *approximation* $U(t)$ *computed using quadrature on terms involving* f *satisfies for* $N = 1, 2, \ldots$

$$|u(t_n) - U_n| \leq S(t_n)|kR(U)|_{(0,t_N)} + \tilde{S}(t_N)C_j|k^j f^{(j)}|_{(0,t_N)}, \tag{5.6.14}$$

where

$$R(U) = \frac{|U_n - U_{n-1}|}{k_n} + |f - aU|, \qquad on \quad I_n \tag{5.6.15}$$

and $j = 1$ *for the rectangle rule,* $j = 2$ *for the midpoint rule,* $C_1 = 1$, $C_2 = \frac{1}{2}$, $f^{(1)} = \dot{f}$ *and* $f^{(2)} = \ddot{f}$.

5.6.1 Short summary of error estimates

In this part we shall derive some short variants for the error estimates above

Lemma 5. *Let* U *be the cG(1) approximation of* u *satisfying the initial value problem*

$$\dot{u} + u = f, \quad t > 0, \quad u(0) = u_0. \tag{5.6.16}$$

Then we have that

$$|(u - U)(T)| \leq \max_{[0,T]} |k(f - \dot{U} - U)|, \tag{5.6.17}$$

where k *is the time step.*

Proof. The error $e = u - U$ satisfies Galerkin orthogonality:

$$\int_0^T (\dot{e} + e) v dt = 0, \qquad \text{for all piecewise constants } v(t). \tag{5.6.18}$$

Let φ satisfy the dual equation

$$-\dot{\varphi} + \varphi = 0, \quad t < T, \quad \varphi(T) = e(T). \tag{5.6.19}$$

Then we have that $\varphi(t) = e(T) \cdot e^{t-T}$: Note that integrating $-\dot{\varphi} + \varphi = 0$ gives

$$\int \frac{\dot{\varphi}}{\varphi} dt = \int 1 \cdot dt \Longrightarrow \ln \varphi = t + C. \tag{5.6.20}$$

Let now $C = \ln C_1$, then (5.6.20) can be written as

$$\ln \varphi - \ln C_1 = \ln \frac{\varphi}{C_1} = t \Longrightarrow \varphi(t) = C_1 \cdot e^t. \tag{5.6.21}$$

Finally, since $\varphi(T) = e(T)$ we have that

$$C_1 \cdot e^T = e(T), \quad \text{i.e.} \quad C_1 = e(T) \cdot e^{-T} \Longrightarrow \varphi(t) = e(T) \cdot e^{t-T}. \tag{5.6.22}$$

To continue we have using $-\dot{\varphi} + \varphi = 0$,

$$|e(T)|^2 = e(T) \cdot e(T) + \int_0^T e(-\dot{\varphi} + \varphi) dt = e(T) \cdot e(T) - \int_0^T e\dot{\varphi}\, dt + \int_0^T e\varphi dt.$$

Note that integration by parts gives

$$\int_0^T e\dot{\varphi} dt = [e \cdot \varphi]_{t=0}^T - \int_0^T \dot{e}\varphi dt = e(T)\varphi(T) - e(0)\varphi(0) - \int_0^T \dot{e}\varphi dt.$$

Using $\varphi(T) = e(T)$, and $e(0) = 0$, we thus have

$$\begin{aligned} |e(T)|^2 &= e(T) \cdot e(T) - e(T) \cdot e(T) + \int_0^T \dot{e}\varphi\, dt + \int_0^T e\varphi\, dt = \int_0^T (\dot{e} + e)\varphi\, dt \\ &= \int_0^T (\dot{e} + e)(\varphi - v) dt = \int_0^T \Big(\underbrace{\dot{u} + u}_{=f} - \dot{U} - U \Big)(\varphi - v) dt. \end{aligned}$$

We have that $\dot{U} + U - f := r(U)$, is the residual and

$$|e(T)|^2 = -\int_0^T r(U) \cdot (\varphi - v) dt \leq \max_{[0,T]} |k \cdot r(U)| \int_0^T \frac{1}{k} |\varphi - v| dt. \tag{5.6.23}$$

Recall that

$$\int_I h^{-1}|\varphi - v|dx \leq \int_I |\varphi'|dx. \tag{5.6.24}$$

Further $-\dot{\varphi} + \varphi = 0$ implies $\dot{\varphi} = \varphi$, and $\varphi(t) = e(T) \cdot e^{t-T}$. Thus

$$\begin{aligned} |e(T)|^2 &\leq \max_{[0,T]} |k \cdot r(U)| \int_0^T |\dot{\varphi}|dt = \max_{[0,T]} |k \cdot r(U)| \int_0^T |\varphi(t)|\, dt \\ &\leq \max_{[0,T]} |kr(U)|e(T)| \int_0^T e^{t-T} dt, \end{aligned} \tag{5.6.25}$$

and since

$$\int_0^T e^{t-T} dt = [e^{t-T}]_0^T = e^0 - e^{-T} = 1 - e^{-T} \leq 1, \quad T > 0,$$

we finally end up with the desired result

$$|e(T)| \leq \max_{[0,T]} |k \cdot r(U)|.$$

□

Problem 45. *Generalize the Lemma to the problem $\dot{u} + au = f$, with $a =$ positive constant.*

Is the statement of Lemma 1 valid for $\dot{u} - u = f$?

Problem 46. *Study the dG(0)-case for $\dot{u} + au = f, \quad a > 0$*

Lemma 6. *Let $\dot{u} + u = f, t > 0$. Show for the cG(1)-approximation $U(t)$ that*

$$|(u - U)(T)| \leq \max_{[0,T]} |k^2 \ddot{u}|T. \tag{5.6.26}$$

Sketchy proof, via the dual equation. Let φ be the dual solution satisfying

$$\dot{\varphi} + \varphi = 0,\ t < T, \quad \varphi(T) = e(T).$$

We compute the error at time T, viz

$$\begin{aligned} |e(T)|^2 = |\Theta(T)|^2 = \Theta(T)\varphi(T) + \underbrace{\int_0^T \bar{\Theta}(-\dot{\Phi} + \Phi)\, dt}_{=0} = \int_0^T (\dot{\Theta} + \Theta)\bar{\Phi} dt \\ = -\int_0^T (\dot{\rho} + \rho)\bar{\Phi}\, dt = -\int_0^T \rho \cdot \bar{\Phi}\, dt \leq \max_{[0,T]} |k^2 \ddot{u}| \int_0^T |\bar{\Phi}|\, dt \\ \leq \max_{[0,T]} |k^2 \ddot{u}| \cdot T \cdot |e(T)|. \end{aligned}$$

Here $\rho = u - \hat{u}, \Theta = \hat{u} - U$ and Φ is cG(1)-approximation of ϕ such that $\int_0^T v(-\dot{\Phi} + \Phi)\,dt = 0$ for all piecewise constant $v(t)$. Furthermore $\hat{u}$ is the piecewise linear interpolant of u and $\bar{w} =$ is the piecewise constant mean value. □

5.7 Exercises

Problem 47. *(a) Derive the stiffness matrix and load vector in piecewise polynomial (of degree q) approximation for the following ODE in population dynamics,*

$$\begin{cases} \dot{u}(t) = \lambda u(t), & \textit{for } 0 < t \leq 1, \\ u(0) = u_0. \end{cases}$$

(b) Let $\lambda = 1$ and $u_0 = 1$ and determine the approximate solution $U(t)$, for $q = 1$ and $q = 2$.

Problem 48. *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

$$|u(t_N) - U_N| \leq \max_{[0,t_N]} |k(\dot{U} + aU - f)|, \quad k = k_n, \ \textit{for } t_{n-1} < t \leq t_n.$$

Problem 49. *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 Cranck-Nicholson approximation, (i.e. $|U_n| = 1$), but NOT in any of the Euler approximations.

Problem 50. *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 51. *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 $a = 1$,

$$|(u-U)(T)| \le \min\Big(||k(f - \dot{U} - U)||_{L^\infty[0,T]}, T||k^2 \ddot{u}||_{L^\infty[0,T]}\Big).$$

Problem 52. *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 $a(t) \ge a_0 > 0$, we have the stability estimate

$$|u(t)| \le e^{-a_0 t}\left(|u_0| + \int_0^t e^{a_0 s}|f(s)|\, ds\right)$$

(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 $a(t) \ge 0$, $f(t) \equiv 0$, and estimate the quantity $\frac{\int_0^T |\dot{u}|\, dt}{|u_0|}$.

Problem 53. *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

$$(Av, v) \ge \alpha ||v||^2, \qquad \forall v,$$

then the solution u of the initial value problem satisfies the stability estimate

$$||u(t)||^2 + \alpha \int_0^t ||u(s)||^2\, ds \le ||u_0||^2 + \frac{1}{\alpha}\int_0^t ||f(s)||^2\, ds.$$

Chapter 6

The heat equation in 1d

In this chapter we focus on some basic stability and finite element error estimates for the one-space dimensional heat equation. A general discussion on classical heat equation can be found in our *Lecture Notes in Fourier Analysis*. We start our study considering an example of an initial boundary value problem with mixed boundary conditions. Higher dimensional case is considered in forthcoming lecture notes based on the present one. Here we consider an example of an initial boundary value problem for the heat equation, viz

$$(IBVP) \qquad \begin{cases} \dot{u} - u'' = f(x), & 0 < x < 1, \quad t > 0, \\ u(x,0) = u_0(x), & 0 < x < 1, \\ u(0,t) = u_x(1,t) = 0, & \qquad\qquad\quad t > 0, \end{cases} \tag{6.0.1}$$

where we have used the following differentiation notation in the $1 - D$ case:

$$\dot{u} := u_t = \frac{\partial u}{\partial t}, \quad u' := u_x = \frac{\partial u}{\partial x}, \quad u'' := u_{xx} = \frac{\partial^2 u}{\partial x^2}.$$

Note that the partial differential equation in (6.0.1) containing three derivatives yields three degrees of freedom and therefore, to determine a unique solution, it is necessary to supply three data: here two boundary condition associated to two spatial derivatives (in u'') and an initial condition corresponding to the time derivative ($\dot{u}$). To have an idea we formulate an example, viz

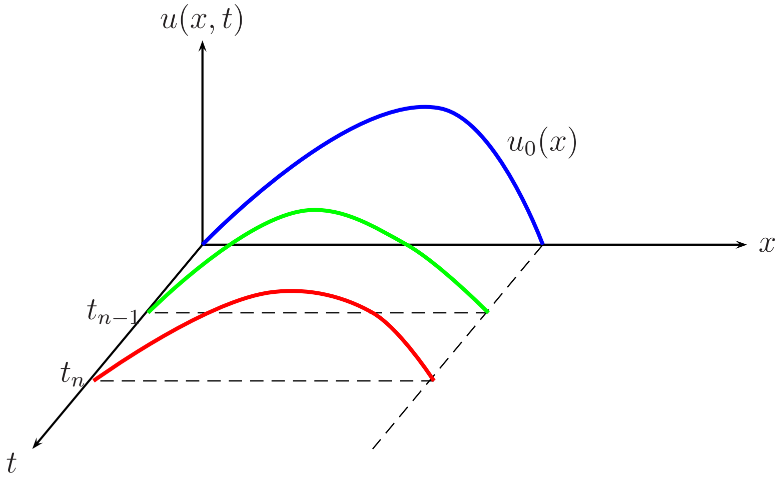


Figure 6.1: A decreasing temperature profile

Problem 54. *Give physical meaning to the IBVP* (6.0.1) *where* $f = 20 - u$.

solution: Heat conduction with

$u(x,t) =$ temperature at x at time t.

$u(x,0) = u_0(x)$, the initial temperature at time $t = 0$.

$u(0,t) = 0$, fixed temperature at time $x = 0$.

$u'(1,t) = 0$, isolated boundary at $x = 1$ (no hear flux).

$f = 20 - u$, heat source, in this case a control system to force $u \Rightarrow 20$.

6.1 Stability estimates

In this part we shall derive a general stability estimate for the mixed IBVP above, prove a $1-D$ version of the Poincare inequality and then derive some homogeneous stability estimates.

Theorem 22. *The IBVP* (6.0.1) *satisfies the stability estimates*

$$||u(\cdot,t)|| \leq ||u_0|| + \int_0^t ||f(\cdot,s)||\,ds, \tag{6.1.1}$$

$$||u_x(\cdot,t)||^2 \leq ||u_0'||^2 + \int_0^t ||f(\cdot,s)||^2\,ds. \tag{6.1.2}$$

Proof. Multiply the equation in (6.0.1) by u and integrate over $(0,1)$ to get

$$\int_0^1 \dot{u}u\,dx - \int_0^1 u''u\,dx = \int_0^1 fu\,dx. \tag{6.1.3}$$

Integrating by parts we get

$$\frac{1}{2}\frac{d}{dt}\int_0^1 u^2\,dx + \int_0^1 (u')^2\,dx - u'(1,t)u(1,t) + u'(0,t)u(0,t) = \int_0^1 fu\,dx,$$

and using the boundary conditions and the Cauchy-Schwartz inequality we end up with

$$||u||\frac{d}{dt}||u|| + ||u'||^2 = \int_0^1 fu\,dx \le ||f||||u||. \tag{6.1.4}$$

Consequently

$$||u||\frac{d}{dt}||u|| \le ||f||||u||, \quad \text{and thus} \quad \frac{d}{dt}||u|| \le ||f||. \tag{6.1.5}$$

Integrating over time we get

$$||u(\cdot,t)|| - ||u(\cdot,0)|| \le \int_0^t ||f||\,ds, \tag{6.1.6}$$

which gives (6.1.1). To prove (6.1.2) we multiply the differential equation by $\dot{u}$ and integrate over $(0,1)$ to obtain

$$\begin{aligned}\int_0^1 (\dot{u})^2\,dx - \int_0^1 u''\dot{u}\,dx &= ||\dot{u}||^2 + \int_0^1 u'\dot{u}'\,dx - u'(1,t)\dot{u}(1,t) + u'(0,t)\dot{u}(0,t)\\ &= \int_0^1 f\dot{u}\,dx.\end{aligned}$$

The expression above gives

$$||\dot{u}||^2 + \frac{1}{2}\frac{d}{dt}||u'||^2 = \int_0^1 f\dot{u}\,dx \le ||f||||\dot{u}|| \le \frac{1}{2}\Big(||f||^2 + ||u'||^2\Big). \tag{6.1.7}$$

Thus

$$\frac{1}{2}||\dot{u}||^2 + \frac{1}{2}\frac{d}{dt}||u'||^2 \le \frac{1}{2}||f||^2, \tag{6.1.8}$$

and hence

$$\frac{d}{dt}||u'||^2 \le ||f||^2. \tag{6.1.9}$$

Now integrating over $(0, t)$ we get the desired result

$$||u'(\cdot,t)||^2 - ||u'(\cdot,0)||^2 \leq \int_0^t ||f(\cdot,s)||^2\,ds, \tag{6.1.10}$$

and the proof is complete. □

To continue we prove a one-dimensional version of the one of the most important inequalities in PDE and analysis.

Theorem 23 (Poincare inequality in $1-D$ case)**.** *Assume that u and u' are square integrable. There exists a constant C, independent of u but dependent of L, such that if $u(0) = u(L) = 0$, then there is constant C, independent of u but dependent of L, such that*

$$\int_0^L u(x)^2\,dx \leq C\int_0^L u'(x)^2\,dx, \quad i.e. \quad ||u|| \leq \sqrt{C}||u'||. \tag{6.1.11}$$

Proof. Note that we can successively write

$$\begin{aligned} u(x) &= \int_0^x u'(y)\,dy \leq \int_0^x |u'(y)|\,dy \leq \int_0^x |u'(y)|\cdot 1\,dy \\ &\leq \Big(\int_0^L |u'(y)|^2\,dy\Big)^{1/2}\cdot\Big(\int_0^L 1^2 dy\Big)^{1/2} = \sqrt{L}\Big(\int_0^L |u'(y)|^2\,dy\Big)^{1/2}. \end{aligned}$$

Thus

$$\int_0^L u(x)^2\,dx \leq \int_0^L L\Big(\int_0^L |u'(y)|^2\,dy\Big) = L^2\int_0^L |u'(y)|^2\,dy, \tag{6.1.12}$$

and hence

$$||u|| \leq L||u'||. \tag{6.1.13}$$

□

Remark 13. *The constant $c = L$ means that the Poincare inequality is valid for arbitrary bounded intervals, but not! for unbounded intervals. It is also unnecessary to have both boundary values equal zero. For instance if $v(0) \neq 0$ and, for simplicity $L = 1$, then by the same argument as above we get the following version of one-dimensional Poincare's' inequality:*

$$||u||^2_{L_2(0,1)} \leq 2\Big(v(0)^2 + ||u'||^2_{L_2(0,1)}\Big). \tag{6.1.14}$$

Theorem 24 (Stability of the homogeneous heat equation)**.** *The homogeneous INBVP for the heat equation*

$$\begin{cases} \dot{u} - u'' = 0, & 0 < x < 1, \quad t > 0 \\ u(0,t) = u_x(1,t) = 0, & \qquad\qquad\quad t > 0 \\ u(x,0) = u_0(x), & 0 < x < 1, \end{cases} \tag{6.1.15}$$

satisfies the stability estimates

$$a) \quad \frac{d}{dt}||u||^2 + 2||u'||^2 = 0, \qquad b) \quad ||u(\cdot,t)|| \le e^{-t}||u_0||.$$

Proof. a) Multiply the equation by u and integrate over $x \in (0,1)$,

$$0 = \int_0^1 (\dot{u} - u'')u\,dx = \int_0^1 \dot{u}u\,dx + \int_0^1 (u')^2\,dx - u'(1,t)u(1,t) + u'(0,t)u(0,t).$$

Using integration by parts and the boundary data we get

$$\frac{1}{2}\frac{d}{dt}\int_0^1 u^2\,dx + \int_0^1 (u')^2\,dx = \frac{d}{dt}||u||^2 + 2||u'||^2 = 0.$$

This gives the proof of a). As for b) using a) together with the Poincare inequality with $L = 1$: $||u|| \le ||u'||$ we have that

$$\frac{d}{dt}||u||^2 + 2||u||^2 \le 0. \tag{6.1.16}$$

Multiplying both sides of (6.1.16) by e^{2t} yields

$$\frac{d}{dt}\Big(||u||^2 e^{2t}\Big) \le \Big(\frac{d}{dt}||u||^2 + 2||u||^2\Big)e^{2t} \le 0. \tag{6.1.17}$$

We replace t by s and integrate over $s \in (0,t)$ to obtain

$$\int_0^t \frac{d}{ds}\Big(||u||^2 e^{2s}\Big)\,ds = ||u(\cdot,t)||^2 e^{2t} - ||u(\cdot,0)||^2 \le 0. \tag{6.1.18}$$

This yields

$$||u(\cdot,t)||^2 \le e^{-2t}||u_0||^2 \Longrightarrow ||u(\cdot,t)|| \le e^{-t}||u_0||, \tag{6.1.19}$$

and completes the proof. □

Remark 14. *For the sake of generality and application of this technical argument in higher dimensions we shall use a general notation for the domain and its boundary: namely Ω and $\partial\Omega$ respectively. The reader may replace Ω by any interval (a,b), for instance $I=(0,1)$ and $\partial\Omega$ by the corresponding boundary. The proof of the general theorem for the energy estimate in higher dimensions is given in part II.*

Theorem 25 (An energy estimate)**.** *For any small $\varepsilon>0$ We have that*

$$\int_\varepsilon^t \|\dot{u}\|(s)ds \le \frac{1}{2}\sqrt{\ln\frac{t}{\varepsilon}}\|u_0\|. \tag{6.1.20}$$

Proof. Multiply the differential equation: $\dot{u}-u''=0$, by $-tu''$ and integrate over Ω to obtain

$$-t\int_\Omega \dot{u}u''\,dx + t\int_\Omega -(u'')^2\,dx = 0. \tag{6.1.21}$$

Integrating by parts and using the fact that $u=0$ on $\partial\Omega$ we get

$$\int_\Omega \dot{u}u''\,dx = -\int_\Omega \dot{u}'\cdot u'\,dx = -\frac{1}{2}\frac{d}{dt}\|u'\|^2, \tag{6.1.22}$$

so that (11.1.11) can be written as

$$t\frac{1}{2}\frac{d}{dt}\|u'\|^2 + t\|u''\|^2 = 0, \tag{6.1.23}$$

and by using the obvious relation $t\frac{d}{dt}\|u'\|^2 = \frac{d}{dt}(t\|u'\|^2) - \|u'\|^2$ we get

$$\frac{d}{dt}(t\|u'\|^2) + 2t\|u''\|^2 = \|u'\|^2. \tag{6.1.24}$$

We now change t to s and integrate over $s\in(0,t)$ to 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. Consequently

$$t\|u'\|^2(t) + 2\int_0^t s\|u''\|^2(s)\,ds \le \frac{1}{2}\|u_0\|^2. \tag{6.1.25}$$

In particular, we have:

$$(I) \quad \|u'\|(t) \le \frac{1}{\sqrt{2t}}\|u_0\| \qquad (II) \quad \Big(\int_0^t s\|u''\|^2(s)\,ds\Big)^{1/2} \le \frac{1}{2}\|u_0\| \quad (6.1.26)$$

Analogously we can show that

$$\|u''\|(t) \le \frac{1}{\sqrt{2}\;t}\|u_0\| \tag{6.1.27}$$

Now using the differential equation $\dot{u} = u''$ and integrating (6.1.27) we obtain

$$\int_\varepsilon^t \|\dot{u}\|(s)ds \le \frac{1}{\sqrt{2}}\|u_0\| \int_\varepsilon^t \frac{1}{s}\,ds = \frac{1}{\sqrt{2}} \ln\frac{t}{\varepsilon}\|u_0\| \tag{6.1.28}$$

or more carefully

$$\begin{aligned}\int_\varepsilon^t \|\dot{u}\|(s)ds &= \int_\varepsilon^t \|u''\|(s)ds = \int_\varepsilon^t 1\cdot\|u''\|(s)ds = \int_0^t \varepsilon\frac{1}{\sqrt{s}}\cdot\sqrt{s}\|u''\|(s)ds\\ &\le \Big(\int_\varepsilon^t s^{-1}\,ds\Big)^{1/2}\cdot\Big(\int_\varepsilon^t s\|u''\|^2(s)\,ds\Big)^{1/2} \le \frac{1}{2}\sqrt{\ln\frac{t}{\varepsilon}}\|u_0\|,\end{aligned}$$

where in the first inequality is just an application of the *Cauchy Schwartz* inequality and the second is an application of (6.1.26) (II) and we have obtained the desired result. □

Problem 55. *Prove* (6.1.27). *Hint: Multiply (1) 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 $t^2\,\dot{u}$.*

6.2 FEM for the heat equation

Consider the one-dimensional heat equation with Dirichlet boundary condition

$$\begin{cases} \dot{u} - u'' = f, & 0 < x < 1, \quad t > 0,\\ u(0,t) = u(1,t) = 0, & \qquad\qquad\quad t > 0,\\ u(x,0) = u_0(x), & 0 < x < 1. \end{cases} \tag{6.2.1}$$

The *Variational formulation* for the problem (6.2.1) reads as follows: For every time interval $I_n = (t_{n-1}, t_n]$ find $u(x,t),\ t \in I_n$ such that

$$\int_{I_n}\int_0^1 (\dot{u}v + u'v')dxdt = \int_{I_n}\int_0^1 fvdxdt, \quad \forall v:\ v(0,t)=v(1,t)=0. \quad \text{(VF)}$$

A piecewise linear Galerkin approximation: For each time interval $I_n = (t_{n-1}, t_n]$, with $t_n - t_{n-1} = k$, let

$$U(x,t) = U_{n-1}(x)\Psi_{n-1}(t) + U_n(x)\Psi_n(t), \tag{6.2.2}$$

where

$$\Psi_n(t) = \frac{t-t_{n-1}}{k}, \qquad \Psi_{n-1}(t) = \frac{t_n - t}{k}, \qquad k = t_n - t_{n-1}, \tag{6.2.3}$$

and

$$U_n(x) = U_{n,1}\varphi_1(x) + U_{n,2}\varphi_2(x) + \ldots + U_{n,m}\varphi_m(x), \tag{6.2.4}$$

with $\varphi(x_j) = \delta_{ij}$ being the usual finite element basis corresponding to a partition of $\Omega = (0,1)$, with $0 = x_1 < \cdots < x_k < x_{k+1} < \cdots < x_m = 1$. In other words U is piecewise linear in both space and time variables and the unknowns are the coefficients $U_{n,k}$ satisfying the following discrete variational formulation:

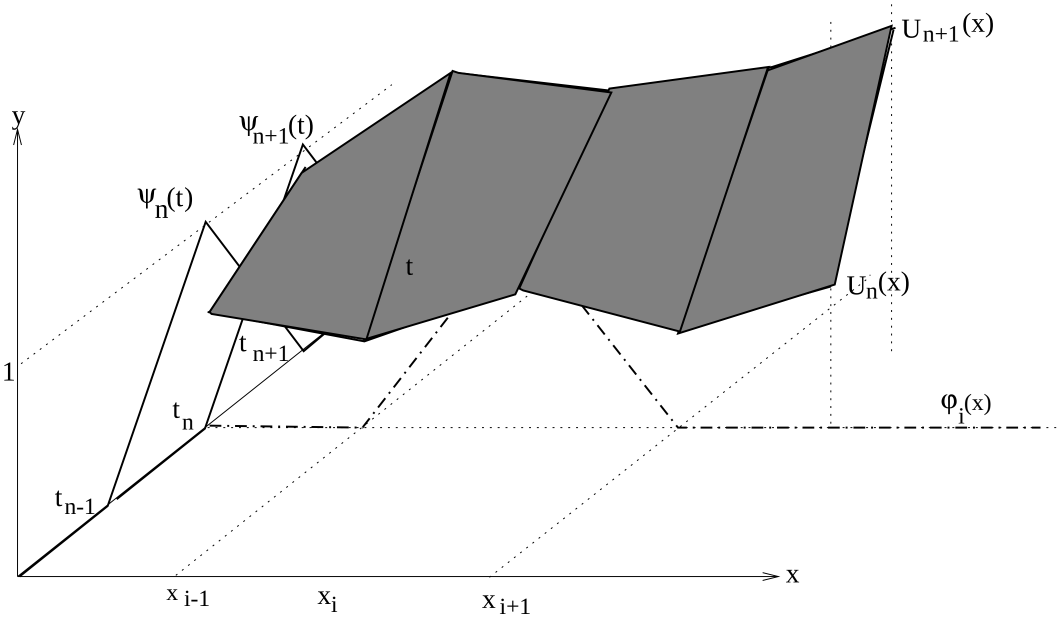
$$\int_{I_n}\int_0^1 (\dot{U}\varphi_j + U'\varphi_j')\,dxdt = \int_{I_n}\int_0^1 f\varphi_j\,dxdt, \qquad j = 1,2,\ldots,m \tag{6.2.5}$$

Note on $I_n = (t_{n-1}, t_n]$ and with $U_n := U(x_n)$ and $U_{n-1} := U(x_{n-1})$ we have

$$\dot{U}(x,t) = U_{n-1}(x)\dot{\Psi}_{n-1}(t) + U_n(x)\dot{\Psi}_n(t) = \frac{U_n - U_{n-1}}{k}. \tag{6.2.6}$$

Further differentiating (6.2.2) with respect to x we get

$$U'(x,t) = U_{n-1}'(x)\Psi_{n-1}(t) + U_n'(x)\Psi_n(t). \tag{6.2.7}$$



Inserting (6.2.6) and (6.2.7) in (6.2.5) we get using $\int_{I_n} dt = k$ and $\int_{I_n} \Psi_n dt = \int_{I_n} \Psi_{n-1} dt = \frac{k}{2}$ that

$$\begin{aligned}\underbrace{\int_0^1 U_n\varphi_j dx}_{M\cdot U_n} - \underbrace{\int_0^1 U_{n-1}\varphi_j\, dx}_{M\cdot U_{n-1}} + \underbrace{\int_{I_n} \Psi_{n-1}\, dt}_{\frac{k}{2}} \underbrace{\int_0^1 U'_{n-1}\varphi'_j\, dx}_{S\cdot U_{n-1}}& \\ + \underbrace{\int_{I_n} \Psi_n\, dt}_{\frac{k}{2}} \underbrace{\int_0^1 U'_n\varphi'_j\, dx}_{S\cdot U_n} = \underbrace{\int_{I_n}\int_0^1 f\varphi_j\, dxdt}_{F}&\end{aligned} \tag{6.2.8}$$

which can be written in a compact form as the *Crank- Nicholson system* (CNS)

$$\Big(M + \frac{k}{2}S\Big)U_n = \Big(M - \frac{k}{2}S\Big)U_{n-1} + F, \tag{CNS}$$

with the solution U_n given by

$$U_n = \underbrace{\Big(M + \frac{k}{2}S\Big)^{-1}}_{B^{-1}} \underbrace{\Big(M - \frac{k}{2}S\Big)}_{A} U_{n-1} + \underbrace{\Big(M + \frac{k}{2}S\Big)^{-1}}_{B^{-1}} F, \tag{6.2.9}$$

where

$$U_n = \begin{bmatrix} U_{n,1} \\ U_{n,2} \\ \dots \\ U_{n,m} \end{bmatrix}. \tag{6.2.10}$$

Thus with a given source term f we can determine the source vector F and then, for each $n = 1, 2, \dots N$, given the vector U_{n-1} we use the CNS to compute the m-dimensional vector U_n (m nodal values of U at the time level t_n).

Example 26. *Derive a corresponding equation system, as above, for the dG(0).*

The matrices S and M introduced in (6.2.8) are known as the *stiffness matrix* and *Mass matrix* respectively. Below we compute these matrices. Note that differentiating (6.2.4):

$$U_n(x) = U_{n,1}\varphi_1(x) + U_{n,2}\varphi_2(x) + \ldots + U_{n,m}\varphi_m(x),$$

we get

$$U_n'(x) = U_{n,1}\varphi_1'(x) + U_{n,2}\varphi_2'(x) + \ldots + U_{n,m}\varphi_m'(x). \tag{6.2.11}$$

Thus for $j = 1, \dots, m$ we have

$$SU_n = \int_0^1 U_n'\varphi_j' = \Big(\int_0^1 \varphi_j'\varphi_1'\Big)U_{n,1} + \Big(\int_0^1 \varphi_j'\varphi_2'\Big)U_{n,2} + \ldots + \Big(\int_0^1 \varphi_j'\varphi_m'\Big)U_{n,m},$$

which can be written in the matrix form as

$$SU_n = \begin{bmatrix} \int_0^1 \varphi_1'\varphi_1' & \int_0^1 \varphi_1'\varphi_2' & \dots & \int_0^1 \varphi_1'\varphi_m' \\ \int_0^1 \varphi_2'\varphi_1' & \int_0^1 \varphi_2'\varphi_2' & \dots & \int_0^1 \varphi_2'\varphi_m' \\ \dots & \dots & \dots & \dots \\ \int_0^1 \varphi_m'\varphi_1' & \int_0^1 \varphi_m'\varphi_2' & \dots & \int_0^1 \varphi_m'\varphi_m' \end{bmatrix} \begin{bmatrix} U_{n,1} \\ U_{n,2} \\ \dots \\ U_{n,m} \end{bmatrix}. \tag{6.2.12}$$

Note that S is just the matrix A_{unif} that we have already computed in Chapter 1:

$$S = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & -1 & 2 & -1 \\ 0 & \dots & \dots & \dots & -1 & 2 \end{bmatrix}. \tag{6.2.13}$$

Similarly, recalling the definition for the mass matrix M introduced in (6.2.8), we have that for $j = 1, \dots, m$

$$MU_n = \int_0^1 U_n \varphi_j. \tag{6.2.14}$$

Thus, to compute the mass matrix M one should drop all derivatives from the general form of the matrix for S given by (6.2.13). In other words unlike the form $SU_n = \int_0^1 U_n' \varphi_j'$, MU_n does not have any derivatives, neither in U_n nor in φ_j. Consequently

$$M = \begin{bmatrix} \int_0^1 \varphi_1\varphi_1 & \int_0^1 \varphi_1\varphi_2 & \dots & \int_0^1 \varphi_1\varphi_m \\ \int_0^1 \varphi_2\varphi_1 & \int_0^1 \varphi_2\varphi_2 & \dots & \int_0^1 \varphi_2\varphi_m \\ \dots & \dots & \dots & \dots \\ \int_0^1 \varphi_m\varphi_1 & \int_0^1 \varphi_m\varphi_2 & \dots & \int_0^1 \varphi_m\varphi_m \end{bmatrix}. \tag{6.2.15}$$

To continue we follow the same procedure as in chapter one recalling that for a uniform partition we have

$$\varphi_j(x) = \frac{1}{h}\begin{cases} x - x_{j-1} & x_{j-1} \le x \le x_j \\ x_{j+1} - x & x_j \le x \le x_{j+1} \\ 0 & x \notin [x_{j-1}, x_{j+1}]. \end{cases} \tag{6.2.16}$$

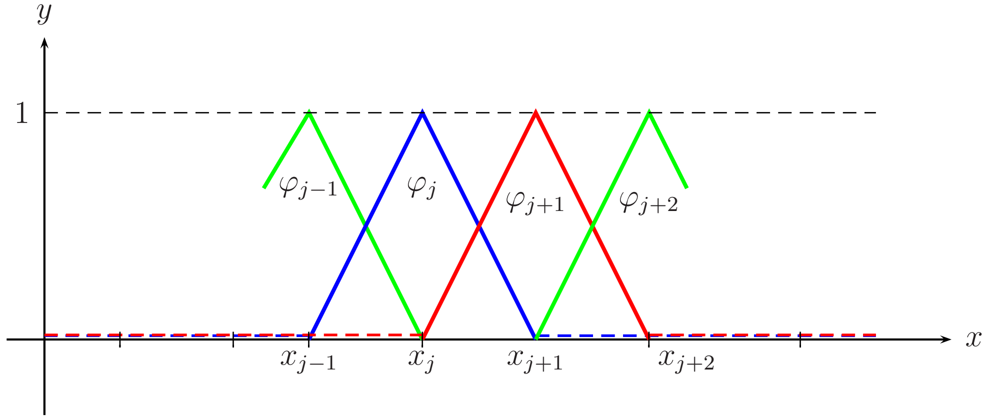


Figure 6.2: φ_j and φ_{j+1}.

Thus

$$\begin{aligned}\int_0^1 \varphi_j(x)^2\,dx &= \frac{1}{h^2}\Big(\int_{x_{j-1}}^{x_j}(x-x_{j-1})^2\,dx + \int_{x_j}^{x_{j+1}}(x_{j+1}-x)^2\Big)\\ &= \frac{1}{h^2}\Big[\frac{(x-x_{j-1})^3}{3}\Big]_{x_{j-1}}^{x_j} + \frac{1}{h^2}\Big[\frac{(x_{j+1}-x)^3}{3}\Big]_{x_j}^{x_{j+1}} \qquad (6.2.17)\\ &= \frac{1}{h^2}\cdot\frac{h^3}{3} + \frac{1}{h^2}\cdot\frac{h^3}{3} = \frac{2}{3}h,\end{aligned}$$

and

$$\begin{aligned}\int_0^1 \varphi_j\varphi_{j+1}\,dx &= \frac{1}{h^2}\int_{x_j}^{x_{j+1}}(x_{j+1}-x)(x_j-x) = [PI]\\ &= \frac{1}{h^2}\Big[(x_{j+1}-x)\frac{(x-x_j)^2}{2}\Big]_{x_j}^{x_{j+1}} - \frac{1}{h^2}\int_{x_j}^{x_{j+1}} -\frac{(x-x_j)^2}{2}\,dx\\ &= \frac{1}{h^2}\Big[\frac{(x-x_j)^3}{6}\Big]_{x_j}^{x_{j+1}} = \frac{1}{6}h.\end{aligned}$$

Obviously we have that

$$\int_0^1 \varphi_j\varphi_i\,dx = 0, \qquad \forall |i-j| > 1. \qquad (6.2.18)$$

Thus the mass matrix in this case is given by

$$M = h \begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 & \dots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & \dots & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \dots & \dots & \dots & \frac{1}{6} & \frac{2}{3} \end{bmatrix}. \tag{6.2.19}$$

6.3 Error analysis

In this section we shall consider a general domain Ω with the boundary $\partial\Omega$. Therefore the analysis are adequate in higher dimensions as well. For our specific one dimensional case this means a general interval $\Omega := [a, b]$ with $\partial\Omega = \{a, b\}$. Then a general for of (6.2.5) can be written as

$$\int_{I_n} \int_\Omega (\dot{U}v + U'v')dxdt = \int_{I_n} \int_\Omega fvdxdt \quad \text{for all } v \in V_h, \tag{6.3.1}$$

where $V_h = \{v(x) : v \text{ is continuous, piecewise linear, and } v(a) = v(b) = 0\}$, and in higher dimensional case $v(a) = v(b) = 0$ is replaced by $v|_{\partial\Omega} = 0$. Note that this *variational formulation* is valid for the exact solution u and for all $v(x, t)$ such that $v(a, t) = v(b, t) = 0$:

$$\int_{I_n} \int_\Omega (\dot{u}v + u'v')dxdt = \int_{I_n} \int_\Omega fv\, dxdt, \quad \forall v \in V_h, \tag{6.3.2}$$

Subtracting (6.3.1) from (6.3.2) we obtain the *Galerkin orthogonality* relation for the error

$$\int_{I_n} \int_\Omega (\dot{e}v + e'v')\, dxdt = 0, \qquad \text{for all } v \in V_h. \tag{6.3.3}$$

Theorem 26 (A posterirori error estimates)**.** *We have the following a posteriori error estimate for the heat conductivity equation given by* (6.2.1)

$$\|e(t)\| \leq \Big(2 + \sqrt{\ln \frac{T}{\varepsilon}}\Big) \max_{[0,T]} \|(k + h^2)r(U)\|. \tag{6.3.4}$$

Sketch. To derive error estimates we let $\varphi(x,t)$ be the solution of the following *dual problem*

$$\begin{cases} -\dot{\varphi} - \varphi'' = 0, & \text{in } \Omega \quad t < T, \\ \varphi = 0, & \text{on } \partial\Omega \quad t < T, \\ \varphi = e, & \text{in } \Omega \quad \text{for } t = T, \end{cases} \tag{6.3.5}$$

where $e = e(t) = e(\cdot, T) = u(\cdot, T) - U(\cdot, T),\ T = t_N$. Note that for $w(x,t) = \varphi(x, T-t),\ (t > 0)$ we can write the backward dual problem (6.3.5) as the following *forward problem*

$$\begin{cases} \dot{w} - w'' = 0, & \text{in } \Omega \quad t > 0, \\ w = 0, & \text{on } \partial\Omega \quad t > 0, \\ w = e, & \text{in } \Omega \quad \text{for } t = 0. \end{cases} \tag{6.3.6}$$

For this problem we have shown in the energy estimate theorem that

$$\int_\varepsilon^T \|\dot{w}\| \leq \frac{1}{2}\sqrt{\ln\frac{T}{\varepsilon}}\ \|e\|, \tag{6.3.7}$$

and consequently (let $s = T - t$, then $\varepsilon \xrightarrow{t} T \Leftrightarrow T - \varepsilon \xrightarrow{s} 0$, and $ds = -dt$) we have for φ:

$$\int_0^{T-\varepsilon} \|\dot{\varphi}\| \leq \frac{1}{2}\sqrt{\ln\frac{T}{\varepsilon}}\ \|e\|. \tag{6.3.8}$$

Now since $-\varphi'' = \dot{\varphi}$ we get also

$$\int_0^{T-\varepsilon} \|\varphi''\| \leq \frac{1}{2}\sqrt{\ln\frac{T}{\varepsilon}}\|e\| \tag{6.3.9}$$

To continue we assume that $u_0 \in V_h$ then, since $(-\dot{\varphi} - \varphi'') = 0$, we can write

$$\begin{aligned}
\|e(T)\|^2 &= \int_\Omega e(T)\cdot e(T)\,dx + \int_0^T \int_\Omega e(-\dot{\varphi} - \varphi'')\;dxdt = [\text{PI in } t] \\
&= \int_\Omega e(T)\cdot e(T)\,dx - \int_\Omega e(T)\cdot e(T)\,dx + \int_\Omega \underbrace{e(0)\cdot\varphi(0)}_{=0}\,dx \\
&+ \int_0^T \int_\Omega (\dot{e}\varphi + e'\varphi')\,dxdt = \{\text{Galerkin Orthogonality (7.1)}\} \\
&= \int_0^T \int_\Omega \dot{e}(\varphi - v) + e'(\varphi - v)'\;dxdt = \{\text{PI in } x\text{, in 2ed term}\} \\
&= \int_0^T \int_\Omega (\dot{e} - e'')(\varphi - v)\;dxdt + \int_0^T e'\,\underbrace{(\varphi - v)\Big|_{\partial\Omega}}_{=0}\,dt \\
&= \int_0^T \int_\Omega (f - \dot{U} + U'')(\varphi - v)\;dxdt = \int_0^T \int_\Omega r(U)(\varphi - v)\;dxdt,
\end{aligned}$$

where we use $\dot{e} = \dot{u} - \dot{U}$ and $e'' = u'' - U''$ to write $\dot{e} - e'' = \dot{u} - u'' - \dot{U} - U'' = f - \dot{U} - U' := r(U)$ which is the residual. Now with mesh variables $h = h(x,t)$ and $k = k(t)$ in x and t, respectively we can derive an interpolation estimate of the form:

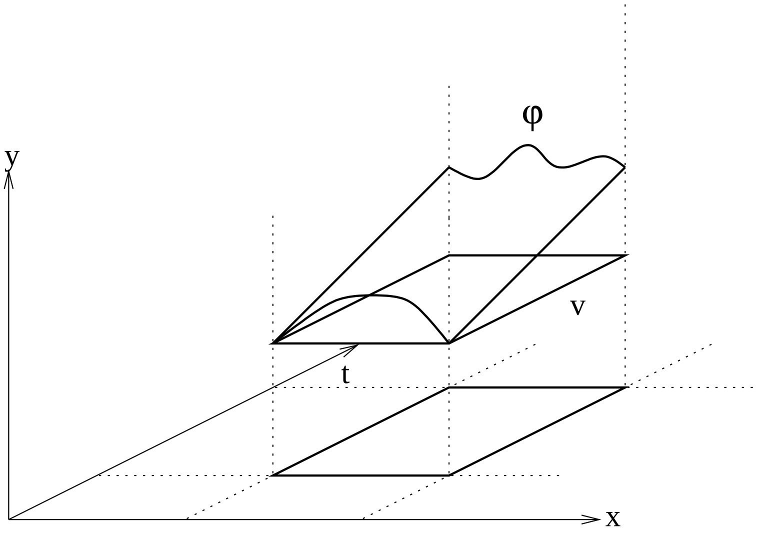
$$||\varphi - v||_{L_2} \leq k||\dot{\varphi}|_{L_2} + h^2||\varphi''||_{L_2} \leq (k + h^2)|\dot{\varphi}|_{L_2} + (k + h^2)||\varphi''||_{L_2}, \quad (6.3.10)$$

Summing up we have using maximum principle and the estimates (7.2.2)-(7.2.3), basically that

$$\begin{aligned}
\|e(T)\|^2 &\leq \int_0^T \|(k + h^2)r(U)\|(\|\dot{\varphi}\| + \|\varphi''\|) \\
&\leq \max_{[0,T]} \|(k + h^2)r(U)\| \Big[\int_0^{T-\varepsilon} (\|\dot{\varphi}\| + \|\varphi''\|) + 2\max_{[T-\varepsilon,T]} \|\varphi\|\Big] \\
&\leq \max_{[0,T]} \|(k + h^2)r(U)\| \Big(\sqrt{\ln\frac{T}{\varepsilon}}\;\|e\| + 2\|e\|\Big).
\end{aligned}$$

This gives our final estimate

$$\|e(t)\| \leq \Big(2 + \sqrt{\ln\frac{T}{\varepsilon}}\Big)\max_{[0,T]} \|(k + h^2)r(U)\|. \qquad (6.3.11)$$



The complete proof is given in general form and for higher dimensions in part II. □

• **Algorithm** Starting from the a posteriori estimate of the error $e = u - U$ for example for

$$\begin{cases} -u'' = f, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{6.3.12}$$

i.e.

$$\|e'\| \leq C\|h\, r(U)\|, \tag{6.3.13}$$

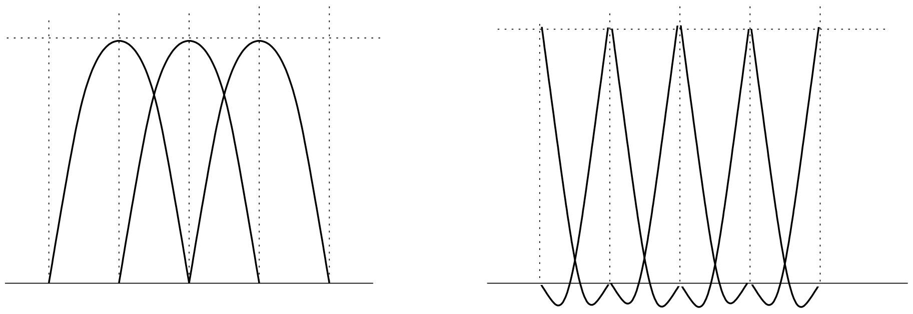
where $r(U) = |f| + \max_{I_k} |[u']|$ and [ ] denotes the jump (over the endpoints of a partition interval I_k), we have the following Algorithm:

(1) Choose an arbitrary $h = h(x)$ and a tolerance Tol > 0.

(2) Given h, compute the corresponding U.

(3) If $C\|hr(U)\| \leq$ Tol, accept U. Otherwise choose a new (refined) $h = h(x)$ and return to step (2) above. □

• **Higher order elements** $cG(2)$, piecewise polynomials of degree 2 is determined by the values of the approximate solution at the end-points of the subintervals. The constructing is through the bases functions of the form:



• Error estimates (a simple case)
For $-u'' = f, \quad 0 < x < 1$ associated with Dirichlet (or Neumann) boundary condition we have

$$\|(u - U)'\| \leq C\|h^2 D^3 u\|. \tag{6.3.14}$$

$$\|u - U\| \leq C \max\Big(h\|h^2 D^3 u\|\Big). \tag{6.3.15}$$

$$\|u - U\| \leq C\|h^2 r(U)\|, \quad \text{where} \quad |r(U)| \leq Ch. \tag{6.3.16}$$

These estimates can be extended to, for example, the space-time discretization of the heat equation.

• The equation of an elastic beam

$$\begin{cases} (au'')'' = f, & \Omega = (0,1) \\ u(0) = 0, & u'(0) = 0 \qquad \text{(Dirichlet)} \\ u''(1) = 0, & (au'')'(1) = 0, \quad \text{(Neumann)} \end{cases} \tag{6.3.17}$$

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

$$\int_0^1 au''v''dx = \int_0^1 fvdx, \qquad \forall v, \quad \text{such that} \quad (0) = v'(0) = 0. \tag{6.3.18}$$

Here, the piecewise linear finite element functions won't work (inadequate).

6.4 Exercises

Problem 56. *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 57. *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 58. *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 59. *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 60. *Let $\|\cdot\|$ denote the $L_2(0,1)$-norm. Consider the following heat equation*

$$\begin{cases} \dot{u} - u'' = 0, & 0 < x < 1, \qquad t > 0, \\ u(0,t) = u_x(1,t) = 0, & \qquad\qquad\qquad t > 0, \\ u(x,0) = u_0(x), & 0 < x < 1. \end{cases}$$

a) Show that the norms: $||u(\cdot,t)||$ and $||u_x(\cdot,t)||$ are non-increasing in time. $||u|| = \Big(\int_0^1 u(x)^2\,dx \Big)^{1/2}$.

b) Show that $||u_x(\cdot,t)|| \to 0$, as $t \to \infty$.

c) Give a physical interpretation for a) and b).

Problem 61. *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 62. *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 $I = (0,1)$.

Problem 63. *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 64. *(a) Prove an a priori error estimate for the cG(1) approximation of the boundary value problem*

$$-u'' + cu' + u = f \quad in\ \ I = (0,1), \qquad u(0) = u(1) = 0,$$

where $c \ge 0$ is constant.

(b) For which value of c is the a priori error estimate optimal?

Problem 65. *We modify problem 2 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

$$\sqrt{\varepsilon}||u'|| \leq C_1||f||, \quad ||cu'|| \leq C_2||f||, \qquad \textit{and} \quad \varepsilon||u''|| \leq C_3||f||,$$

where $||\cdot||$ is the $L_2(I)$-norm.

Problem 66. *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.

Chapter 7

The wave equation in 1d

We start with the homogeneous wave equation: Consider the initial-boundary value problem

$$\begin{cases} \ddot{u} - u'' = 0, & 0 < x < 1 \qquad t > 0 \qquad (DE) \\ u(0,t) = 0, & u(1,t) = 0 \qquad t > 0 \qquad (BC) \\ u(x,0) = u_0(x), & \dot{u}(x,0) = v_0(x), \quad 0 < x < 1 \quad (IC). \end{cases} \tag{7.0.1}$$

Below we shall derive the most important property of the wave equation

Theorem 27 (Conservation of energy)**.** *For the wave equation* (7.0.1) *we have that*

$$\frac{1}{2}||\dot{u}||^2 + \frac{1}{2}||u'||^2 = \frac{1}{2}||v_0||^2 + \frac{1}{2}||u_0'||^2 = \mathit{Constant}, \tag{7.0.2}$$

where

$$||w||^2 = ||w(\cdot, x)||^2 = \int_0^1 |w(x,t)|^2 \, dx. \tag{7.0.3}$$

Proof. We multiply the equation by $\dot{u}$ and integrate over $I = (0,1)$ to get

$$\int_0^1 \ddot{u}\,\dot{u} dx - \int_0^1 u'' u\,\dot{u}\, dx = 0. \tag{7.0.4}$$

Using partial integration and the boundary data we obtain

$$\begin{aligned}
&\int_0^1 \frac{1}{2}\frac{d}{dt}\Big(\dot{u}\Big)^2\,dx + \int_0^1 u'\,(\dot{u})'\,dx - \Big[u'(x,t)\dot{u}(x,t)\Big]_0^1 \\
&= \int_0^1 \frac{1}{2}\frac{d}{dt}\Big(\dot{u}\Big)^2\,dx + \int_0^1 \frac{1}{2}\frac{d}{dt}\Big(u'\Big)^2\,dx \qquad (7.0.5)\\
&= \frac{1}{2}\frac{d}{dt}\Big(||\dot{u}||^2 + \frac{1}{2}||u'||^2\Big) = 0.
\end{aligned}$$

Thus, the quantity

$$\frac{1}{2}||\dot{u}||^2 + \frac{1}{2}||u'||^2 = \text{Constant, independent of} \quad t. \tag{7.0.6}$$

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 67. *Show that $\|(\dot{u})'\|^2 + \|u''\|^2 =$ constant, independent of t.*
Hint: Multiply (DE): $\ddot{u} - u'' = 0$ by $-(\dot{u})''$ and integrate over I.
Alternatively: differentiate the equation with respect to x and multiply by $\dot{u}, \ldots$.

Problem 68. *Derive a total conservation of energy relation using the Robin type boundary condition: $\dfrac{\partial u}{\partial n} + u = 0$.*

7.1 FEM for the wave equation

We seek the finite element solution $u(x,t)$ for the following problem

$$\begin{cases}
\ddot{u} - u'' = 0, & 0 < x < 1 \qquad t > 0 \qquad (DE)\\
u(0,t) = 0, & u'(1,t) = g(t,) \quad t > 0 \qquad (BC)\\
u(x,0) = u_0(x), & \dot{u}(x,0) = v_0(x), \quad 0 < x < 1 \quad (IC).
\end{cases} \tag{7.1.1}$$

We let $\dot{u} = v$, and reformulate the problem as a system of PDEs:

$$\begin{cases}
\dot{u} - v = 0 & \text{(Convection)}\\
\dot{v} - u'' = 0 & \text{(Diffusion)}.
\end{cases} \tag{7.1.2}$$

Remark 15. *We can rewrite the above system as $\dot{w} + Aw = 0$ with*

$$w = \begin{pmatrix} u \\ v \end{pmatrix} \Longrightarrow \dot{w} + Aw = \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} + \underbrace{\begin{pmatrix} a & b \\ c & d \end{pmatrix}}_{A} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad (7.1.3)$$

thus we get the following system of equations

$$\begin{cases} au + bv = -\dot{u} \\ cu + dv = -\dot{v}. \end{cases} \quad (7.1.4)$$

Recalling that $\dot{u} = v$ and $\dot{v} = u''$ (7.1.4) can be written as

$$\begin{cases} au + bv = -v \\ cu + dv = -u''. \end{cases} \quad (7.1.5)$$

Consequently we have $a = 0, b = -1$ and $c = -\frac{\partial^2}{\partial x^2}$, $d = 0$, i.e.

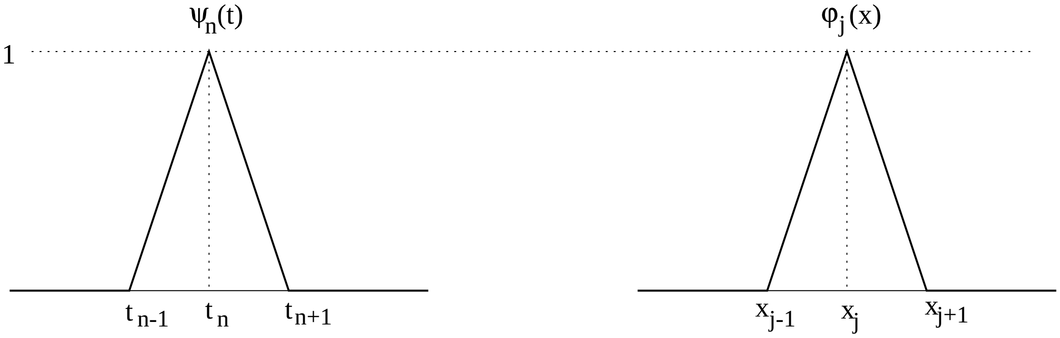
$$\underbrace{\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix}}_{\dot{w}} + \underbrace{\begin{pmatrix} 0 & -1 \\ -\frac{\partial^2}{\partial x^2} & 0 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} u \\ v \end{pmatrix}}_{w} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (7.1.6)$$

•The finite element discretization procedure
For each n we define the piecewise linear approximations as

$$\begin{cases} U(x,t) = U_{n-1}(x)\Psi_{n-1}(t) + U_n(x)\Psi_n(t), \\ V(x,t) = V_{n-1}(x)\Psi_{n-1}(t) + V_n(x)\Psi_n(t), \end{cases} \quad 0 < x < 1, \quad t \in I_n, \quad (7.1.7)$$

where

$$\begin{cases} U_n(x) = U_{n,1}(x)\varphi_1(x) + \ldots + U_{n,m}(x)\varphi_m(x), \\ V_{n-1}(x) = V_{n-1,1}(x)\varphi_1(x) + \ldots + V_{n-1,m}(x)\varphi_m(x). \end{cases} \quad (7.1.8)$$



Note that since $\dot{u} - v = 0,\ t \in I_n = (t_{n-1}, t_n]$ we have

$$\int_{I_n}\int_0^1 \dot{u}\varphi\, dxdt - \int_{I_n}\int_0^1 v\varphi\, dxdt = 0, \qquad \text{for all} \ \ \varphi(x,t). \tag{7.1.9}$$

Similarly $\dot{v} - u'' = 0$ yields

$$\int_{I_n}\int_0^1 \dot{v}\varphi\, dxdt - \int_{I_n}\int_0^1 u''\varphi\, dxdt = 0, \tag{7.1.10}$$

where, in the second term we use partial integration in x and the boundary condition $u'(1,t) = g(t)$ to obtain

$$\int_0^1 u''\varphi dx = [u'\varphi]_0^1 - \int_0^1 u'\varphi'\, dx = g(t)\varphi(1,t) - u'(0,t)\varphi(0,t) - \int_0^1 u'\varphi'\, dx.$$

Inserting in (7.1.10) we get

$$\int_{I_n}\int_0^1 \dot{v}\varphi\, dxdt + \int_{I_n}\int_0^1 u'\varphi'\, dxdt = \int_{I_n} g(t)\varphi(1,t)\, dt, \tag{7.1.11}$$

for all φ such that $\varphi(0,t) = 0$. We therefore seek $U(x,t)$ and $V(x,t)$ such that

$$\begin{aligned} &\int_{I_n}\int_0^1 \underbrace{\frac{U_n(x) - U_{n-1}(x)}{k}}_{\dot{U}} \varphi_j(x)\, dxdt - \\ &\qquad - \int_{I_n}\int_0^1 \Big(V_{n-1}(x)\Psi_{n-1}(t) + V_n(x)\Psi_n(t)\Big)\varphi_j(x)\, dxdt = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{for } j = 1, 2, \ldots, m, \end{aligned} \tag{7.1.12}$$

and

$$\int_{I_n}\int_0^1 \underbrace{\frac{V_n(x)-V_{n-1}(x)}{k}}_{\dot{V}} \varphi_j(x)\,dxdt$$
$$+\int_{I_n}\int_0^1 \underbrace{\Big(U'_{n-1}(x)\Psi_{n-1}(t)+U'_n(x)\Psi_n(t)\Big)}_{U'}\varphi'_j(x)\,dxdt \tag{7.1.13}$$
$$=\int_{I_n} g(t)\varphi_j(1)\,dt, \qquad \text{for } j=1,2,\ldots,m.$$

The equations (7.1.12) and (7.1.13) is reduced to the *iterative forms*:

$$\underbrace{\int_0^1 U_n(x)\varphi_j(x)dx}_{MU_n}-\frac{k}{2}\underbrace{\int_0^1 V_n(x)\varphi_j(x)dx}_{MV_n}$$
$$=\underbrace{\int_0^1 U_{n-1}(x)\varphi_j(x)dx}_{MU_{n-1}}+\frac{k}{2}\underbrace{\int_0^1 V_{n-1}(x)\varphi_j(x)\,dx}_{MV_{n-1}}, \text{ for } j=1,2,\ldots,m,$$

and

$$\underbrace{\int_0^1 V_n(x)\varphi_j(x)dx}_{MV_n}+\frac{k}{2}\underbrace{\int_0^1 U'_n(x)\varphi'_j(x)\,dx}_{SU_n}$$
$$=\underbrace{\int_0^1 V_{n-1}(x)\varphi_j(x)\,dx}_{MV_{n-1}}-\frac{k}{2}\underbrace{\int_0^1 U'_{n-1}(x)\varphi'_j(x)\,dx}_{SU_{n-1}}+g_n, \text{ for } j=1,2,\ldots,m,$$

respectively, where as we computed earlier

$$S=\frac{1}{h}\begin{bmatrix} 2 & -1 & \ldots & 0\\ -1 & 2 & -1 & \ldots\\ \ldots & \ldots & \ldots & \ldots\\ 0 & -1 & 2 & -1\\ 0 & 0 & -1 & 2\end{bmatrix}, \quad M=h\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & \ldots & 0\\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \ldots\\ \ldots & \ldots & \ldots & \ldots\\ \ldots & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}\\ 0 & \ldots & \frac{1}{6} & \frac{2}{3}\end{bmatrix},$$

and we use the vector functions:

$$U_n = \begin{pmatrix} U_{n,1} \\ U_{n,2} \\ \dots \\ U_{n,m} \end{pmatrix}, \quad \text{and} \quad g_n = \begin{pmatrix} 0 \\ \dots \\ 0 \\ g_{n,m} \end{pmatrix} \quad \text{where} \quad g_{n,m} = \int_{I_n} g(t)\,dt.$$

In compact form the vectors U_n and V_n are determined through solving the linear system of equations:

$$\begin{cases} MU_n - \frac{k}{2}MV_n = MU_{n-1} + \frac{k}{2}MV_{n-1} \\ \frac{k}{2}SU_n + MV_n = -\frac{k}{2}SU_{n-1} + MV_{n-1} + g_n. \end{cases} \tag{7.1.14}$$

This is a system of $2m$ equations with $2m$ unknowns:

$$\underbrace{\begin{bmatrix} M & -\frac{k}{2}S \\ \frac{k}{2}S & M \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} U_n \\ V_n \end{bmatrix}}_{W} = \underbrace{\begin{bmatrix} M & \frac{k}{2}M \\ -\frac{k}{2}S & M \end{bmatrix} \begin{bmatrix} U_{n-1} \\ V_{n-1} \end{bmatrix} + \begin{bmatrix} 0 \\ g_n \end{bmatrix}}_{b}, \tag{7.1.15}$$

with $W = A \setminus b$, $U_n = W(1:m)$, $\quad V_n = W(m+1:2m)$.

7.2 Exercises

Problem 69. *Derive the corresponding linear system of equations in the case of time discretization with dG(0).*

Problem 70 (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.*

$$\|U_n'\|^2 + \|V_n\|^2 = \|U_{n-1}'\|^2 + \|V_{n-1}\|^2. \tag{7.2.1}$$

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 $U_n^t SU_n = \|U_n'\|^2$, *where*

$$U_n = \begin{pmatrix} U_{n,1} \\ U_{n,2} \\ \dots \\ U_{n,m} \end{pmatrix}, \quad \textit{and} \quad U_n = U_n(x) = U_{n,1}(x)\varphi_1(x) + \ldots + U_{n,m}(x)\varphi_m(x).$$

Problem 71. *Apply cG(1) time discretization directly to the wave equation by letting*

$$U(x,t) = U_{n-1}\Psi_{n-1}(t) + U_n(x)\Psi_n(t), \qquad t \in I_n. \tag{7.2.2}$$

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 72. *Show that the FEM with the mesh size h for the problem:*

$$\begin{cases} -u'' = 1 & 0 < x < 1 \\ u(0) = 1 & u'(1) = 0, \end{cases} \tag{7.2.3}$$

with

$$U(x) = 7\varphi_0(x) + U_1\varphi_1(x) + \ldots + U_m\varphi_m(x). \tag{7.2.4}$$

leads to the linear system of equations: $\tilde{A} \cdot \tilde{U} = \tilde{b}$, where

$$\tilde{A} = \frac{1}{h}\underset{m \times (m+1)}{\begin{bmatrix} -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \ldots \end{bmatrix}}, \qquad \tilde{U} = \underset{(m+1) \times 1}{\begin{bmatrix} 7 \\ U_1 \\ \ldots \\ U_m \end{bmatrix}}, \qquad \tilde{b} = \underset{m \times 1}{\begin{bmatrix} h \\ \ldots \\ h \\ \frac{h}{2} \end{bmatrix}}$$

which is reduced to $AU = b$, with

$$A = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}, \quad U = \begin{bmatrix} U_1 \\ U_2 \\ \dots \\ U_m \end{bmatrix}, \quad b = \begin{bmatrix} h + \frac{7}{h} \\ h \\ \dots \\ h \\ \frac{h}{2} \end{bmatrix}$$

Problem 73. *Construct a FEM for the problem*

$$\begin{cases} \ddot{u} + \dot{u} - u'' = f, & 0 < x < 1 \quad t > 0, \\ u(0,t) = 0, & u'(1,t) = 0, \quad t > 0, \\ u(x,0) = 0, & \dot{u}(x,0) = 0, \quad 0 < x < 1. \end{cases} \tag{7.2.5}$$

Problem 74. *Assume that $u = u(x)$ satisfies*

$$\int_0^1 u'v'dx = \int_0^1 fv\,dx, \quad \textit{for all } v(x) \textit{ such that } v(0) = 0. \tag{7.2.6}$$

Show that $-u'' = f$ for $0 < x < 1$ and $u'(1) = 0$.
Hint: See Lecture notes, previous chapters.

Problem 75. *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) $f = 1, \quad u_0 = 0, \quad v_0 = 0$.

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 7 (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

$$\int_\Omega (\Delta u) v dx = \int_\Omega (\nabla u \cdot \mathbf{n}) v ds - \int_\Omega \nabla u \cdot \nabla v dx. \tag{8.1.3}$$

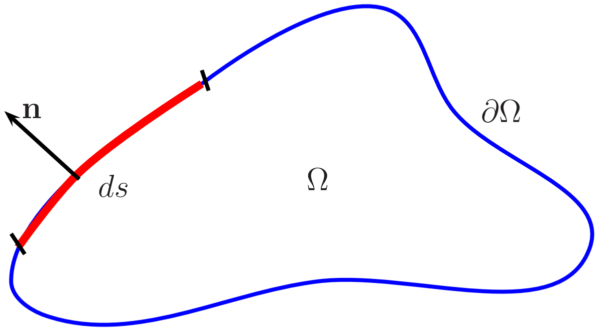


Figure 8.1: A smooth domain Ω with an outward unit normal $\mathbf{n}$

In the case that Ω is a rectangular domain. Then we have that

$$\begin{aligned}\iint_\Omega \frac{\partial^2 u}{\partial x^2} v dx dy = & \int_0^b \int_0^a \frac{\partial^2 u}{\partial x^2}(x,y) \cdot v(x,y) dx dy = [P.I.] \\ = & \int_0^b \Big(\Big[\frac{\partial u}{\partial x}(x,y) \cdot v(x,y)\Big]_{x=0}^a - \int_0^a \frac{\partial u}{\partial x}(x,y) \cdot \frac{\partial v}{\partial x}(x,y) dx \Big) dy \\ = & \int_0^b \Big(\frac{\partial u}{\partial x}(a,y) \cdot v(a,y) - \frac{\partial u}{\partial x}(0,y) \cdot v(0,y) \Big) dy - \\ & - \iint_\Omega \frac{\partial u}{\partial x} \cdot \frac{\partial v}{\partial x}(x,y) dx dy.\end{aligned}$$

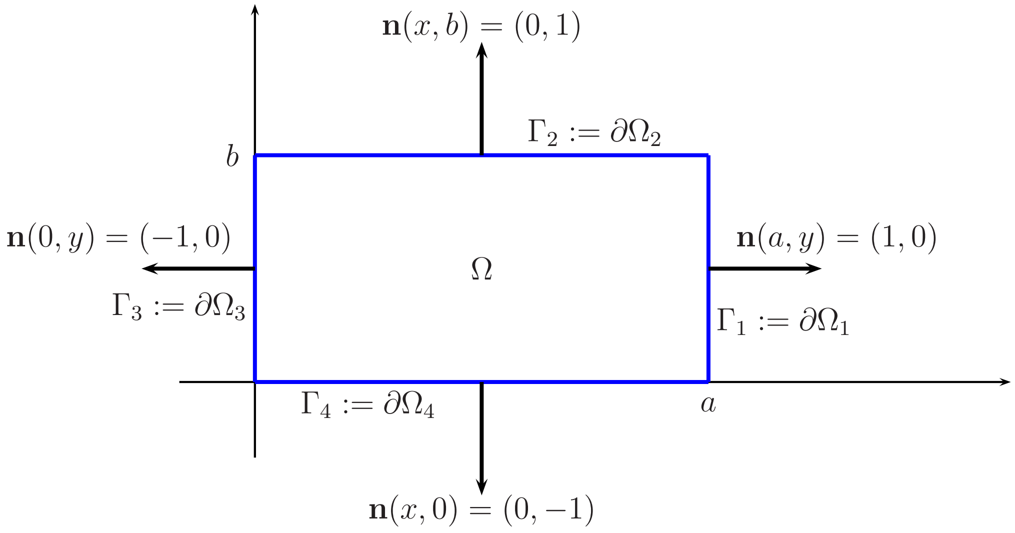

$\mathbf{n}(x,b)=(0,1)$
$\Gamma_2 := \partial\Omega_2$
b
$\mathbf{n}(0,y)=(-1,0)$
$\mathbf{n}(a,y)=(1,0)$
Ω
$\Gamma_3 := \partial\Omega_3$
$\Gamma_1 := \partial\Omega_1$
$\Gamma_4 := \partial\Omega_4$
a
$\mathbf{n}(x,0)=(0,-1)$

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

$$\iint_\Omega\frac{\partial^2 u}{\partial y^2}vdxdy=\int_{\Gamma_2\cup\Gamma_4}\Big(\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\Big)\cdot\mathbf{n}(x,y)v(x,y)ds-\iint_\Omega\frac{\partial u}{\partial y}\cdot\frac{\partial v}{\partial y}dxdy.$$

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)

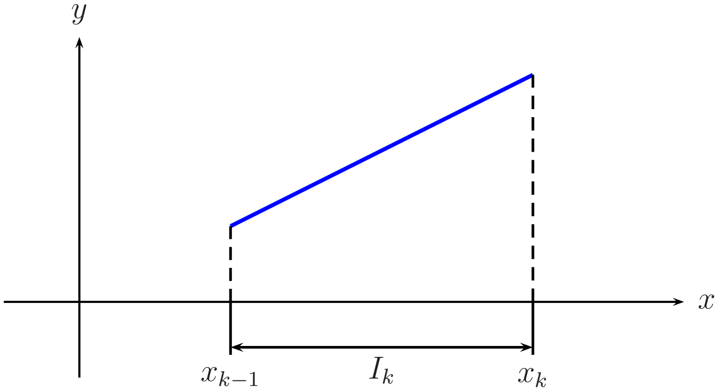
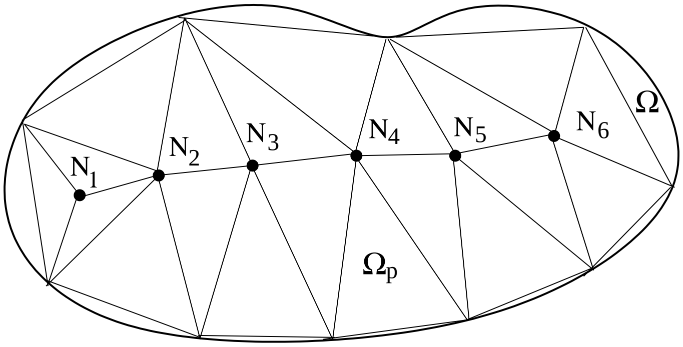


Figure 8.3: A picewise linear function on a subinterval $I_k = (x_{k-1}, x_k)$.

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 $N_i,\ 1 \le i \le 6$ 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

$$U(\mathbf{x}) = U_1\varphi_1(\mathbf{x}) + U_1\varphi_2(\mathbf{x}) + \ldots + U_6\varphi_6(\mathbf{x}), \qquad (8.2.1)$$

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

$$\varphi_i(N_j) = \left\{ \begin{array}{ll} 1, & j = i \\ 0, & j \neq i \end{array} \right\} = \delta_{ij} \quad \text{(affin)} \qquad (8.2.2)$$

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

$$U(N_5) = U_5\varphi_5(N_5) = U_5 \qquad (8.2.3)$$

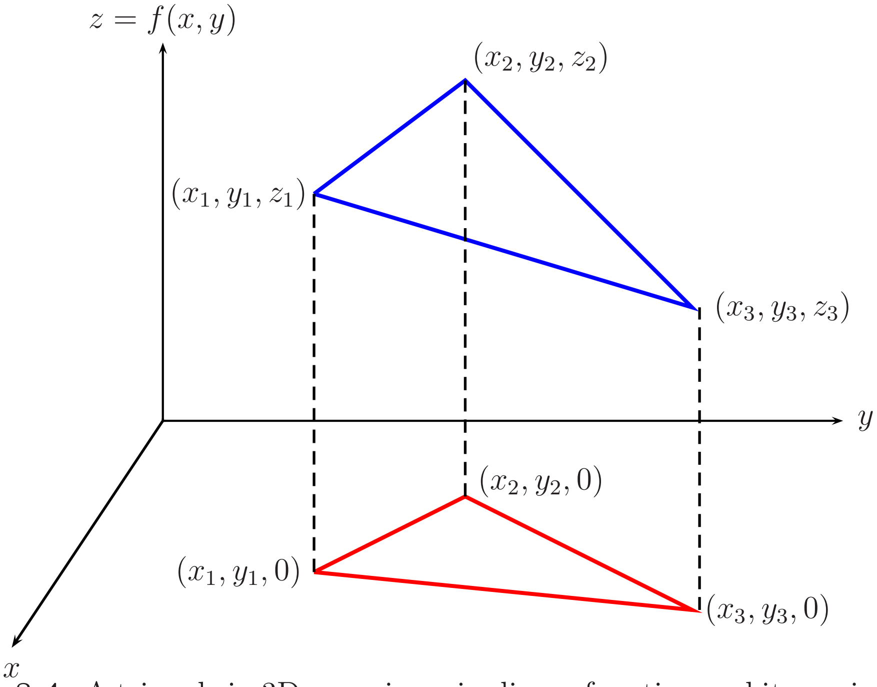
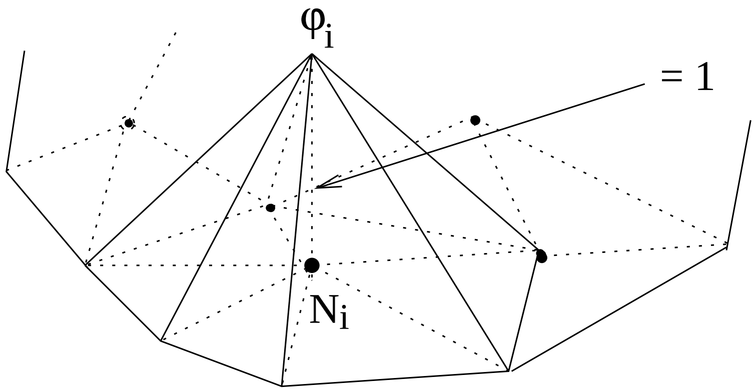


Figure 8.4: A triangle in 3D as a piecewise linear function and its projection in 2D.



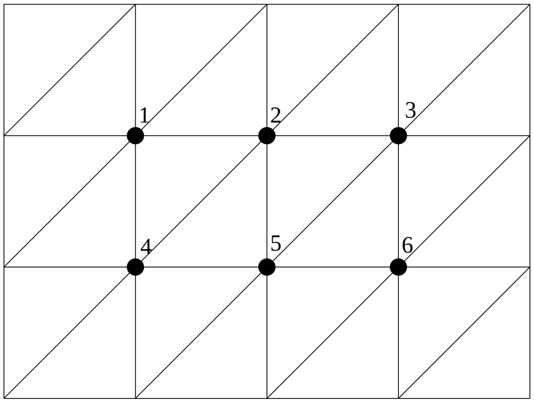
Example 27. *, let $\Omega = \{(x,y) : 0 < x < 4, 0 < y < 3\}$ and make a FEM discretization of the following boundary value problem:*

$$\begin{cases} -\Delta u = f & in\ \Omega \\ u = 0 & on\ \partial\Omega \end{cases} \tag{8.2.4}$$

The variational formulation reads as follows: Find a function u vanishing at the boundary $\Gamma = \partial\Omega$ of Ω, such that

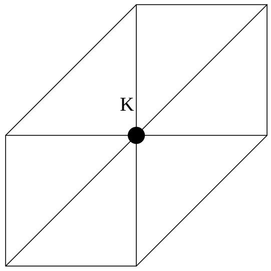
$$\iint_{\Omega} (\nabla u \cdot \nabla v) dxdy = \iint_{\Omega} f v dxdy, \qquad \forall v \in H_0^1(\Omega). \tag{8.2.5}$$

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



$$V_h^0 = \{v \in \mathcal{C}(\Omega) : v \textit{is linear on each sub-triangle and is 0 at the boundary.}\}$$

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 28. *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

$$D^2 f = \Big(\sum_{i,j=1}^{2} (\frac{\partial^2 f}{\partial x_i \partial x_j})^2 \Big)^{1/2}.$$

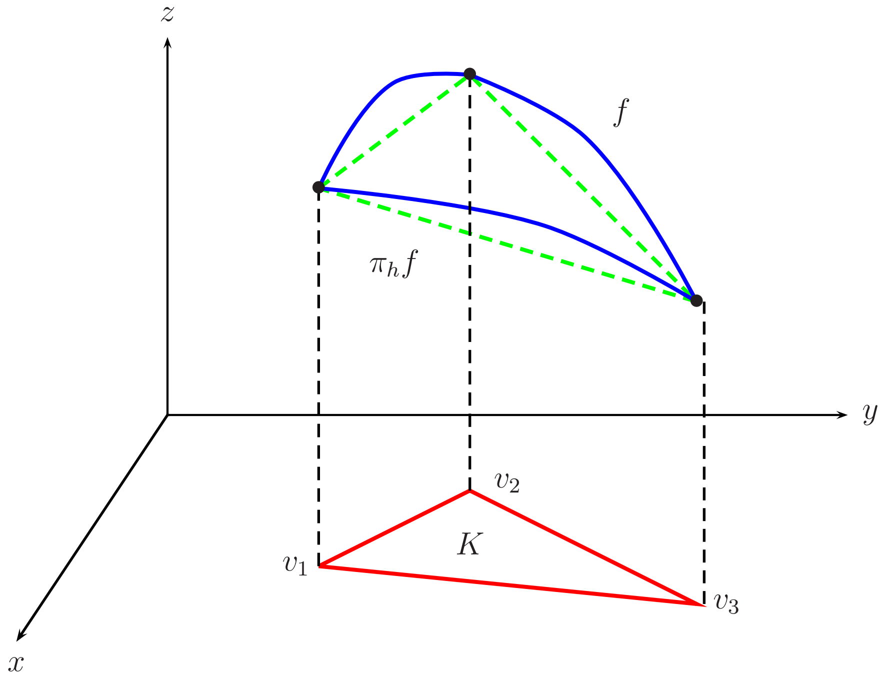


Figure 8.5: The nodal interpolant of f in 2D case

Remark 16. *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 15. *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 29.

$$\|u - \pi_h u\| \le C_i \|h^2 D^2 u\|. \tag{8.2.12}$$

Proof. We have using (8.2.10) and the Cauchy's inequality that

$$\begin{aligned} \|u - \pi_h u\|^2 &= (u - \pi_h u, u - \pi_h u) \\ &(u - \pi_h u, u - v) + (u - \pi_h u, v - \pi_h u) = (u - \pi_h u, u - v) \\ &\le \|u - \pi_h u\| \|u - v\|. \end{aligned} \tag{8.2.13}$$

This yields

$$\|u - \pi_h u\| \le \|u - v\|, \qquad \forall v \in V_h. \tag{8.2.14}$$

Now choosing $v = \pi_h u$ and recalling the interpolation theoren above we get the desired result. □

8.3 Exercises

Problem 76. *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 77. *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 78. *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 79. *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 80. *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 81. *Let* K *be a triangle with nodes* $\{a^i\}$, $i = 1, 2, 3$ *and let the midpoints of the edges be denoted* $\{a^{ij}, 1 \le i < j \le 3\}$.
a) Show that a function $v \in \mathcal{P}^1(K)$ *is uniqely determined by the degrees of freedom* $\{v(a^{ij}), 1 \le i < j \le 3\}$.
b) Are functions continuous in the corresponding finite element space of piecewise linear functions?

Problem 82. *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 83. *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 84. *Assume that the triangle* K *has nodes at* $\{v^1, v^2, v^3\}$, $v^i = (v^i_1, v^i_2)$, *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 85. *Let* K *be a triangular element. Show the following identities, for* j, $k = 1, 2$, *and* $x \in K$,

$$\sum_{i=1}^{3} \lambda_i(x) = 1, \qquad \sum_{i=1}^{3} (v^i_j - x_j)\lambda_i(x) = 0, \tag{8.3.2}$$

$$\sum_{i=1}^{3}\frac{\partial}{\partial x_k}\lambda_i(x)=0, \qquad \sum_{i=1}^{3}(v_j^i-x_j)\frac{\partial\lambda_i}{\partial x_k}=\delta_{jk}, \tag{8.3.3}$$

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 86. *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

$$|r_i(x)|\le\frac{1}{2}h_K\|D^2f\|_{L_\infty(K)}, \qquad i=1,2,3. \tag{8.3.5}$$

Hint: (I) Note that $|v^i-x|\le h_K$. (II) Start applying Cauchy's inequality to show that

$$\sum_{ij}x_ic_{ij}x_j=\sum_i x_i\sum_j c_{ij}x_j.$$

Problem 87. *τ_K is the smallest angle of a triangular element K. Show that*

$$\max_{x\in K}|\nabla\lambda_i(x)|\le\frac{2}{h_K\sin(\tau_K)}.$$

Problem 88. *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 89. *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 90. *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):

$$\text{BVP} \quad \Longleftrightarrow \quad \text{VF} \quad \Longleftrightarrow \quad \text{MP}.$$

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 28. *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

$$(u,u) \geq 0, \quad \textit{and } (u,u) = 0 \iff u \equiv 0.$$

Note that

$$(u,u) = \int_0^1 u'(x)^2 dx = 0 \iff u'(x) = 0,$$

thus $u(x)$ is constant and since $u(0) = u(1) = 0$ we have $u(x) \equiv 0$.

Definition 16. *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 29. *Let*

$$\ell(v) = \int_0^1 fv\, dx, \qquad \forall v \in \mathcal{H}_0^1, \tag{9.1.9}$$

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 17. *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:*

$$\begin{aligned} a(v,v) &\geq c_1 \|v\|_V^2, \quad \forall v \in V && (\textit{coercivity}) && (9.1.12)\\ |a(u,v)| &\leq c_2 \|u\|_V \|v\|_V, \quad \forall u, v \in V && (a \textit{ is continuous}) && (9.1.13)\\ |L(v)| &\leq c_3 \|v\|_V, \quad \forall v \in V && (L \textit{ is continuous}). && (9.1.14) \end{aligned}$$

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 18. *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 30. *For the scalar product*

$$(u,v)=\int_0^1 u'(x)v'(x)dx, \quad \textit{in} \quad \mathcal{H}_0^1, \tag{9.1.16}$$

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): $|(u,v)|\leq\|u\|\|v\|$ *is the Cauchy's inequality sketched below:*

Proof of the Cauchy's inequality. Using the obvious inequality $2ab\leq a^2+b^2$, we have

$$2|(u,w)|\leq\|u\|^2+\|w\|^2. \tag{9.1.18}$$

We let $w=(u,v)\cdot v/\|v\|^2$, then

$$2|(u,w)|=2\Big|\Big(u,(u,v)\frac{v}{\|v\|^2}\Big)\Big|\leq\|u\|^2+|(u,v)|^2\frac{\|v\|^2}{\|v\|^4} \tag{9.1.19}$$

Thus

$$2\frac{|(u,v)|^2}{\|v\|^2}\leq\|u\|^2+|(u,v)|^2\frac{\|v\|^2}{\|v\|^4}, \tag{9.1.20}$$

which multiplying by $\|v\|^2$, gives

$$2|(u,v)|^2\leq\|u\|^2\cdot\|v\|^2+|(u,v)|^2, \tag{9.1.21}$$

and hence

$$|(u,v)|^2\leq\|u\|^2\cdot\|v\|^2, \tag{9.1.22}$$

and the proof is complete. □

Definition 19. *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 20. *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 30. *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 21. *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 31. *$\mathcal{H}_0^1 = \{f : [0,1] \to \mathbb{R} : \int_0^1 f'(x)^2 dx < \infty, \wedge f(0) = f(1) = 0\}$ 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 8 (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

$$\begin{aligned} u(x) &= \int_0^x u'(y)dy \leq \int_0^x |u'(y)|dy \leq \int_0^L |u'(y)| \cdot 1dy \\ &\leq \Big(\int_0^L u'(y)^2 dy\Big)^{1/2} \Big(\int_0^L 1^2 dy\Big)^{1/2} = \sqrt{L}\Big(\int_0^L u'(y)^2 dy\Big)^{1/2}. \end{aligned} \tag{9.1.25}$$

Consequently

$$u(x)^2 \leq L \int_0^L u'(y)^2 dy, \tag{9.1.26}$$

and hence

$$\int_0^L u(x)^2dx \le L\int_0^L \Big(\int_0^L u'(y)^2dy\Big)dx = L^2\int_0^L u'(x)^2dx, \qquad (9.1.27)$$

i.e. $C_L = L$. Thus Poincare inequality deteriorates in unbounded domains. □

Definition 22. *We define a functional ℓ as a mapping from a (linear) function space V into $\mathbb{R}$, i.e.,*

$$\ell : V \to \mathbb{R}. \qquad (9.1.28)$$

- *A funcitonal ℓ is called <u>linear</u> if*

$$\begin{cases} \ell(u+v) = \ell(u) + \ell(v) & \textit{for all } \ u, v \in V \\ \ell(\alpha u) = \alpha \cdot \ell(u) & \textit{for all } \ u \in V \ \textit{and} \ \alpha \in \mathbb{R}. \end{cases} \qquad (9.1.29)$$

- *A functional is called <u>bounded</u> if there is a constant C such that*

$$|\ell(u)| \le C \cdot \|u\| \quad \textit{for all } u \in V \quad (C \textit{ is independnet of } u)$$

Example 31. *If $f \in L^2(0,1)$, i.e. $\int_0^1 f(x)^2dx$ is bounded, then*

$$\ell(v) = \int_0^1 u(x)v(x)dx \qquad (9.1.30)$$

is a bounded linear functional.

Problem 91. *Show that ℓ, defined in example above is linear.*

Problem 92. *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 32. *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:

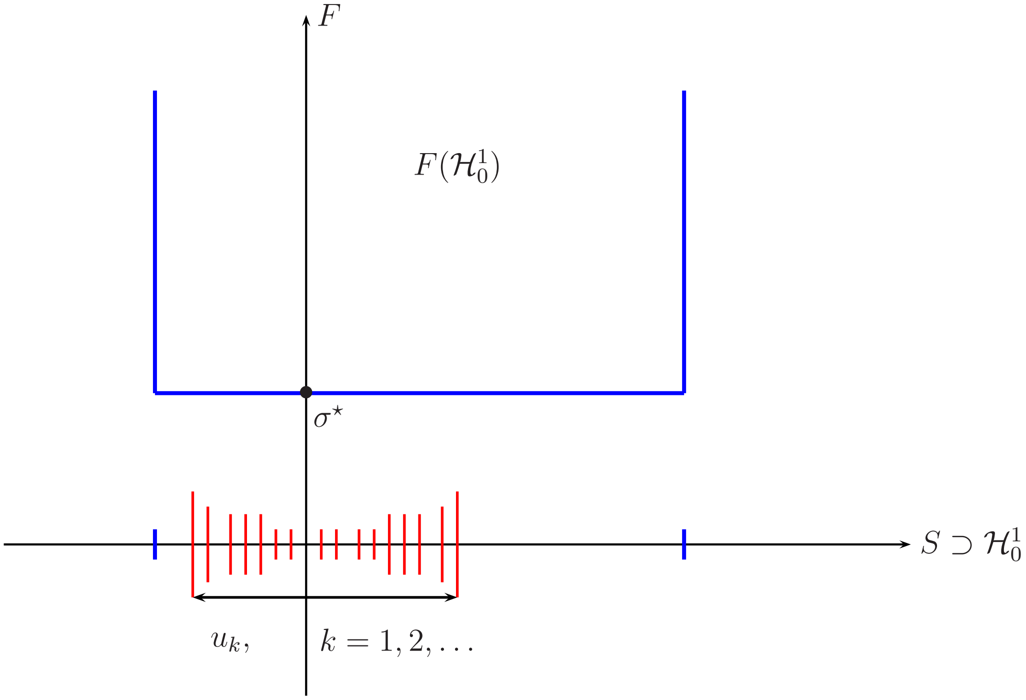


Figure 9.1: The axiom of choice for existence of a solution of (V) and (M).

(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 3. *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 33 (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 34 (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 17. 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:*

$$a(v,v) \geq \alpha\|v\|^2, \quad \textit{for some } \alpha > 0. \tag{9.2.10}$$

<u>Note</u>

$$\textit{If } a(u,v) = (u,v), \ \textit{then } \alpha = \beta = 1.$$

9.3 Exercises

Problem 93. *Verify that the assumptions of the Lax-Milgram theorem are satisfied for the following problems with appropriate assumptions on α and f.*

$$\text{(I)} \qquad \begin{cases} -u'' + \alpha u = f, & \textit{in } (0,1), \\ u(0) = u'(1) = 0, & \alpha = 0 \textit{ and } 1. \end{cases}$$

$$\text{(II)} \qquad \begin{cases} -u'' + \alpha u = f, & \textit{in } (0,1), \\ u(0) = u(1) & u'(0)0u'(1) = 0. \end{cases}$$

$$\text{(III)} \qquad \begin{cases} -u'' = f, & \textit{in } (0,1), \\ u(0) - u'(0) = u(1) + u'(1)a = 0. \end{cases}$$

Problem 94. *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 95. *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 96. *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 97. *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 9 (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 98. *Derive corresponding estimates for following Neumann problem:*

$$\begin{cases} -\Delta u + u = f, & \textit{in } \Omega \\ \frac{\partial u}{\partial n} = 0, & \textit{on } \Gamma = \partial\Omega. \end{cases} \tag{10.1.7}$$

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

$$(VF): \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega fv \, dx, \quad \forall v \ \text{ such that } v = 0 \text{ on } \Gamma. \tag{10.2.1}$$

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 $\varphi_j, j = 1, 2, \dots, n$ be the corresponding basis functions, such that $\varphi_j(x)$ is continuous, linear in x on each K and

$$\varphi_j(N_i) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \tag{10.2.2}$$

where $N_1, N_2, \dots, N_n$ are the inner nodes in the triangulation.

Now we set the approximate solution $U(x)$ to be a linear combination of the basis functions $\varphi_j,\ j = 1, \dots, n$:

$$U(x) = U_1\varphi_1(x) + U_2\varphi_2(x) + \ldots + U_n\varphi_n(x), \tag{10.2.3}$$

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

$$(FEM) \qquad \int_\Omega \nabla U \cdot \nabla \varphi_i \, dx = \int_\Omega f \cdot \varphi_i \, dx, \quad i = 1, 2, \ldots n, \tag{10.2.4}$$

or equivalently

$$(V_h^0) \qquad \int_\Omega \nabla U \cdot \nabla v \, dx = \int_\Omega f \cdot v \, dx, \quad \forall v \in V_h^0. \tag{10.2.5}$$

We recall that

$V_h^0 = \{v(x) : v \text{ is continuous, piecewise linear(on } \mathcal{T}), \text{ and } v = 0 \text{ on } \Gamma = \partial\Omega\}$.

Note that every $v \in V_h^0$ can be represented by

$$v(x) = v(N_1)\varphi_1(x) + v(N_2)\varphi_2(x) + \ldots + v(N_n)\varphi_n(x). \tag{10.2.6}$$

Theorem 35 (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*

$$\|\nabla e\| = \|\nabla(u - U) \leq C\|h\, D^2 u\|. \tag{10.2.7}$$

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

$$\|\nabla e\|^2 = \int_\Omega \nabla e \cdot \nabla u dx - \int_\Omega \nabla e \cdot \nabla v dx = \int_\Omega \nabla e \cdot \nabla (u - v) dx \le \|\nabla e\| \cdot \|\nabla (u - v)\|.$$

Thus

$$\|\nabla (u - U)\| \le \|\nabla (u - v)\|, \quad \forall v \in V_h^0, \tag{10.2.9}$$

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.

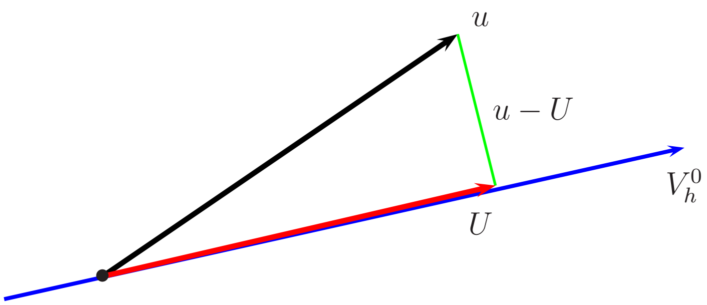


Figure 10.1: The orthogonal (L_2) projection of u on V_h^0.

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

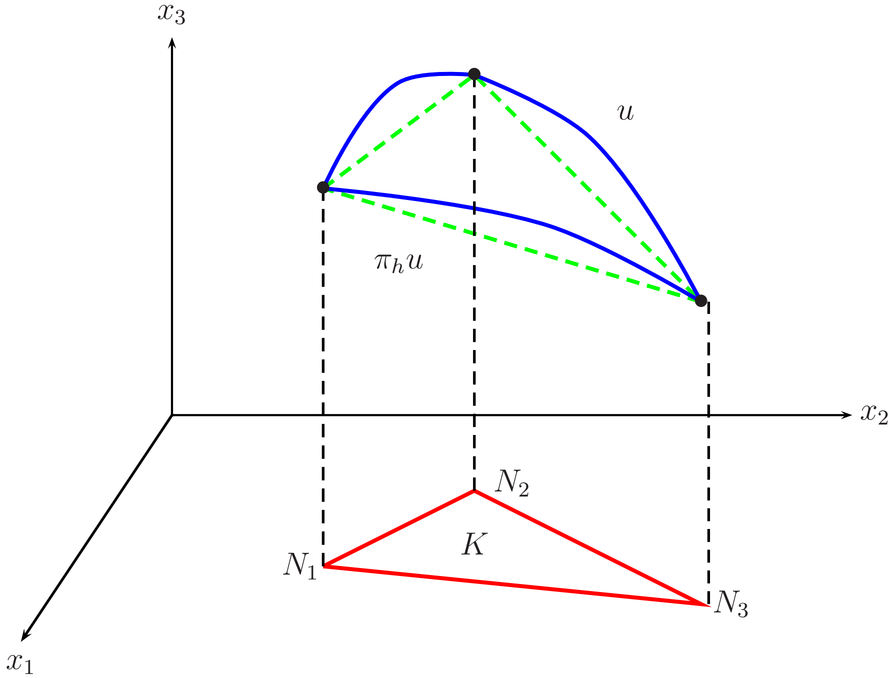


Figure 10.2: The nodal interpolant of u in 2D case

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:

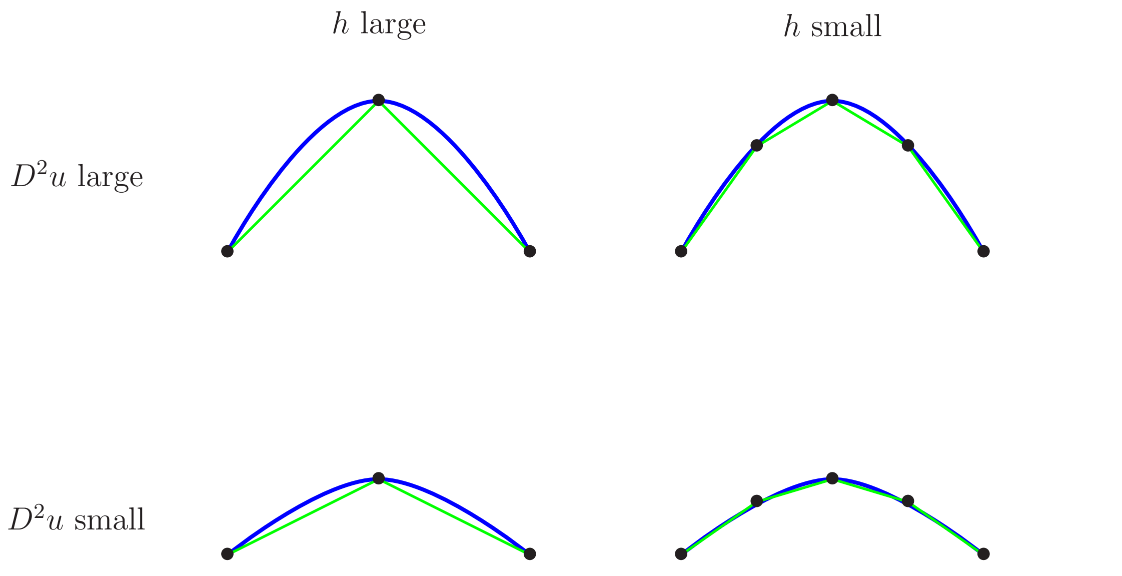


Figure 10.3: The adaptivity priciple: to refine mesh for large D^2u

Lemma 10 (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 36 (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

$$\begin{aligned}\|e\|^2 = \int_\Omega e(-\Delta\varphi)dx = \int_\Omega \nabla e\cdot\nabla\varphi\ dx, \int_\Omega \nabla e\cdot\nabla(\varphi - v)\,dx \\ \leq \|\nabla e\|\cdot\|\nabla(\varphi - v)\|, \quad \forall v\in V_h^0,\end{aligned} \tag{10.2.16}$$

where in the last equality we have used the Galerkin orthogonality. We now choose v such that

$$\|\nabla(\varphi - v)e\| \leq C\|h\cdot D^2\varphi C\| \leq C(\max_\Omega h)\|h\cdot D^2\varphi C\|. \tag{10.2.17}$$

Applying the lemma to φ, we get

$$\|D^2\varphi\| \leq C_\Omega\cdot\|\Delta\varphi\| = C_\Omega\|e\|. \tag{10.2.18}$$

Now (10.2.11)-(10.2.18) implies that

$$\begin{aligned}\|e\|^2 \leq \|\nabla e\|\cdot\|\nabla(\varphi - v)\| \leq \|\nabla e\|\cdot C\max_\Omega h\,\|D^2\varphi\| \\ \leq \|\nabla e\|\cdot C\max_\Omega hC_\Omega\|e\| \leq C^2\,C_\Omega\,\max_\Omega h\|e\|\|h\,D^2u\|.\end{aligned} \tag{10.2.19}$$

Thus we have obtained the desired result: *a priori error estimate:*

$$\|e\| = \|u - U\| \leq C^2\,C_\Omega\,(\max_\Omega h)\cdot\|h\,D^2u\|. \tag{10.2.20}$$

□

Corollary 3 (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,*

$$\|u - U\| \leq C^2\,C_\Omega^2\,(\max_\Omega h)^2\,\|f\|. \tag{10.2.21}$$

Theorem 37 (a posteriori error estimate)**.** *For the solution of the Poisson equation* (10.0.1) *we have that*

$$\|u - U\| \le C\,\|h^2 r\|, \tag{10.2.22}$$

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

$$\begin{aligned} \|e\|^2 &= \int_\Omega \nabla e \cdot \nabla\varphi\, dx - \int_\Omega \nabla e \cdot \nabla v dx = \int_\Omega \nabla e \cdot \nabla(\varphi - v)\, dx \\ &= \int_\Omega (-\Delta e)(\varphi - v)\, dx \le \|h^2 r\| \cdot \|h^{-2}(\varphi - v)\| \\ &\le C \cdot \|h^2 r\| \cdot \|\Delta\varphi\| \le C \cdot \|h^2 r\| \cdot \|e\|, \end{aligned} \tag{10.2.26}$$

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:

$$\|u - U\| \le C\,\|h^2 r\|. \tag{10.2.27}$$

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,

$$\|e\| \le C\,\|h^2 f\|. \tag{10.2.28}$$

Note that now is $\nabla e(\varphi - v) \neq 0$ on the enter-element boundaries. □

Problem 99. *Show that* $\|(u-U)'\| \leq C\|hr\|$.

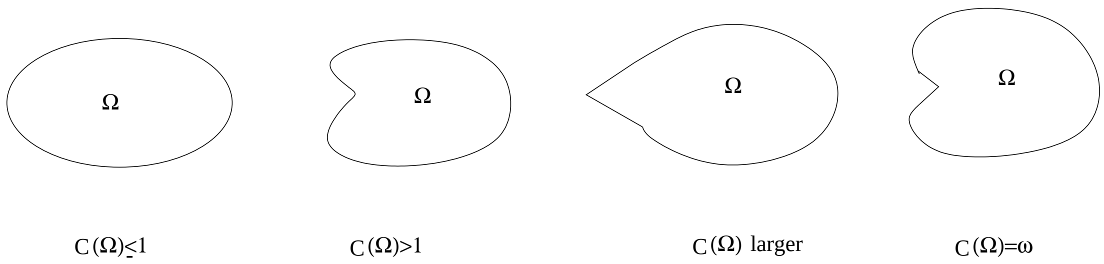
Problem 100. *Verify that for* v *being the interpolant of* φ, *we have*

$$\|e\| \leq C\,\|h^2 f\| \times \begin{cases} \|h^{-2}\,(\varphi - v)\| & \leq C\,\|\Delta\varphi\|, \quad \textit{and} \\ \|h^{-1}(\varphi - v)\| & \leq C\,\|\nabla\varphi\|. \end{cases} \tag{10.2.29}$$

Problem 101. *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

$$\int_\Omega u_x u_{xyy} dxdy = \int_{\partial\Omega} u_x(u_{yx} \cdot n_y) ds - \int_\Omega u_{xy} u_{xy}\, dxdy, \qquad (10.2.32)$$

which inserting in (10.2.31) gives that

$$\int_\Omega u_{xx} u_{yy}\, dxdy = \int_{\partial\Omega} (u_x u_{yy} n_x - u_x u_{yx} n_y) ds + \int_\Omega u_{xy} u_{xy}\, dxdy. \qquad (10.2.33)$$

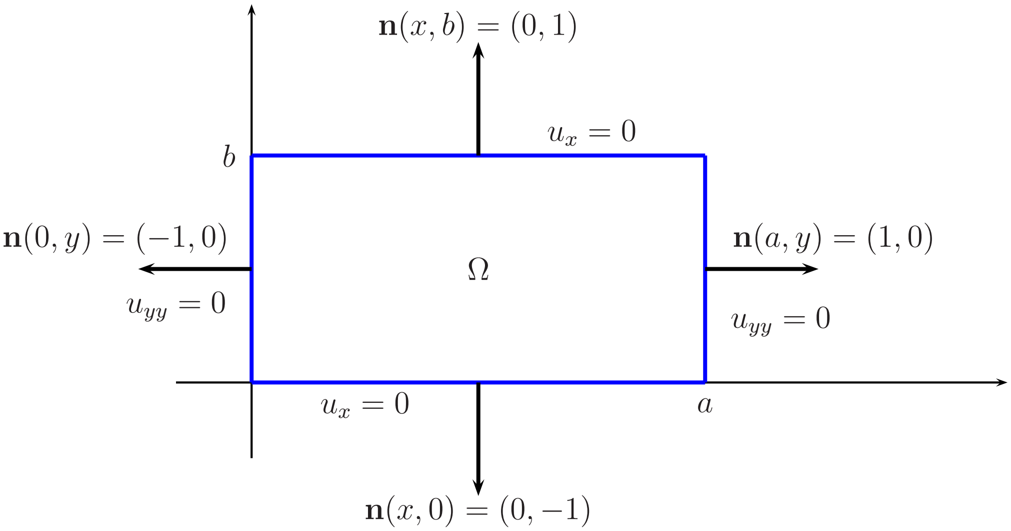

$\mathbf{n}(x, b) = (0, 1)$
$u_x = 0$
b
$\mathbf{n}(0, y) = (-1, 0)$
Ω
$\mathbf{n}(a, y) = (1, 0)$
$u_{yy} = 0$
$u_{yy} = 0$
$u_x = 0$
a
$\mathbf{n}(x, 0) = (0, -1)$

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

$$\int_\Omega u_{xx} u_{yy} dxdy = \int_\Omega u_{xy} u_{xy} dxdy = \int_\Omega u_{xy}^2\, dxdy. \qquad (10.2.34)$$

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 $\equiv 1$.

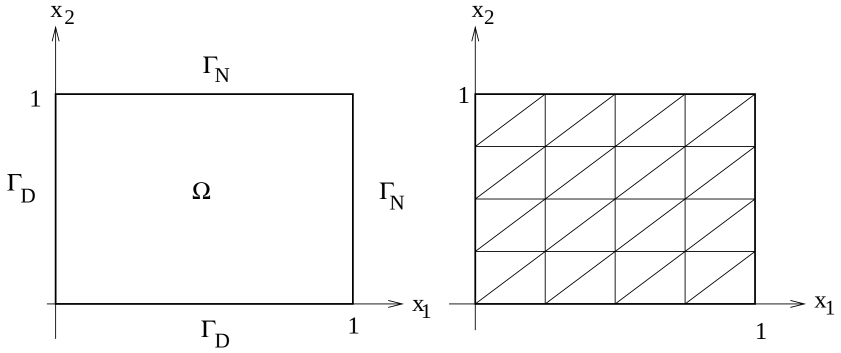
□

10.3 Exercises

Problem 102. *Consider the following two dimensional problem:*

$$\begin{cases} -\Delta u = 1, & \textit{in } \Omega \\ u = 0, & \textit{on } \Gamma_D \\ \frac{\partial u}{\partial n} = 0, & \textit{on } \Gamma_N \end{cases} \tag{10.3.1}$$

See figure below

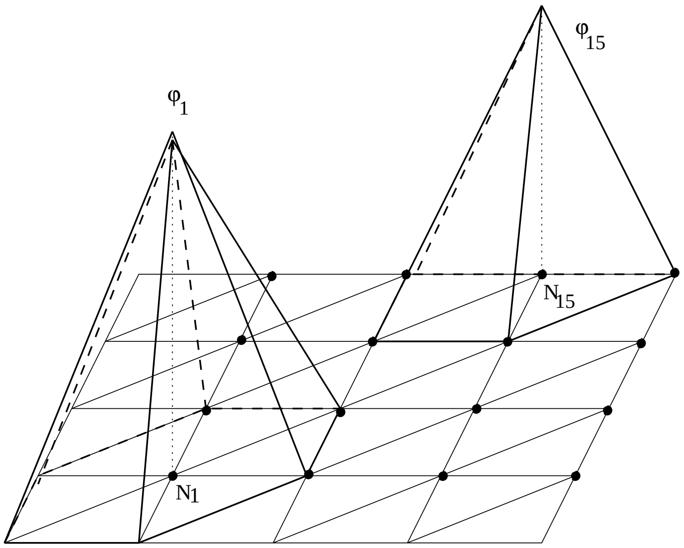


Triangulate Ω as in the figure and let

$$U(x) = U_1\varphi_1(x) + \ldots + U_{16}\varphi_{16}(x),$$

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 103. *Generalize the procedure in the previous problem to the following case*

$$\begin{cases} -\nabla(a\nabla u) = f, & in\ \Omega \\ u = 0, & on\ \Gamma_D \\ a\frac{\partial u}{\partial n} = 7, & on\ \Gamma_N \end{cases}, \quad where \quad \begin{cases} \quad a = 1 \quad for\ x_1 < \frac{1}{2} \\ \quad a = 2 \quad for\ x_1 > \frac{1}{2} \\ f = x_2. \quad mesh\text{-}size{=}h. \end{cases}$$

Problem 104. *Consider the Dirichlet problem*

$$-\nabla\cdot(a(x)\nabla u) = f(x), \quad x \in \Omega \subset \mathbb{R}^2, \qquad u = 0,\ for\ x \in \partial\Omega.$$

Assume that c_0 and c_1 are constants such that $c_0 \le a(x) \le c_1,\ \forall x \in \Omega$ 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

$$||u - U||_{H_0^1(\Omega)} \le C \inf_{\chi\in M} ||u - \chi||_{H_0^1(\Omega)}.$$

Problem 105. *Consider the following* Schrödinger *equation*

$$\dot{u} + i\Delta u = 0, \quad in\ \Omega, \qquad u = 0, \quad on\ \partial\Omega,$$

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 106. *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:*

$$\begin{cases} -\Delta u = 1, & \text{in} \quad \Omega = \{(x_1, x_2) : \ 0 < x_1 < 2, \ \ 0 < x_2 < 1\}, \\ u = 0, & \text{on} \quad \Gamma_1 = \{(0, x_2)\} \cup \{(x_1, 0)\} \cup \{(x_1, 1)\}, \\ \frac{\partial u}{\partial n} = 0, & \text{on} \quad \Gamma_2 = \{(2, x_2) : 0 \leq x_2 \leq 1\}. \end{cases}$$

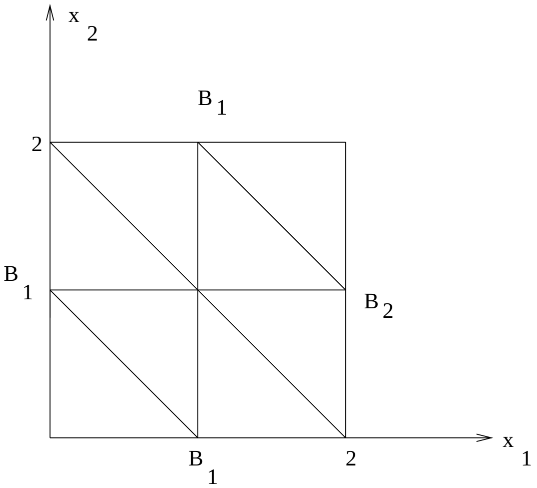
Problem 107. *Let $\Omega = (0,2) \times (0,2)$, $B_1 = \partial\Omega \setminus B_2$ and $B_2 = \{2\} \times (0,2)$. Determine the stiffness matrix and load vector in the cG(1) solution for the problem*

$$\begin{cases} -\frac{\partial^2 u}{\partial x_1^2} - 2\frac{\partial^2 u}{\partial x_2^2} = 1, & \text{in } \Omega = (0,2) \times (0,2), \\ u = 0, \quad \text{on } B_1, & \frac{\partial u}{\partial x_1} = 0, \quad \text{on } B_2, \end{cases}$$

with piecewise linear approximation applied on the triangulation below:

Problem 108. *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:*

$$\begin{cases} -\Delta u = 1, & \text{on} \quad \Omega = (0,1) \times (0,1), \\ \frac{\partial u}{\partial n} = 0, & \text{for} \quad x_1 = 1, \\ u = 0, & \text{for} \quad x \in \partial\Omega \setminus \{x_1 = 1\}, \end{cases}$$

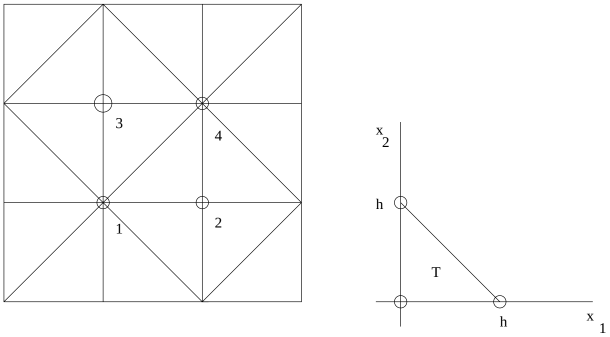


on a triangulation with triangles of side length $1/4$ *in the* x_1*-direction and* $1/2$ *in the* x_2*-direction.*

Problem 109. *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 38 (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 110. *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 $u_0 = constant$*?*

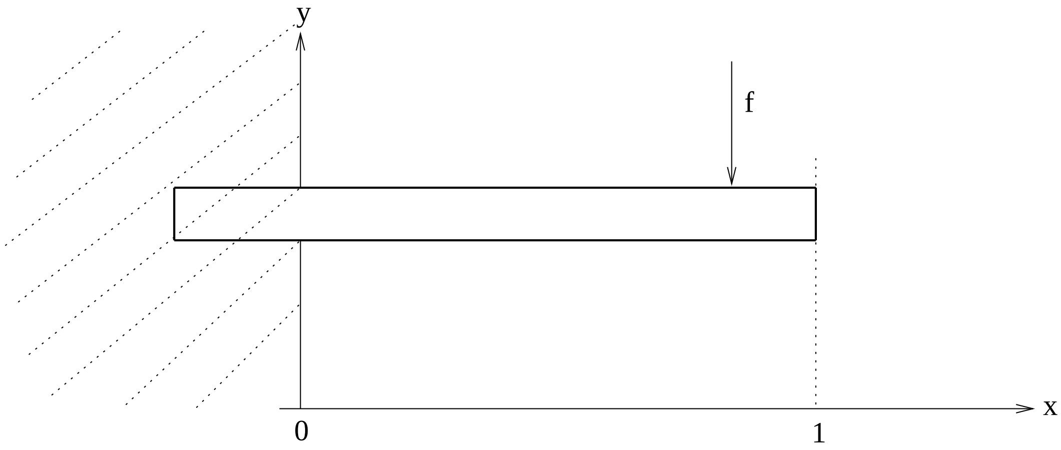
Problem 111. *Derive the corresponding estimate for Neuman boundary condition:*

$$\frac{\partial u}{\partial n} = 0. \tag{11.1.17}$$

Problem 112. *Prove the stability estimate* (11.1.5).

Example 32 (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)*

$$\begin{cases} (au'')'' = f, & \Omega = (0,1), \\ u(0) = 0, & u'(0) = 0, \qquad (\textit{Dirichlet}) \\ u''(1) = 0, & (au'')'(1) = 0, \qquad (\textit{Neumann}) \end{cases} \tag{11.1.18}$$



where a *is the bending stiffness*

au'' *is the moment*

f *is the function load*

$u = u(x)$ *is the vertical deflection*

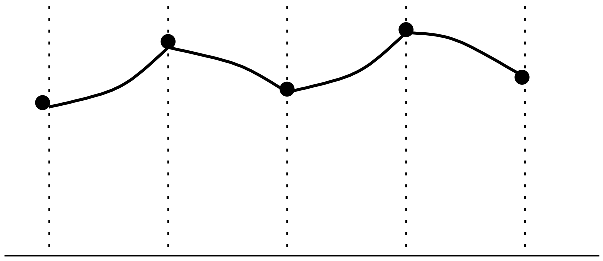
Variational form:

$$\int_0^1 au''v''dx = \int_0^1 fvdx, \quad \forall v(x) \textit{ such that } v(0) = v'(0) = 0. \tag{11.1.19}$$

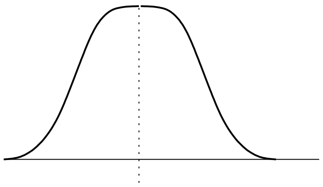
<u>*FEM:*</u> *Piecewise linear functions won't work (inadequate).*

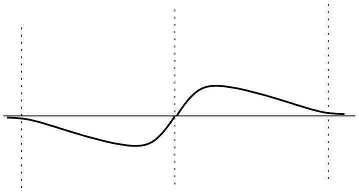
11.2 Exercises

Problem 113. *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 114. *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 115. *Formulate the equation for* $cG(1)dG(1)$ *for the two-dimensional heat equation using the discrete Laplacian.*

Problem 116. *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*

$$f = \begin{cases} 1/(\pi\varepsilon)^2, & \textit{for} \quad r < \varepsilon, \\ 0, & \textit{otherwise.} \end{cases}$$

e) *Determine the fundamental solution corresponding to* $f = \delta$, *letting* $\varepsilon \to 0$.

Problem 117. *Consider the Schrödinger equation*

$$i\dot{u} - \Delta u = 0, \quad \textit{in } \Omega, \qquad u = 0, \quad \textit{on } \partial\Omega.$$

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 118. *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 119. *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 39 (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*

$$\frac{1}{2}\|\dot{u}\|^2 + \frac{1}{2}\|\nabla u\|^2 = \textit{constant, independent of } t, \tag{12.0.2}$$

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

$$\int_\Omega \frac{1}{2}\frac{d}{dt}(\dot{u}^2)\,dx + \int_\Omega \frac{1}{2}\frac{d}{dt}(|\nabla u|^2)\,dx = 0 \iff \frac{1}{2}\frac{d}{dt}(\|\dot{u}\|^2 + \|\nabla u\|^2) = 0,$$

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 120. *Show that*

$$\|\dot{u}\|^2 + \|\nabla u\|^2 = \textit{ constant, independent of}\, t.$$

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 121. *Derive a total conservation of energy relation using the Robin type boundary condition: $\dfrac{\partial u}{\partial n} + u = 0$.*

Problem 122. *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 123. *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 124. *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

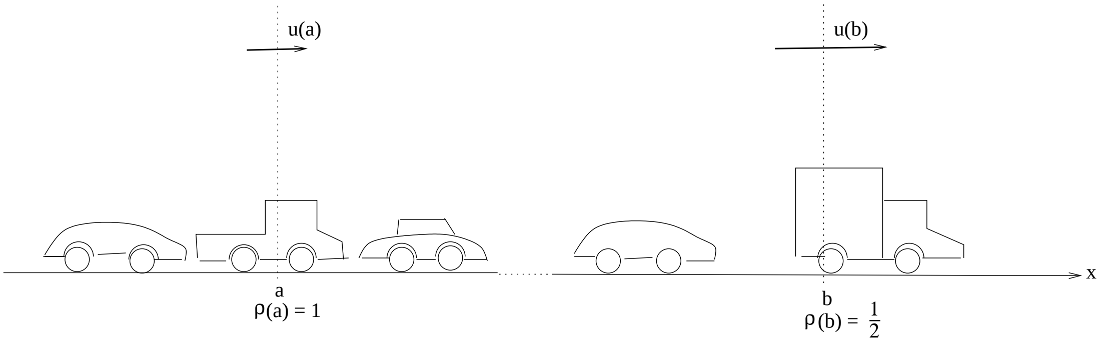
We illustrate the convection-diffusion phenomenon by an example:

Example 33 (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

$$\dot{\rho} + (1 - 2\rho)\rho' = 0 \qquad \textit{(A non-linear convection equation).} \tag{13.1.4}$$

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*

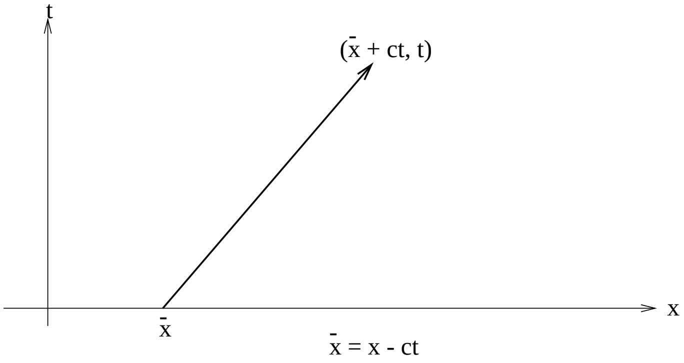
$$\dot{\rho} + \Big((c - \varepsilon\frac{\rho'}{\rho})\rho\Big)' = 0, \tag{13.1.5}$$

i.e.,

$$\dot{\rho} + c\rho' - \varepsilon\rho'' = 0 \quad (\textit{A convection - diffusion equation}). \tag{13.1.6}$$

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 18. *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.*

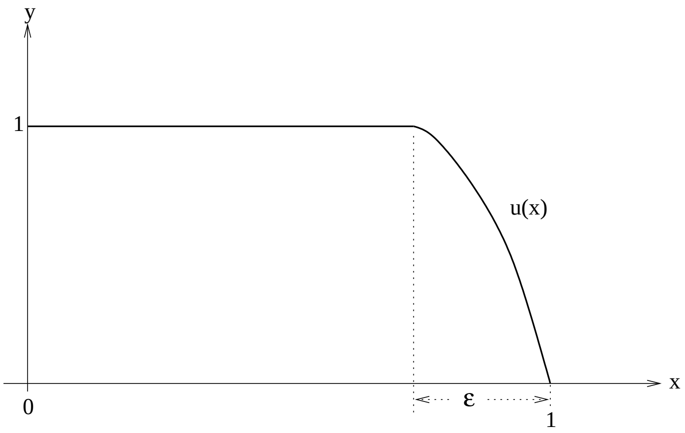
Example 34 (The boundary layer)**.** *Consider the following boundary value problem*

$$\text{(BVP)} \quad \begin{cases} u' - \varepsilon u'' = 0, & 0 < x < 1 \\ u(0) = 1, & u(1) = 0. \end{cases} \tag{13.1.10}$$

The exact solution to this problem is given by

$$u(x) = C\Big(e^{1/\varepsilon} - e^{x/\varepsilon}\Big), \qquad \textit{with} \quad C = \frac{1}{e^{1/\varepsilon} - 1}. \tag{13.1.11}$$

which has an outflow boundary layer of width $\sim \varepsilon$, as seen in the Fig. below

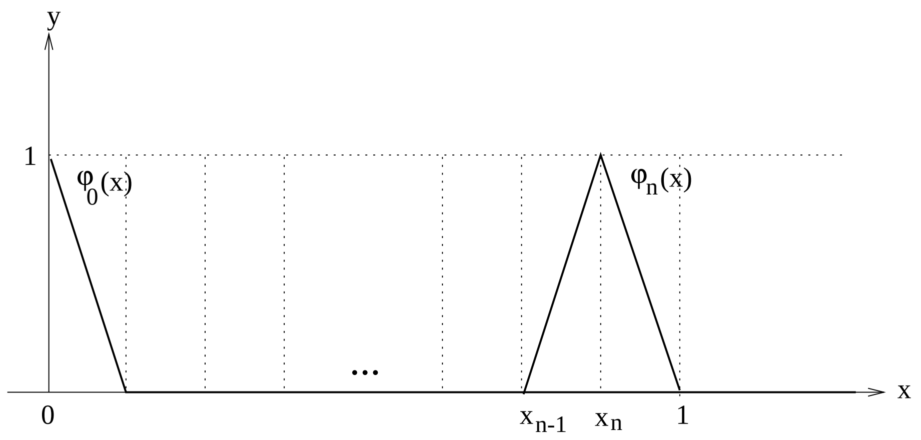


13.1.1 Finite Element Method

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

$$\int_0^1 \Big(U'\varphi_j dx + \varepsilon U'\varphi_j'\Big)\,dx = 0, \quad j = 1, 2, \dots n. \tag{13.1.13}$$

This yields the equations

$$\frac{1}{2}\Big(U_{j+1} - U_{j-1}\Big) + \frac{\varepsilon}{h}\Big(2U_j - U_{j-1} - U_{j+1}\Big) = 0, \quad j = 1, 2, \dots, n, \tag{13.1.14}$$

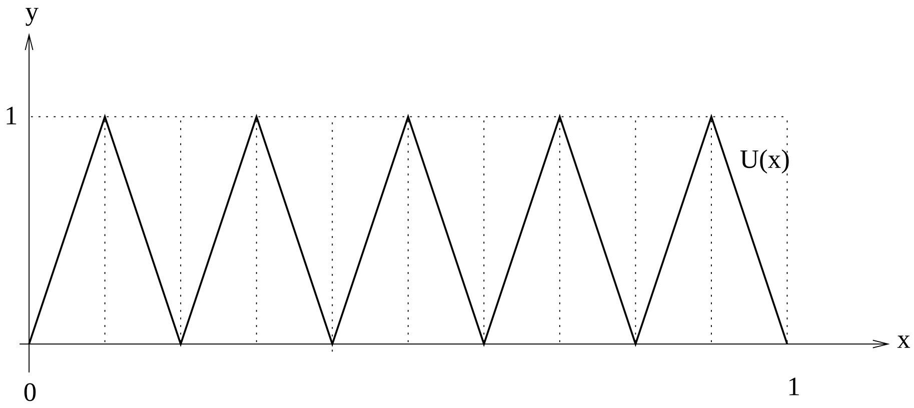
where $U_0 = 1$ and $U_{n+1} = 0$.

Note that, using *Central -differencing* we may also write

$$\underbrace{\frac{U_{j+1} - U_{j-1}}{2h}}_{\text{corresp. to } u'(x_j)} \quad -\varepsilon \underbrace{\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2}}_{\text{corresp. to } u''(x_j)} = 0 \quad \Big(\Longleftrightarrow \frac{1}{h}\times \text{ equation(13.1.14)}\Big).$$

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

$$u'(x_j) \approx \frac{U_j - U_{j-1}}{h}, \qquad (13.1.15)$$

which, formally, gives a better stability, however, with low accuracy.

Remark 19. *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 ($\beta \equiv 1$)

$$\int_0^1 \Big[u'(v+\frac{1}{2}hv'\Big) - \varepsilon \cdot u''\Big(v+\frac{1}{2}hv'\Big)\Big]dx = \int_0^1 f\Big(v+\frac{1}{2}hv'\Big)\,dx. \quad (13.1.16)$$

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 20. *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*

$$A^tAu = A^tf \quad \Longleftrightarrow \quad -u'' = -f, \qquad \textit{(the continuous form)}. \tag{13.1.20}$$

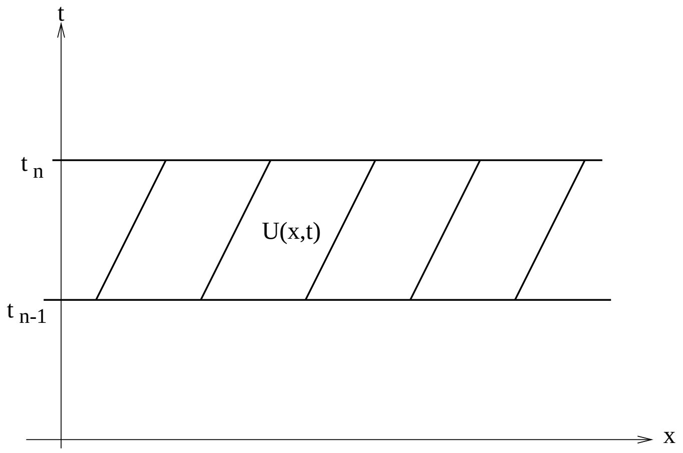
While multiplying $u' = Au = f$ *by* v *and integrating over* $(0,1)$ *we have*

$$\int_0^1 U'v'dx = \int_0^1 fv'dx \qquad \textit{(the weak form)}, \tag{13.1.21}$$

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:

$$\int_{I_n}\int_{\Omega}\Big[(\dot{U}+\beta U)\Big(v+\frac{\beta}{2}h\dot{v}\Big)+\hat{\varepsilon}\;U'v'\Big]\,dxdt=\int_{I_n}\int_{\Omega}f\Big(v+\frac{\beta}{2}hv'\Big)\,dxdt.\quad\square$$

13.2 Exercises

Problem 125. *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

$$\Big(\int_I u^2\phi\,dx\Big)^{1/2}\leq\Big(\int_I f^2\phi\,dx\Big)^{1/2}.$$

where $\phi(x)$ is a positive weight function defined on $(0,1)$ satisfying $\phi_x(x)\leq 0$ and $-\phi_x(x)\leq\phi(x)$ for $0\leq x\leq 1$.

Problem 126. *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

$$|\phi\prime(0)|\leq C\|e\|,\qquad\|\varepsilon\phi''\|\leq C\|e\|.$$

Problem 127. *Consider the convection-diffusion-absorption problem*

$$-\varepsilon u'' + xu' + u = f, \quad \textit{in } I = (0,1), \quad u(0) = u'(1) = 0,$$

where ε *is a positive constant, and* $f \in L_2(I)$. *Prove that*

$$\|\varepsilon u''\| \le \|f\|,$$

where $\|\cdot\|$ *denotes the* $L_2(I)$*-norm.*

Problem 128. *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*

$$-u'' + u' = f, \quad \textit{in } I = (0,1), \quad u(0) = u(1) = 0.$$

Problem 129. *Prove an a priori and an a posteriori error estimate for the* $cG(1)$ *finite element method for the problem*

$$-u'' + u' + u = f, \quad \textit{in } I = (0,1), \quad u(0) = u(1) = 0.$$

Problem 130. *Consider the convection-diffusion-absorption problem*

$$-\varepsilon u_{xx} + u_x + u = f, \quad \textit{in} \quad I = (0,1), \quad u(0) = 0, \qquad \sqrt{\varepsilon} u_x + u(1) = 0,$$

where ε *is a positive constant, and* $f \in L_2(I)$. *Prove the following stability estimates for the solution* u

$$\|\sqrt{\varepsilon} u_x\| + \|u\| + |u(1)| \le C\|f\|,$$

$$\|u_x\| + \|\varepsilon u_{xx}\| \le C\|f\|,$$

where $\|\cdot\|$ *denotes the* $L_2((0,))$*-norm and* C *is an appropriate constant.*

Problem 131. *Consider the convection problem*

$$\beta \cdot \nabla u + \alpha u = f, \qquad x \in \Omega, \quad u = g, quad x \in \Gamma_-, \tag{13.2.1}$$

Define the outflow Γ_+ *ans inflow* Γ_- *boundaries. Assume that* $\alpha - \frac{1}{2}\nabla \cdot \beta \ge c > 0$. *Show the following stability estimate*

$$c\|u\|^2 \int_{\Gamma_+} n \cdot \beta u^2 \, ds \, dt \le \|u_0\|^2 + \frac{1}{c}\|f\|^2 + \int_{\Gamma_-} |n \cdot \beta| g^2 \, ds. \tag{13.2.2}$$

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 132. *Consider the convection problem*

$$\begin{aligned} &\dot{u} + \beta \cdot \nabla u + \alpha u = f, && x \in \Omega, \quad t > 0, \\ &u = g, && x \in \Gamma_-, \quad t > 0, \\ &u(x,0) = u_0(x), && x \in \Omega, \end{aligned} \tag{13.2.3}$$

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*

$$\begin{aligned} \|u(\cdot,T)\|^2 + c\int_0^T \|u(\cdot,t)\|^2\,dt + \int_0^T \int_{\Gamma_+} n \cdot \beta u^2\,ds\,dt \\ \leq \|u_0\|^2 + \frac{1}{c}\int_0^T \|f(\cdot,t)\|^2\,dt + \int_0^T \int_{\Gamma_-} |n \cdot \beta| g^2\,ds\,dt, \end{aligned} \tag{13.2.4}$$

where $\|u(\cdot,T)\|^2 = \int_\Omega u(x,T)^2\,dx$.

Answers to Exercises

Piecewise Polynomial Approximation in 1D

Linear Least Squares

1.

   a. $x_1 = -7$, $x_2 = 4$

   b. $x_1 = 1.66$, $x_2 = 4.42$

   c. $x_1 = 2$, $x_2 = 1$

   d. $x_1 = 1.6$, $x_2 = 0.6$, $x_3 = 1.2$

   e. $x_1 = 1$, $x_2 = 1$, $x_3 = 3$

2.

   a. $y = 2t - 1$

   b. $y = 3t + 1$

   c. $y = 4 - t$

3. a. $y = \frac{1}{70}(25x^2 + 21x + 76)$

   b. $y = \frac{1}{20}(-5x^2 - 9x + 37)$

4. c. because r must be orthogonal against all columns of A.

5. $x_1 = 1.5942$, $x_2 = 0.0088$

6. Yes!

Galerkin's Method

7.

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. -

8.

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 $C = \dfrac{2}{3\pi(9\pi^2+1)}$.

9.

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 $\xi_1 = \frac{8}{9}(\frac{2}{9} - \frac{2}{3}\pi)/\pi$.

10. $U(x) = (16 \sin x + \frac{16}{27} \sin 3x)/\pi^3 + 2x^2/\pi^2$.

Polynomial Interpolation in 1D

12. (a) x, (b) 0.

13.

$$\Pi_1 f(x) = \begin{cases} 4 - 11(x+\pi)/(2\pi), & -\pi \le x \le -\frac{\pi}{2}, \\ 5/4 - (x + \frac{\pi}{2})/(2\pi), & -\frac{\pi}{2} \le x \le 0, \\ 1 - 7x/(2\pi), & 0 \le x \le \frac{\pi}{2}, \\ 3(x-\pi)/(2\pi), & \frac{\pi}{2} \le x \le \pi. \end{cases}$$

18. Check the conditions required for a Vector space.

19.
$$\Pi_1 f(x) = f(a)\frac{2x-a-b}{a-b} + f(\frac{a+b}{2})\frac{2(x-a)}{b-a}.$$

20. Hint: Use Theorem 5.1 from PDE Lecture Notes.

21.
$$\pi_4\Big(e^{-8x^2}\Big) \approx 0.25x^4 - 1.25x^2 + 1.$$

22. For example we may choose the following basis:

$$\varphi_{i,j}(x) = \begin{cases} 0, & x \in [x_{i-1}, x_i], \\ \lambda_{i,j}(x), & i = 1, \dots, m+1, \quad j = 0, 1, 2. \end{cases}$$

$$\lambda_{i,0}(x) = \frac{(x-\xi_i)(x-x_i)}{(x_{i-1}-\xi_i)(x_{i-1}-x_i)}, \quad \lambda_{i,1}(x) = \frac{(x-x_{i-1})(x-x_i)}{(\xi_i-x_{i-1})(\xi_i-x_i)},$$

$$\lambda_{i,2}(x) = \frac{(x-x_{i-1})(x-\xi_i)}{(x_i-x_{i-1})(x_i-\xi_i)}, \qquad \xi_i \in (x_{i-1}, x_i).$$

23. Trivial

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

Numerical linear algebra

26.
$$LU = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 2 \\ 0 & -1 & 3 \\ 0 & 0 & 5 \end{bmatrix}.$$

27.
$$x = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

28.

$$LDU = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & -4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 14 \end{bmatrix} \begin{bmatrix} 1 & 1 & -3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

30. The exact solution is $(1/15, -11/15) = (0.066666, -0.733333)$.

(a) $(u_1^3, u_2^3) = (5/64, -47/64)$, $\quad \rho(J) = 1/4$ and $||e_3||_\infty = 0.011$.

(b) $(u_1^3, u_2^3) = (0.0673828, -0.7331543)$, $\quad \rho(G) = 1/16$ and $||e_3||_\infty = 7 \times 10^{-4}$.

(c) $(u_1^3, u_2^3) = (0.066789, -0.733317)$, $\quad \rho(\omega_0) = 0.017$ and $||e_3||_\infty = 1 \times 10^{-4}$.

Two-Point BVPs

32. 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.

34. 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.

37. 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).

38. 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).

39. 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.

40. d) $||U||_E^2 = \xi^T A\xi$ (check spectral theorem, linear algebra!)

41. a) For an $M+1$ partition (here $M = 2$) we get $a_{ii} = 2/h$, $a_{i,i+1} = -1/h$ except $a_{M+1,M+1} = 1/h - 1$, $b_i = 0$, $i = 1, \dots, M$ and $b_{M+1} = -1$.

42. c)

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = 2/3 \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}.$$

43.

$$3 \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \frac{1}{18} \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\Longleftrightarrow \text{(MATLAB)} \quad \xi_1 = \xi_2 = 0.102.$$

44. Check the theory.

Scalar Initial Value Problems

47. a)

$$a_{ij} = \frac{j}{j+i} - \frac{1}{j+i+1}, \qquad b_i = \frac{1}{i+1}, \qquad i,j = 1,2,\dots,$$

b) $q = 1: \quad U(t) = 1 + 3t. \qquad q = 2: \quad U(t) = 1 + \frac{8}{11}t + \frac{10}{11}t^2.$

49. 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-Nicholson: $U_n = \frac{1}{3}U_{n-1}, \quad U_0 = 1.$

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-Nicholson: $|U_n| = |\frac{1-0.2i/2}{1+0.2i/2}||U_{n-1}| = |U_{n-1}|.$

Heat Equation in 1D

54. Heat conduction with

$u(x,t) =$	temperature at x at time t.
$u(x,0) = u_0(x),$	the initial temperature at time $t = 0$.
$u(0,t) = 0,$	fixed temperature at time $x = 0$.
$u'(1,t) = 0,$	isolated boundary at $x = 1$ (no hear flux).
$f = 20 - u,$	heat source, in this case a control system to force $u \Rightarrow 20$.

57. $||e||_E \leq C_i\Big(||hu''||_{L_2(I)} + \sqrt{K}||h^2u''||_{L_2(I)}\Big).$

58. a) $||(u-U)'||_a \leq C_i||hR(U)||_{1/a}.$

b)We have the matrix equation

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 3 & -2 \\ 0 & 0 & -2 & 4 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

which yields the approximate solution $U = -3(1/2, 1, 2, 3)^t$.

62. $||e|| \leq ||h^2 u_{xx}||$

63. $||e||_{H^1} \leq C_i\Big(||hu''|| + ||h^2 u''||\Big)$.

64. a) $||e||_E \leq (1+c)||u-v||_E$. b) $c = 0$.

Wave Equation in 1D

73. Follow the procedure as in the lecture notes.

75. 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) $u(x,t) = \frac{1}{2c}\int_0^t 2c(t-s)\,ds = t^2/2$.

Calculs in Several variables/Piecewise Polynomials

78. No! There are no other rotation invariant solutions.

79. $\rho_K \leq \frac{\tau_K h_K}{2}$.

81. b) No!

84.

$$\lambda_1(x) = 1 - D^{-1}(v^3 - v^2)^t \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} (x - v^1),$$

where $D = (v_1^2 - v_1^1)(v_2^2 - v_2^1) - (v_1^3 - v_1^1)(v_2^3 - v_2^1)$.

Riesz and Lax-Milgram Theorems

93. (I) and (II) $\alpha > 0$ and $f \in L_2(0,1)$. (III) $f \in L_2(0,1)$.

Poisson equation

105. c) $1/\sqrt{\lambda_1}$.

106.

$$A = \begin{pmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \qquad b = \frac{1}{8}\begin{pmatrix} 2 \\ 2 \\ 2 \\ 1 \end{pmatrix}.$$

107.

$$A = \begin{bmatrix} 6 & -1 \\ -1 & 3 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 1/2 \end{bmatrix}.$$

108.

$$A = \begin{pmatrix} 5 & -2 & 0 & 0 \\ -2 & 5 & -2 & 0 \\ 0 & -2 & 5 & -2 \\ 0 & 0 & -2 & 5/2 \end{pmatrix} \qquad b = \frac{1}{16}\begin{pmatrix} 2 \\ 2 \\ 2 \\ 1 \end{pmatrix}.$$

109.

$$M = \frac{h^2}{12}\begin{bmatrix} 8 & 1 & 1 & 1 \\ 1 & 4 & 0 & 1 \\ 1 & 0 & 4 & 1 \\ 1 & 1 & 1 & 8 \end{bmatrix}, \qquad S = \begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}.$$

Heat equation in ND

114. b)

$$\|w(T)\|^2 + 2\int_0^T \|\nabla w\|^2\,dt \leq \|\varepsilon\|^2.$$

Wave equation in ND

122. $v = \frac{1}{4\pi}\frac{\cos(r)}{r}$ and the corresponding solution $u = e^{it}\frac{1}{4\pi}\frac{\cos(r)}{r}$.

Convection-Diffusion Equations

128. a priori: $||e||_{H^1} \leq C_i\Big(||hu''|| + ||h^2u''||\Big).$

a posteriori: $||e||_{H^1} \leq C_i||hR(U)||.$

129. 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] K. Atkinson, *An Introduction to Numerical Analysis*, 2ed Ed. John Wiley & Sons, Inc, New York, 1989.

[4] D. Braess, *Finite Elements*. Theory, fast solvers, and application in solid mechanics, 2ed Ed. Cambridge Universityu Press, 2001.

[5] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[6] L. R. Burden and J. D. Faires, *Numerical Analysis*, Fifth Ed. Brook/Cole, CA, 1998.

[7] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, New York, 1980.

[8] P. Ciarlet, *Introduction to Matrix Numerical Analysis and Optimization* (in French), Masson, Paris, 1982.

[9] K. Eriksson, D. Estep, P. Hansbo and C. Johnson, *Computational Differ-*

ential Equations, Studentlitteratur, Lund, 1996.

[10] L. C. Evans, *Partial Differential Equations*, Graduate Studies in Mathematics, 19. Americam Mathematical Society, Providance, RI, 1998.

[11] G. B. Folland, *Fourier Analysis and its Applications*, Waswoth & Cole 1992.

[12] G. B. Folland, *Intorduction to Partial Differential Equations*, Princeton University Press, 1976.

[13] G. Golub and C. V. Loan, *Matrix Computations*, John Hopkins University Press, Maryland, 1983.

[14] K. E. Gustafson, *Partial Differential Equations and Hilbert Space Methods*, John Wiley & Sons, New York, 1980.

[15] T. J. R. Hughes, *The Finite Element Method.* Linear Static and Dynamic Finite Element Analysis, Prentice-Hall, Englewood Cliffs, New Jersy, 1987.

[16] C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Element Method*, Studentlitteratur, Lund, 1991.

[17] S. Larsson and V. Thomee, *Partial differential equations with numerical methods.* Texts in Applied Mathematics, 45. Springer-Verlag, Berlin, 2003.

[18] J. T. Oden, *Finite elements: an introduction.* Handbook of numerical analysis, Vol. II, 3–15, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991.

[19] G. Strang, *Introduction to Applied Mathematics*, Wellesely-Cambridge Press, Cambridge, Mass, 1986.

[20] 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.

[21] W. Strauss, *Partial Differential Equations.* An Introduction, 2ed Ed.

John Wiley & Sons, Ltd, 2008.

[21] M. E. Taylor, *Partial Differential Equations. I. Basic Theory*. Applied Mathematical Sciences, 115. Springer-Verlag, New York, 1996.

[23] V. Thomee, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Mathematics 1054. Springer-Verlag, New York, Tokyo, 1984.

[24] S. Yakowitz and F. Szidarovsky, *An Introduction to Numerical Computations*, 2ed Ed. Macmillan Co. New York, 1989.

[25] O. C. Zienkiewicz, *The Finite Element Method in Structural and Continum Mechanics*. McGraw-Hill, London, 1971.