

# FINITE ELEMENT METHODS FOR THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

Vassilios A. Dougalis

Department of Mathematics, University of Athens, Greece

and

Institute of Applied and Computational Mathematics, FORTH, Greece

Revised edition 2013

## PREFACE

This is the current version of notes that I have used for the past thirty-five years in graduate courses at the University of Tennessee, Knoxville, the University of Crete, the National Technical University of Athens, and the University of Athens. They have evolved, and aged, over the years but I hope they may still prove useful to students interested in learning the basic theory of Galerkin - finite element methods and some facts about Sobolev spaces.

I first heard about approximation with cubic Hermite functions and splines from George Fix in the Numerical Analysis graduate course at Harvard in the fall of 1971, and also, subsequently, from Garrett Birkhoff of course. But most of the basic techniques of the analysis of Galerkin methods I learnt from courses and seminars that Garth Baker taught at Harvard during the period 1973-75.

Over the years I was fortunate to be associated with and learn more about Galerkin methods from Max Gunzburger, Ohannes Karakashian, Larry Bales, Bill McKinney, George Akrivis, Vidar Thomée and my students; for my debt to the latter it is apt to say that *διδάσκω ἀεὶ διδασκόμενος*.

I would like to thank very much Dimitri Mitsoudis and Gregory Kounadis for transforming the manuscript into  $\text{\TeX}$ .

V. A. Dougalis

Athens, March 2012

In the revised 2013 edition, two new chapters 6 and 7 on Galerkin finite element methods for parabolic and second-order hyperbolic equations were added. These had previously existed in hand-written form, and I would like to thank Gregory Kounadis and Leetha Saridakis for writing them in  $\text{\TeX}$ .

V. A. Dougalis

Athens, February 2013

# Contents

<b>1</b>	<b>Some Elements of Hilbert Space Theory</b>	<b>1</b>
1.1	Vector spaces . . . . .	1
1.2	Inner product, Norm . . . . .	2
1.3	Some topological concepts . . . . .	3
1.4	Hilbert space . . . . .	4
1.5	Examples of Hilbert spaces . . . . .	5
1.6	The Projection Theorem . . . . .	10
1.7	Bounded (continuous) linear functionals on a Hilbert space . . . . .	14
1.8	Bounded (continuous) linear operators on a Hilbert space . . . . .	17
1.9	The Lax–Milgram and the Galerkin theorems . . . . .	19
<b>2</b>	<b>Elements of the Theory of Sobolev Spaces and Variational Formulation of Boundary–Value Problems in One Dimension</b>	<b>26</b>
2.1	Motivation . . . . .	26
2.2	Notation and preliminaries . . . . .	28
2.3	The Sobolev space $H^1(I)$ . . . . .	31
2.4	The Sobolev spaces $H^m(I)$ , $m = 2, 3, 4, \dots$ . . . . .	46
2.5	The space $\overset{0}{H}^1(I)$ . . . . .	47
2.6	Two–point boundary–value problems . . . . .	53
2.6.1	Zero Dirichlet boundary conditions. . . . .	53
2.6.2	Neumann boundary conditions. . . . .	58
<b>3</b>	<b>Galerkin Finite Element Methods for Two–Point Boundary–Value Problems</b>	<b>63</b>
3.1	Introduction . . . . .	63

3.2	The Galerkin–finite element method with piecewise linear, continuous functions . . . . .	65
3.3	An indefinite problem . . . . .	78
3.4	Approximation by Hermite cubic functions and cubic splines . . . . .	84
3.4.1	Hermite, piecewise cubic functions . . . . .	84
3.4.2	Cubic splines . . . . .	90
<b>4</b>	<b>Results from the Theory of Sobolev Spaces and the Variational Formulation of Elliptic Boundary–Value Problems in <math>\mathbb{R}^N</math></b>	<b>99</b>
4.1	The Sobolev space $H^1(\Omega)$ . . . . .	99
4.2	The Sobolev space $\overset{0}{H}{}^1(\Omega)$ . . . . .	103
4.3	The Sobolev spaces $H^m(\Omega)$ , $m = 2, 3, 4, \dots$ . . . . .	104
4.4	Sobolev’s inequalities. . . . .	106
4.5	Variational formulation of some elliptic boundary–value problems. . . . .	107
4.5.1	(a) Homogeneous Dirichlet boundary conditions. . . . .	107
4.5.2	(b) Homogeneous Neumann boundary conditions. . . . .	110
<b>5</b>	<b>The Galerkin Finite Element Method for Elliptic Boundary–Value Problems</b>	<b>112</b>
5.1	Introduction . . . . .	112
5.2	Piecewise linear, continuous functions on a triangulation of a plane polygonal domain . . . . .	114
5.3	Implementation of the finite element method with $P_1$ triangles . . . . .	128
<b>6</b>	<b>The Galerkin Finite Element Method for the Heat Equation</b>	<b>139</b>
6.1	Introduction. Elliptic projection . . . . .	139
6.2	Standard Galerkin semidiscretization . . . . .	141
6.3	Full discretization with the implicit Euler and the Crank-Nicolson method	147
6.4	The explicit Euler method. Inverse inequalities and stiffness . . . . .	154
<b>7</b>	<b>The Galerkin Finite Element Method for the Wave Equation</b>	<b>161</b>
7.1	Introduction . . . . .	161
7.2	Standard Galerkin semidiscretization . . . . .	162

7.3 Fully discrete schemes . . . . .	167
<b>References</b>	<b>179</b>

# Chapter 1

## Some Elements of Hilbert Space Theory

### 1.1 Vector spaces

A set  $V$  of elements  $u, v, w, \dots$  is called a *vector space* (over the complex numbers) if

1. For every pair of elements  $u \in V, v \in V$  we define a new element  $w \in V$ , their *sum*, denoted by  $w = u + v$ .
2. For every complex number  $\lambda$  and every  $u \in V$  we define an element  $w = \lambda u \in V$ , the *product* of  $\lambda$  and  $u$ .
3. Sum and product obey the following laws:
  - i.  $\forall u, v \in V : u + v = v + u$ .
  - ii.  $\forall u, v, w \in V : (u + v) + w = u + (v + w)$ .
  - iii.  $\exists 0 \in V$  such that  $u + 0 = u, \forall u \in V$ .
  - iv.  $\forall u \in V \exists (-u) \in V$  such that  $u + (-u) = 0$ .
  - v.  $1 \cdot u = u, \forall u \in V$ .
  - vi.  $\lambda(\mu u) = (\lambda\mu)u, \forall u \in V$  and complex  $\lambda, \mu$ .
  - vii.  $(\lambda + \mu)u = \lambda u + \mu u$ , for  $u \in V, \lambda, \mu$  complex.
  - viii.  $\lambda(u + v) = \lambda u + \lambda v$ , for  $u, v \in V, \lambda$  complex.

The elements  $u, v, w, \dots$  of  $V$  are called *vectors*.

An expression of the form

$$\lambda_1 u_1 + \lambda_2 u_2 + \dots + \lambda_n u_n,$$

where  $\lambda_i$  complex numbers and  $u_i \in V$  is called a *linear combination* of the vectors  $u_i$ ,  $1 \leq i \leq n$ . The vectors  $u_1, \dots, u_n$  are called *linearly dependent* if there exist complex numbers  $\lambda_i$ , not all zero, for which:

$$\lambda_1 u_1 + \lambda_2 u_2 + \dots + \lambda_n u_n = 0.$$

They are called *linearly independent* if they are not linearly dependent, i.e. if  $\lambda_1 u_1 + \lambda_2 u_2 + \dots + \lambda_n u_n = 0$  holds only in the case  $\lambda_1 = \lambda_2 = \dots = \lambda_n = 0$ .

A vector space  $V$  is called *finite-dimensional* (of dimension  $n$ ) if  $V$  contains  $n$  linearly independent elements and if any  $n + 1$  vectors in  $V$  are linearly dependent. As a consequence, a set of  $n$  linearly independent vectors forms a *basis* of  $V$ , i.e. it is a set of linearly independent vectors that *spans*  $V$ , i.e. such that any  $u$  in  $V$  can be written uniquely as a linear combination of the basis vectors.

## 1.2 Inner product, Norm

A vector space  $V$  is called an *inner product space* if for every pair of elements  $u \in V$ ,  $v \in V$  we can define a complex number, denoted by  $(u, v)$  and called the *inner product* of  $u$  and  $v$ , with the following properties:

1.  $\forall u \in V : (u, u) \geq 0$ . If  $(u, u) = 0$  then  $u = 0$ .
2.  $(u, v) = \overline{(v, u)}$ ,  $\forall u, v \in V$ , where  $\bar{z}$  is the complex conjugate of the complex number  $z$ .
3.  $(\lambda u + \mu v, w) = \lambda(u, w) + \mu(v, w)$  for  $u, v, w \in V$ ,  $\lambda, \mu$  complex.

As a consequence of (2) and (3)  $(u, \lambda u) = \bar{\lambda} (u, u)$  for  $u, v \in V$  and  $\lambda$  complex. The vectors  $u, v$  are called *orthogonal* if  $(u, v) = 0$ .

For every  $u \in V$  we define the nonnegative number  $\|u\|$  by

$$\|u\| = (u, u)^{\frac{1}{2}},$$

which is called the *norm* of  $u$ . As a consequence of the properties of the inner product we see that:

- i.  $\|u\| \geq 0$  and if  $\|u\| = 0$ , then  $u = 0$
- ii.  $\forall$  complex  $\lambda, u \in V$ :  $\|\lambda u\| = |\lambda| \|u\|$
- iii.  $\forall u, v \in V$ :  $\|u + v\| \leq \|u\| + \|v\|$  (*Triangle inequality*).

To prove (iii) we first prove the *Cauchy–Schwarz Inequality*:

$$\text{iv. } |(u, v)| \leq \|u\| \|v\|, \forall u, v \in V.$$

To prove (iv) we may assume that  $(u, v) \neq 0$ . We let now  $\theta = \frac{(u, v)}{|(u, v)|}$ . We find then for any real  $\lambda$  that

$$0 \leq (\bar{\theta}u + \lambda v, \bar{\theta}u + \lambda v) = \lambda^2(v, v) + 2\lambda |(u, v)| + (u, u).$$

Hence for any real  $\lambda$  the quadratic on the right hand side of the above inequality is nonnegative. Hence, necessarily,

$$|(u, v)|^2 \leq (u, u)(v, v),$$

which gives (iv). To prove now the triangle inequality (iii) we see that

$$\begin{aligned} \|u + v\|^2 &= (u + v, u + v) = (u, u) + (v, v) + (u, v) + (v, u) \\ &\leq \|u\|^2 + \|v\|^2 + 2 |(u, v)| \leq \|u\|^2 + \|v\|^2 + 2 \|u\| \|v\| \\ &= (\|u\| + \|v\|)^2, \end{aligned}$$

from which (iii) follows. (Supplied only with a norm that just satisfies properties (i)–(iii),  $V$  becomes a *normed vector space*).

### 1.3 Some topological concepts

In  $V$  we define the *distance*  $\rho$  of two vectors  $u$  and  $v$  as  $\rho(u, v) = \|u - v\|$ . If  $u_0$  is a fixed vector in  $V$  and  $\delta$  a given positive number, then the set of vectors  $v$  in  $V$  which satisfy  $\|v - u_0\| < \delta$  is called the *open ball with center  $u_0$  and radius  $\delta$* . The set  $\|v - u_0\| \leq \delta$  is the *closed ball with center  $u_0$  and radius  $\delta$* . We say that a sequence of



vectors  $u_1, u_2, u_3, \dots$  in  $V$  is *convergent* if there exists a vector  $u \in V$  such that, given  $\epsilon > 0$  there exists a positive integer  $N = N(\epsilon)$  for which:

$$\|u_n - u\| < \epsilon, \text{ for all } n \geq N.$$

We call  $u$  the *limit* of the sequence  $\{u_i\}_{i \geq 1}$  and write  $\lim_n u_n = u$  or  $u_n \rightarrow u$  in  $V$  as  $n \rightarrow \infty$ . It is easy to see that a convergent sequence has only one limit. Obviously  $u_n \rightarrow u$  in  $V \Leftrightarrow \|u_n - u\| \rightarrow 0$  as  $n \rightarrow \infty$ .

A sequence of vectors  $u_1, u_2, u_3, \dots$  is said to be a *Cauchy sequence* if given any  $\epsilon > 0$ , there exists an integer  $N = N(\epsilon)$  such that

$$\|u_n - u_m\| < \epsilon, \text{ for all } m, n \geq N.$$

It is easy to see that every convergent sequence is Cauchy. The converse is not always true. We will say that  $V$  is *complete* whenever every Cauchy sequence in  $V$  is convergent. A subset  $A$  of  $V$  is called a *dense* subset of  $V$  if for every  $u \in V$  there exists a sequence  $u_1, u_2, u_3, \dots \in A$  such that  $u_n \rightarrow u$  as  $n \rightarrow \infty$ .

**Exercise:** If  $u_n \rightarrow u, v_n \rightarrow v$  in  $V$  then:

- (a)  $\lim_n(\lambda u_n + \mu v_n) = \lambda u + \mu v$  for complex  $\lambda, \mu$ .
- (b)  $\lim_n(u_n, v_n) = (u, v)$  (and as a consequence we say that the inner product is a continuous function of its arguments).
- (c)  $\lim_n \|u_n\| = \|u\|$ .
- (d)  $\lim_n \lambda_n u_n = \lambda u$  for every convergent sequence  $\lambda_n \rightarrow \lambda$  of complex numbers.  $\diamond$

## 1.4 Hilbert space

A *complete inner product space*  $V$  is called a *Hilbert space*. In other words, a Hilbert space is an inner product space in which a Cauchy sequence is always convergent. We will usually denote Hilbert spaces by  $H$ .

A subset  $S$  of a Hilbert space  $H$  is called a *subspace* of  $H$  if  $u \in S, v \in S$  imply that  $\lambda u + \mu v \in S$ , for any complex numbers  $\lambda, \mu$ .  $S$  is said to be a *dense subspace* of  $H$  if it is a dense subset of  $H$  and a subspace of  $H$ .  $S$  is said to be a *closed subspace* of  $H$

if  $S$  is a subspace of  $H$  with the following property: let  $\{u_n\}$  be a convergent sequence in  $H$  such that  $u_n \in S$ ,  $n = 1, 2, 3, \dots$ . Then  $u = \lim_n u_n$  belongs to  $S$  too.

**Exercise:** A dense, closed subspace of  $H$  coincides with  $H$ . ◇

Given any (noncomplete) inner product space  $V$  we can prove that by adding new elements to  $V$  we can extend  $V$  to a (complete) Hilbert space  $H$  such that  $V$  is a dense subspace of  $H$ . The process is referred to as *completion* of  $V$  or as *closure* of  $V$  in  $H$ .

## 1.5 Examples of Hilbert spaces

**A.**  $H = \mathbb{C}^n$  with the Euclidean inner product  $(u, v) = \sum_{i=1}^n u_i \bar{v}_i$  and norm  $\|u\| = (\sum_{i=1}^n |u_i|^2)^{1/2}$ .

**B.**  $l_2$

We denote by  $l_2$  the set of all complex sequences  $u = \{u_1, u_2, u_3, \dots\} \equiv \{u_i\}_{i=1}^{\infty}$  which satisfy the inequality

$$\sum_{j=1}^{\infty} |u_j|^2 < \infty.$$

In  $l_2$  we define  $u + v$ ,  $\lambda u$  and  $(u, v)$  in the following way:

$$w = u + v \text{ with } w = \{w_i\}_{i=1}^{\infty}, w_i = u_i + v_i.$$

$$z = \lambda u \text{ (}\lambda \text{ complex number)} z = \{z_i\}_{i=1}^{\infty}, z_i = \lambda u_i.$$

$$(u, v) = \sum_{j=1}^{\infty} u_j \bar{v}_j.$$

Then it follows that  $\sum_{j=1}^{\infty} |z_j|^2 < \infty$  and  $\sum_{j=1}^{\infty} |w_j|^2 < \infty$  because

$$|w_j|^2 = |u_j + v_j|^2 \leq 2|u_j|^2 + 2|v_j|^2.$$

The convergence of the series defining the inner product follows from

$$|u_j \bar{v}_j| = |u_j| |v_j| \leq \frac{1}{2} \{|u_j|^2 + |v_j|^2\}.$$

By verifying the axioms one by one we easily conclude that the set of all vectors  $u, v, w, \dots$  with the above mentioned properties and the given operations form an inner product space. (The zero vector is the sequence  $\mathbf{0} = \{0, 0, \dots\}$ ). The space  $l_2$  is a

complete inner product space and hence a Hilbert space. To show that let  $u^{(1)}, u^{(2)}, \dots$  be a Cauchy sequence in  $l_2$  with

$$u^{(n)} = \{u_1^{(n)}, u_2^{(n)}, \dots\}.$$

Then given  $\epsilon > 0$

$$\|u^{(n)} - u^{(m)}\| = (u^{(n)} - u^{(m)}, u^{(n)} - u^{(m)})^{\frac{1}{2}} = \left( \sum_{j=1}^{\infty} |u_j^{(n)} - u_j^{(m)}|^2 \right)^{\frac{1}{2}} < \epsilon \quad (1.1)$$

for all  $n, m \geq N(\epsilon)$ . In particular it follows that

$$|u_j^{(n)} - u_j^{(m)}| < \epsilon, \text{ for all } n, m \geq N(\epsilon) \text{ and every } j = 1, 2, 3, \dots$$

Fix  $j$ . Then the sequence  $u_j^{(1)}, u_j^{(2)}, \dots$  is convergent. We denote the limit of this sequence by  $u_j$ , i.e.

$$\lim_{n \rightarrow \infty} u_j^{(n)} = u_j \text{ for } j = 1, 2, 3, \dots \quad (1.2)$$

Now it follows from (1.1) that for every positive integer  $k$

$$\sum_{j=1}^k |u_j^{(n)} - u_j^{(m)}|^2 < \epsilon^2 \text{ for all } n, m \geq N(\epsilon).$$

Letting  $m \rightarrow \infty$  in the above, since it is a finite sum, we obtain by (1.2) that

$$\sum_{j=1}^k |u_j^{(n)} - u_j|^2 \leq \epsilon^2 \text{ for all } n, m \geq N(\epsilon).$$

Letting now  $k \rightarrow \infty$  in the above, we obtain that

$$\sum_{j=1}^{\infty} |u_j^{(n)} - u_j|^2 \leq \epsilon^2 \text{ for all } n, m \geq N(\epsilon). \quad (1.3)$$

We set now  $u = \{u_1, u_2, \dots\}$ . By (1.3),  $u - u^{(n)} \in l_2$ . Hence  $u = (u - u^{(n)}) + u^{(n)} \in l_2$ . By (1.3) it also follows that

$$\|u^{(n)} - u\| = \left( \sum_{j=1}^{\infty} |u_j^{(n)} - u_j|^2 \right)^{\frac{1}{2}} < \epsilon \text{ for all } n \geq N(\epsilon).$$

Hence there exists  $u \in l_2$  such that  $u^{(n)} \rightarrow u$  as  $n \rightarrow \infty$ . We conclude that the Cauchy sequence  $\{u^{(1)}, u^{(2)}, \dots\}$  is actually a convergent sequence. Therefore  $l_2$  is

complete and hence a Hilbert space. We remark that the Cauchy–Schwarz inequality  $|(u, v)| \leq \|u\|\|v\|$  becomes

$$\left| \sum_{j=1}^{\infty} u_j \bar{v}_j \right| \leq \left( \sum_{j=1}^{\infty} |u_j|^2 \right)^{\frac{1}{2}} \left( \sum_{j=1}^{\infty} |v_j|^2 \right)^{\frac{1}{2}}.$$

The triangle inequality  $\|u + v\| \leq \|u\| + \|v\|$  becomes

$$\left( \sum_{j=1}^{\infty} |u_j + v_j|^2 \right)^{\frac{1}{2}} \leq \left( \sum_{j=1}^{\infty} |u_j|^2 \right)^{\frac{1}{2}} + \left( \sum_{j=1}^{\infty} |v_j|^2 \right)^{\frac{1}{2}}.$$

### C. $L_2(\Omega)$

Let  $\Omega$  be an open set of  $\mathbb{R}^n$ . We describe the points in  $\mathbb{R}^n$  by  $n$ -tuples  $x = (x_1, x_2, \dots, x_n)$  and denote the (Euclidean) length of the vector  $x$  by:

$$|x| = \left( \sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}.$$

We now consider the set of complex-valued continuous functions  $u(x) = u(x_1, \dots, x_n)$  defined on  $\Omega$ . Addition  $u+v$  and multiplication  $\lambda u$  by a complex number  $\lambda$  are defined, as usual, by:

$$\begin{aligned} w = u + v, \quad w(x) &= u(x) + v(x), \\ z = \lambda u, \quad z(x) &= \lambda u(x) \end{aligned}$$

We define now an *inner product* for such functions by:

$$(u, v) = \int_{\Omega} u(x) \overline{v(x)} dx, \tag{1.4}$$

where  $dx$  is the volume element in  $\Omega$ , i.e.  $dx = dx_1 dx_2 \dots dx_n$  and  $\int_{\Omega} \dots dx$  is the multiple integral (in the Riemann sense)

$$\int_{\Omega} \dots dx = \underbrace{\int \int \int \dots \int}_{\Omega} \dots dx_1 dx_2 \dots dx_n.$$

Since  $\Omega$  is an arbitrary open set of  $\mathbb{R}^n$ , the integral in (1.4) defining the inner product may not exist. We restrict therefore our attention to complex-valued functions  $u(x)$ , defined on  $\Omega$ , with the property that

$$\int_{\Omega} |u(x)|^2 dx < \infty.$$

Let now  $V$  be the above-described vector space, i.e. let

$$V = \{u \mid u \text{ continuous on } \Omega, \int_{\Omega} |u(x)|^2 dx < \infty\}.$$

$V$  is an inner product space with the inner product defined by (1.4). To see this, note that

$$|u(x) + v(x)|^2 \leq 2(|u(x)|^2 + |v(x)|^2).$$

It follows that  $u \in V, v \in V \Rightarrow u + v \in V$  and easily,  $\lambda u \in V$  for  $\lambda$  complex. Finally the existence of the integral in (1.4) is proved by noting that  $\forall x \in \Omega$ :

$$2|u(x)||v(x)| \leq |u(x)|^2 + |v(x)|^2.$$

Hence, integrating:

$$\begin{aligned} |(u, v)| &= \left| \int_{\Omega} u(x)\overline{v(x)}dx \right| \leq \int_{\Omega} |u(x)||v(x)|dx \\ &\leq \frac{1}{2} \int_{\Omega} |u(x)|^2 dx + \frac{1}{2} \int_{\Omega} |v(x)|^2 dx. \end{aligned}$$

By verifying now the axioms one by one we confirm that  $V$  is an inner product space. The norm on  $V$  is given by

$$\|u\| \equiv (u, u)^{\frac{1}{2}} = \left( \int_{\Omega} |u(x)|^2 dx \right)^{\frac{1}{2}}.$$

The zero element in  $V$  is the function  $u(x) \equiv 0, x \in \Omega$ . For  $u, v \in V$ , the Cauchy-Schwarz and the triangle inequalities take the form:

$$\begin{aligned} \left| \int_{\Omega} u(x)\overline{v(x)}dx \right| &\leq \left( \int_{\Omega} |u(x)|^2 dx \right)^{\frac{1}{2}} \left( \int_{\Omega} |v(x)|^2 dx \right)^{\frac{1}{2}} \\ \left( \int_{\Omega} |u(x) + v(x)|^2 dx \right)^{\frac{1}{2}} &\leq \left( \int_{\Omega} |u(x)|^2 dx \right)^{\frac{1}{2}} + \left( \int_{\Omega} |v(x)|^2 dx \right)^{\frac{1}{2}}. \end{aligned}$$

The functions  $u_1, u_2, \dots \in V$  form a *Cauchy sequence* in  $V$  if  $\forall \epsilon > 0$

$$\|u_m - u_n\| = \left( \int_{\Omega} |u_m(x) - u_n(x)|^2 dx \right)^{\frac{1}{2}} < \epsilon,$$

for all  $m, n \geq N = N(\epsilon)$ . The sequence is *convergent* if there exists a function  $u \in V$  such that for every  $\epsilon > 0$  there exists an integer  $N = N(\epsilon)$  such that

$$\|u_n - u\| = \left( \int_{\Omega} |u_n(x) - u(x)|^2 dx \right)^{\frac{1}{2}} < \epsilon$$

holds for all  $n \geq N(\epsilon)$ .

The space  $V$  is not complete. To see that let  $\Omega = (-1, 1)$  in  $\mathbb{R}$  and let  $V$  be the set of continuous, real-valued functions  $u$  defined on  $(-1, 1)$  such that  $\int_{-1}^1 |u(x)|^2 dx < \infty$ . Let  $(u, v)$  be defined as  $\int_{-1}^1 u(x)v(x)dx$  and let  $\|u\| = (u, u)^{\frac{1}{2}}$ . Consider the sequence  $u_1, u_2, \dots$ , where

$$u_j(x) = \begin{cases} -1 & \text{for } -1 < x < -\frac{1}{j} \\ jx & \text{for } -\frac{1}{j} \leq x \leq \frac{1}{j} \\ 1 & \text{for } \frac{1}{j} < x < 1. \end{cases}$$

**Exercise:** Prove that  $\{u_j\}_{j=1}^{\infty}$  is a Cauchy sequence in  $V$ . ◇

However, there is *no* continuous function  $u$  on  $(-1, 1)$  for which  $\|u_n - u\| \rightarrow 0$  as  $n \rightarrow \infty$ . It is easy to see that  $\|u_n - f\| \rightarrow 0$  as  $n \rightarrow \infty$  where  $f$  is the discontinuous function

$$f(x) = \begin{cases} -1 & \text{for } -1 < x < 0 \\ 0 & \text{for } x = 0 \\ 1 & \text{for } 0 < x < 1. \end{cases}$$

Thus,  $V$  is a non-complete inner product space. By our assertion in §1.4 we can *complete* the space  $V$  by ‘adding’ new elements to it. This extended *complete* space we call  $L_2(\Omega)$ . The elements that we add to  $V$  can be considered as representatives of those Cauchy sequences of  $V$  for which there do not exist functions  $u \in V$  such that  $\|u_n - u\| \rightarrow 0$ . Some of those ‘additional’ functions may be piecewise continuous but in general they will be highly discontinuous functions defined on  $\Omega$ . (For example  $f$ , above, belongs to  $L_2(\Omega)$ ).

It is well-known that  $L_2(\Omega)$  is isometrically isomorphic to the set of (equivalence classes of) complex-valued Lebesgue measurable functions  $u$  on  $\Omega$  for which the (Lebesgue) integral  $\int_{\Omega} |u(x)|^2 dx$  is finite. The inner product (understood in the Lebesgue sense) is given again by (1.4).

As a consequence of the process of completion of  $V$ :

- (i)  $L_2(\Omega)$  is a complete inner product space (a Hilbert space).
- (ii)  $V$  is dense in  $L_2(\Omega)$ , i.e. for every  $u \in L_2(\Omega)$  there exists a sequence  $\{u_n\}_{n=1}^{\infty}$  of functions in  $V$  such that  $\|u_n - u\| = \left(\int_{\Omega} |u_n(x) - u(x)|^2 dx\right)^{\frac{1}{2}} \rightarrow 0$  as  $n \rightarrow \infty$ .

## 1.6 The Projection Theorem

The following result provides important information about geometric properties of a Hilbert space.

**Theorem 1.1** (Projection Theorem). *Let  $G$  be a closed subspace of a Hilbert space  $H$ , properly included in  $H$ . Then, given  $h \in H$  there exists a unique element  $g \in G$  such that:*

$$(i) \quad \|h - g\| = \inf_{\phi \in G} \|h - \phi\|.$$

Moreover

$$(ii) \quad (h - g, \phi) = 0 \text{ for each } \phi \in G.$$

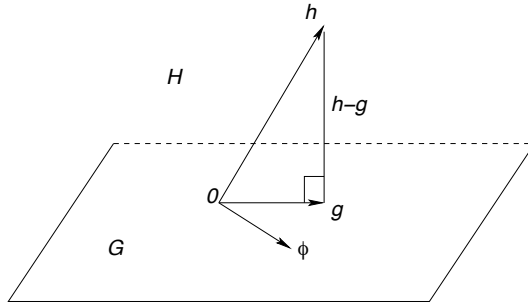


Figure 1.1:

**Proof.** Let  $h \in H$  such that  $h \notin G$  (if  $h \in G$  pick  $g = h$  and the theorem is proved). Now, since for all  $\phi \in G$ ,  $\|h - \phi\| \geq 0$ , there exists a sequence  $\phi_n \in G$  such that

$$\lim_{n \rightarrow \infty} \|h - \phi_n\| = \inf_{\phi \in G} \|h - \phi\| \equiv \delta \quad (1.5)$$

We first show that  $\{\phi_n\}_{n \geq 1}$  is a Cauchy sequence. Let  $f_1, f_2 \in H$ . Then the *parallelogram law* holds:

$$2\|f_1\|^2 + 2\|f_2\|^2 = \|f_1 + f_2\|^2 + \|f_1 - f_2\|^2.$$

(Proof: **Exercise**). Set  $f_1 = h - \phi_m$ ,  $f_2 = h - \phi_n$ . Then

$$\|\phi_m - \phi_n\|^2 = 2\|h - \phi_m\|^2 + 2\|h - \phi_n\|^2 - 4\|h - \frac{\phi_m + \phi_n}{2}\|^2. \quad (1.6)$$

Now

$$\|h - \frac{\phi_m + \phi_n}{2}\| \leq \frac{1}{2}\|h - \phi_m\| + \frac{1}{2}\|h - \phi_n\|$$

and by (1.5),

$$\limsup_{m,n \rightarrow \infty} \left\| h - \frac{\phi_m + \phi_n}{2} \right\| \leq \frac{1}{2}\delta + \frac{1}{2}\delta = \delta.$$

But by definition of  $\delta$ ,

$$\liminf_{m,n \rightarrow \infty} \left\| h - \frac{\phi_m + \phi_n}{2} \right\| \geq \delta.$$

Hence  $\lim_{m,n \rightarrow \infty} \left\| h - \frac{\phi_m + \phi_n}{2} \right\| = \delta$ . Then by (1.6) and (1.5) we conclude that

$$\lim_{m,n \rightarrow \infty} \|\phi_m - \phi_n\|^2 = 2\delta^2 + 2\delta^2 - 4\delta^2 = 0.$$

Hence  $\{\phi_n\}_{n \geq 1}$  is a Cauchy sequence in  $G$ . Since  $G$  is closed, the sequence is convergent in  $G$ , i.e. there exists  $g \in G$  such that  $\phi_n \rightarrow g$  as  $n \rightarrow \infty$ . We show that this  $g$  satisfies  $\|h - g\| = \inf_{\phi \in G} \|h - \phi\| \equiv \delta$ . Obviously  $\|h - g\| \leq \|h - \phi_n\| + \|g - \phi_n\|$  and taking the limit of both sides as  $n \rightarrow \infty$  we get that  $\|h - g\| \leq \delta$ . By definition of  $\delta$ ,  $\|h - g\| \geq \delta$ . Hence  $\|h - g\| = \delta$  as required. We also show that  $g$  is *unique*. Indeed, let  $g_1, g_2 \in G$ ,  $g_1 \neq g_2$  have the property that

$$\delta \equiv \inf_{g \in G} \|h - g\| = \|h - g_1\| = \|h - g_2\|.$$

Then, since  $\frac{1}{2}(g_1 + g_2) \in G \Rightarrow \delta \leq \|h - \frac{1}{2}(g_1 + g_2)\|$ . But by the triangle inequality

$$\left\| h - \frac{1}{2}(g_1 + g_2) \right\| \leq \frac{1}{2}\|h - g_1\| + \frac{1}{2}\|h - g_2\| = \frac{1}{2}\delta + \frac{1}{2}\delta = \delta.$$

Hence

$$\left\| h - \frac{1}{2}(g_1 + g_2) \right\| = \delta = \frac{1}{2}\|h - g_1\| + \frac{1}{2}\|h - g_2\|,$$

i.e. the triangle inequality holds as *equality*. Now for any  $\chi, \psi \in H$  such that  $\chi \neq 0$ ,  $\psi \neq 0$ ,  $\|\chi + \psi\| = \|\chi\| + \|\psi\| \Leftrightarrow \chi = \lambda\psi$ , for some  $\lambda > 0$ . (Proof: **Exercise**.) Hence there exists  $\lambda$  such that  $h - g_1 = \lambda(h - g_2)$ , i.e.  $h(1 - \lambda) = g_1 - \lambda g_2$ . If  $\lambda = 1$ ,  $g_1 = g_2$  (contradiction). If  $\lambda \neq 1$ ,  $h = \frac{(g_1 - \lambda g_2)}{1 - \lambda}$  i.e.  $h \in G$  (contradiction). Hence  $g$  is unique and we proved (i) above.

To prove (ii), with  $g$  constructed as above, suppose that there exists  $\phi_* \neq 0$  in  $G$  for which (ii) fails, i.e.  $(h - g, \phi_*) \neq 0$ . Define the element  $g_* \in G$  by:

$$g_* = g + \frac{(h - g, \phi_*)}{(\phi_*, \phi_*)} \phi_*.$$



Then

$$\begin{aligned}
\|h - g_*\|^2 &= (h - g - \frac{(h - g, \phi_*)}{\|\phi_*\|^2} \phi_*, h - g - \frac{(h - g, \phi_*)}{\|\phi_*\|^2} \phi_*) \\
&= (h - g, h - g) - \frac{(h - g, \phi_*)}{(\phi_*, \phi_*)} (\phi_*, h - g) \\
&\quad - \frac{\overline{(h - g, \phi_*)}}{(\phi_*, \phi_*)} (h - g, \phi_*) + \frac{|(h - g, \phi_*)|^2}{\|\phi_*\|^4} (\phi_*, \phi_*) \\
&= \|h - g\|^2 - \frac{|(h - g, \phi_*)|^2}{\|\phi_*\|^2}.
\end{aligned}$$

Hence

$$\|h - g_*\| < \|h - g\| = \inf_{\phi \in G} \|h - \phi\| \quad (\text{contradiction}).$$

Hence  $(h - g, \phi) = 0, \forall \phi \in G$  and we have (ii).  $\square$

**Exercise:** Prove that if for some  $g \in G, (h - g, \phi) = 0 \forall \phi \in G$ , then (i) in Theorem 1.1 holds.  $\diamond$

Given  $h \in H$  we call  $g$ , the existence and uniqueness of which is guaranteed by Theorem 1.1, the *orthogonal projection* of  $h$  on the closed subspace  $G$  or the *best approximation* of  $h$  in  $G$ . If we denote by  $f = h - g$ , then we can write

$$h = f + g \quad \text{where } g \in G \text{ and } (f, \phi) = 0, \forall \phi \in G.$$

Hence  $f$  is *orthogonal* to all vectors of the closed subspace  $G$ . Let  $G^\perp$  be the set of all such vectors, i.e. let

$$G^\perp = \{u \in H : (u, \phi) = 0 \forall \phi \in G\}.$$

$G^\perp$  is called the *orthogonal complement* of  $G$  and it is a *closed subspace* of  $H$ . To see that let  $u_n \in G^\perp$  such that  $u_n \rightarrow u$ . Then  $(u, \phi) = (u, \phi) - (u_n, \phi)$ , for all  $\phi \in G$  since  $(u_n, \phi) = 0$ . Hence for all  $\phi \in G$   $|(u, \phi)| = |(u - u_n, \phi)| \leq \|u - u_n\| \|\phi\| \rightarrow 0$  as  $n \rightarrow \infty$ , i.e.  $(u, \phi) = 0 \Rightarrow u \in G^\perp$ . Hence  $G^\perp$  is closed. (To show that it is a subspace is trivial). It is easy to see that  $G \cap G^\perp = \{0\}$ . Hence we can write  $H$  as the *direct sum* of  $G$  and  $G^\perp$

$$H = G \oplus G^\perp,$$

meaning by that that there exist two disjoint closed subspaces  $G$  and  $G^\perp$  such that every element  $h \in H$  can be written uniquely as the sum of  $g \in G$  and  $f \in G^\perp$ ,

$h = g + f$ , as above.

**Exercise:** With  $g, f$  defined as above prove the Pythagorean theorem:

$$\|h\|^2 = \|g\|^2 + \|f\|^2.$$

◇

A particular case of importance occurs when  $G$  is *finite-dimensional*. Then  $G$  is *closed* (Proof: **Exercise**). Let  $\{\varphi_1, \varphi_2, \dots, \varphi_s\}$  be a basis of  $G$ . Given  $h \in H$  we can *explicitly* construct the best approximation  $g$  of  $h$  in  $G$  as follows: By (ii),  $g$  satisfies:

$$(h - g, \varphi) = 0 \quad \forall \varphi \in G.$$

Hence

$$(h - g, \varphi_i) = 0, \quad i = 1, 2, \dots, s. \quad (1.7)$$

Let  $g = \sum_{i=1}^s c_i \varphi_i$ . We seek the coefficients  $\{c_i\}_{i=1}^s$ . By (1.7), the  $c_i$ 's satisfy the following linear system of equations:

$$\sum_{j=1}^s M_{ij} c_j = (h, \varphi_i) \quad 1 \leq i \leq s, \quad (1.8)$$

where  $M = \{M_{ij}\}$  is the  $s \times s$  *Gram matrix* (or *mass matrix*) associated with the basis  $\{\varphi_i\}_{i=1}^s$  of  $G$  and defined by

$$M_{ij} = (\varphi_j, \varphi_i), \quad 1 \leq i, j \leq s.$$

To see that  $M$  is invertible, suppose that for some complex  $s$ -vector  $d = [d_1, \dots, d_s]^T$  we have that  $M d = 0$ . Thus

$$\sum_{j=1}^s M_{ij} d_j = 0 \Rightarrow \left( \sum_{j=1}^s d_j \varphi_j, \varphi_i \right) = 0 \quad \forall i: 1 \leq i \leq s.$$

Hence we conclude easily that the vector  $u = \sum_{j=1}^s d_j \varphi_j \in G$  is orthogonal to *all*  $\varphi \in G$ , i.e.  $u \in G^\perp \cap G \Rightarrow u = 0$ . Hence  $\sum_{j=1}^s d_j \varphi_j = 0$  and by the linear independence of the  $\varphi_j$ 's  $\Rightarrow d_j = 0, \forall j \Rightarrow d = 0$ . Hence  $M d = 0 \Rightarrow d = 0$ , i.e.  $M$  is invertible. Actually,  $M$  is positive-definite (**Exercise**).

**Example:** Let  $\Omega = (0, 1)$  and  $H = L^2(0, 1)$  (real-valued). Suppose  $f$  is a given element of  $L^2(0, 1)$  and let  $G$  be the subspace of  $H$  consisting of all real-valued polynomials of degree  $\leq n - 1, n > 1$ . Find the best approximation to  $f$  in  $G$ .

*Solution.* A basis for  $G$  obviously consists of the functions  $\varphi_j(x) = x^{j-1}$ ,  $1 \leq j \leq n$ . Let  $g$  be the best approximation (orthogonal projection) of  $f$  in  $G$ . Suppose that  $g = \sum_{j=1}^n a_j \varphi_j$ . Then  $g$  satisfies:

$$(g - f, \varphi_i) = 0, \quad 1 \leq i \leq n,$$

from which

$$\sum_{j=1}^n M_{ij} a_j = (f, \varphi_i), \quad 1 \leq i \leq n, \quad (1.9)$$

where  $M$  is the  $n \times n$  Gram matrix,

$$M_{ij} = (\varphi_i, \varphi_j) = \int_0^1 \varphi_i \varphi_j = \int_0^1 x^{i+j-2} dx = \frac{1}{i+j-1}.$$

$M$  is positive-definitive but very ill conditioned. □

## 1.7 Bounded (continuous) linear functionals on a Hilbert space

Let  $H$  be a Hilbert space. By a *functional*  $F$  on  $H$  we mean a function from  $H$  into the complex numbers  $\mathbb{C}$ , i.e. a map which assigns to every  $\phi \in H$  a unique complex number  $F(\phi)$ ,

$$F : H \rightarrow \mathbb{C}, \quad \phi \mapsto F(\phi).$$

A functional  $F$  on  $H$  is called a *linear functional* if for every  $\phi, \psi \in H$  and  $\lambda, \mu \in \mathbb{C}$  :

$$F(\lambda\phi + \mu\psi) = \lambda F(\phi) + \mu F(\psi).$$

A functional  $F$  on  $H$  is called *bounded* if

$$\sup_{0 \neq \phi \in H} \frac{|F(\phi)|}{\|\phi\|} < \infty.$$

If a functional  $F$  on  $H$  is bounded we define its *norm*, denoted by  $\|F\|$  (do not confuse with the norm of  $\phi \in H$ ,  $\|\phi\|$ ) by

$$\|F\| = \sup_{0 \neq \phi \in H} \frac{|F(\phi)|}{\|\phi\|}. \quad (1.10)$$

Let  $F$  be a bounded, linear functional on  $H$ . Then it is easy to see that  $F$  is a continuous function of its argument. Indeed, let  $\phi_n \rightarrow \phi$  in  $H$ . Then

$$|F(\phi_n) - F(\phi)| = |F(\phi_n - \phi)| \leq \|F\| \|\phi_n - \phi\| \xrightarrow{n \rightarrow \infty} 0.$$

Hence  $F(\phi_n) \rightarrow F(\phi)$  in  $\mathbb{C}$ , i.e.  $F$  is continuous. (Note that the inequality

$$|F(\phi)| \leq \|F\| \|\phi\|$$

follows from the definition of the norm (1.10)  $\|F\|$  of  $F$ .)

Conversely, let  $F$  be a linear functional on  $H$  and suppose that  $F$  is a continuous function of its argument on  $H$ . We shall show that  $F$  is bounded. Indeed, if  $F$  is continuous at  $\phi_0 \in H$ , then for each  $\epsilon > 0$  there exists a  $\delta > 0$  such that  $|F(\phi_0) - F(h)| < \epsilon$  for  $\|h - \phi_0\| \leq \delta$ . Now let  $\phi \neq 0$  be an arbitrary element of  $H$ . By the linearity of  $F$  we obtain that

$$F(\phi) = \frac{\|\phi\|}{\delta} F\left(\frac{\delta\phi}{\|\phi\|}\right) = \frac{\|\phi\|}{\delta} \left\{ F\left(\frac{\delta\phi}{\|\phi\|} + \phi_0\right) - F(\phi_0) \right\}.$$

Since the vector  $\frac{\delta\phi}{\|\phi\|} + \phi_0 = h$ , satisfies the relation  $\|h - \phi_0\| \leq \delta$  we have that  $|F(\phi)| < \frac{\epsilon}{\delta} \|\phi\|$ , i.e.  $\frac{|F(\phi)|}{\|\phi\|} < \frac{\epsilon}{\delta}$ . Fix  $\epsilon = \epsilon_0 > 0$ . Then  $\delta = \delta(\epsilon_0) \equiv \delta_0$ , and since  $\epsilon_0, \delta_0$  are independent of  $\phi$  we see that

$$\|F\| = \sup_{0 \neq \phi \in H} \frac{|F(\phi)|}{\|\phi\|} \leq \frac{\epsilon_0}{\delta_0} < \infty,$$

i.e.  $F$  is bounded.

Hence we proved that for a linear functional  $f$  on  $H$ , boundedness  $\Leftrightarrow$  continuity. We speak thus of a *bounded (continuous) linear functional* (b.l.f.).

**Exercise:** Let  $F$  be a b.l.f. on  $H$ . With the norm  $\|F\|$  defined by (1.10) show that

$$\|F\| = \sup_{\phi \in H: \|\phi\| \leq 1} |F(\phi)| = \sup_{\phi \in H: \|\phi\| = 1} |F(\phi)|.$$

◇

Now let  $F, G$  be b.l.f.'s on a Hilbert space  $H$ . We can define the sum of two b.l.f.'s  $F + G = L$  by  $L(\phi) = F(\phi) + G(\phi)$  for each  $\phi \in H$  and the scalar product  $\lambda F$  as  $\lambda F = G, G(\phi) = \lambda F(\phi)$ .

**Exercise:** We denote by  $H'$  the space of bounded linear functionals  $F, G, \dots$  on a

Hilbert space  $H$ . ( $H'$  is called the *dual* of  $H$ ). With addition and scalar multiplication defined as above show that  $H'$  forms a vector space. Then  $\|F\|$ , defined by (1.10) is a *norm* on  $H'$ , i.e.  $H'$  is a normed linear space. Finally show that  $H'$  is *complete*, i.e. every Cauchy sequence in  $H'$  converges to an element in  $H'$ .  $\diamond$

An example of a bounded, linear functional on  $H$  is furnished by the *inner product* on  $H$  of the elements of  $H$  with a fixed element  $f \in H$ . Given  $f \in H$ , define for every  $\phi \in H$ ,  $F(\phi)$  by

$$F(\phi) = (\phi, f).$$

Clearly  $F$  is a linear functional on  $H$ . To see that it is bounded observe that

$$|F(\phi)| = |(\phi, f)| \leq \|\phi\| \|f\|.$$

So for all  $\phi \neq 0$ :

$$\frac{|F(\phi)|}{\|\phi\|} \leq \|f\| < \infty.$$

Hence

$$\|F\| = \sup_{0 \neq \phi \in H} \frac{|F(\phi)|}{\|\phi\|} \leq \|f\|.$$

In fact, since  $F(f) = \|f\|^2$  we see that the sup is attained for  $\phi = f \in H$ . Hence  $\|F\| = \|f\|$ .

It turns out that the *converse* of the above statement is also true. Namely that *every* bounded linear functional on  $H$  has the form  $(\phi, f)$  for *some*  $f \in H$ . This is the content of:

**Theorem 1.2** (Riesz Representation Theorem). *Every bounded (continuous) linear functional  $F$  on a Hilbert space  $H$  can be expressed in the form  $F(\phi) = (\phi, f)$ , for each  $\phi \in H$ , where  $f$  is an element of  $H$  which is uniquely determined by  $F$ ; moreover,  $\|F\| = \|f\|$ .*

**Proof.** We denote by  $G$  the set of all elements  $g \in H$  such that  $F(g) = 0$ , i.e.  $G = \text{Ker}F$ . Obviously  $G$  is a subspace of  $H$ . Moreover  $G$  is a *closed* subspace of  $H$ . To see that let  $g_n \rightarrow g$  with  $g_n \in G$ . Then  $F(g) = F(g - g_n) + F(g_n) = F(g - g_n)$ . Hence

$$|F(g)| \leq \|F\| \|g - g_n\| \rightarrow 0 \quad \text{as } n \rightarrow \infty, \text{ i.e. } F(g) = 0 \Leftrightarrow g \in G.$$

There are two possibilities now: either  $G = H$  or  $G \subsetneq H$  (properly included in  $H$ ). In the first case  $F$  is the *zero functional* on  $H$  and the theorem is proved with  $f = 0$ .

Hence, assume that  $G \subsetneq H$ . In this case  $G^\perp$  contains non-zero elements. Let  $f_0 \in G^\perp$ ,  $f_0 \neq 0$ . For  $\phi \in H$ , consider the vector  $F(\phi)f_0 - F(f_0)\phi$ . This vector belongs to  $G$  because  $F(F(\phi)f_0 - F(f_0)\phi) = F(\phi)F(f_0) - F(f_0)F(\phi) = 0$ . Hence, since  $f_0 \in G^\perp$ , we see that for all  $\phi \in H$ :

$$(F(\phi)f_0 - F(f_0)\phi, f_0) = 0.$$

Hence  $F(\phi)\|f_0\|^2 = F(f_0)(\phi, f_0)$  from which

$$F(\phi) = \left( \phi, \frac{\overline{F(f_0)}}{\|f_0\|^2} f_0 \right), \quad \text{for all } \phi \in H.$$

We set now

$$f = \frac{\overline{F(f_0)}}{\|f_0\|^2} f_0$$

and the equality above provides the required representation, i.e. we have *existence* of  $f$ .

To prove that  $f$  is *unique*, suppose that there exist two vectors  $f_1 \neq f_2$  such that for all  $\phi \in H$ :  $f(\phi) = (\phi, f_1) = (\phi, f_2)$ . Hence  $(\phi, f_1 - f_2) = 0$  for all  $\phi \in H$ . In particular, set  $\phi = f_1 - f_2$  from which it follows that  $f_1 - f_2 = 0$ . It remains to prove that  $\|F\| = \|f\|$ . We immediately obtain from  $F(\phi) = (\phi, f)$  that

$$|F(\phi)| = |(\phi, f)| \leq \|\phi\| \|f\| \xrightarrow{\phi \neq 0} \frac{|F(\phi)|}{\|\phi\|} \leq \|f\|.$$

Hence

$$\|F\| = \sup_{0 \neq \phi \in H} \frac{|F(\phi)|}{\|\phi\|} \leq \|f\|.$$

On the other hand taking  $\phi = f$  we see that  $F(f) = (f, f) = \|f\|^2$ , from which

$$\|f\|^2 \leq \|F\| \|f\|, \quad \text{i.e. } \|f\| \leq \|F\|.$$

Hence  $\|F\| = \|f\|$ . □

## 1.8 Bounded (continuous) linear operators on a Hilbert space

Let  $H$  be a Hilbert space as usual. Let  $M$  be a subspace of  $H$ . By a *linear operator*  $T : M \rightarrow H$  we mean a function defined on  $M$  with values in  $H$  which assigns to the

vector  $u \in M$  the (unique) vector  $Tu \in H$  and which satisfies:

$$T(\lambda u + \mu v) = \lambda T(u) + \mu T(v) \quad \text{for } u, v \in M, \lambda, \mu \in \mathbb{C}.$$

The subspace  $M$  of  $H$  on which  $T$  is defined is called the *domain* of  $T$  and is denoted by  $D(T)$ . The *range* of  $T$  is the set of vectors  $v \in H$  to each one of which there corresponds at least one  $u \in D(T)$  such that  $Tu = v$ , i.e.

$$\text{Range}(T) \equiv \text{Ran}(T) = \{v \in H : \exists u \in D(T) \text{ such that } Tu = v\}.$$

We also define

$$\text{Ker}(T) \equiv \{u \in D(T) : Tu = 0\}.$$

The operator  $T$  is called *one-to-one* (1-1) if  $u_1 \neq u_2 \Rightarrow Tu_1 \neq Tu_2$ . Equivalently,  $T$  is one-to-one if  $Tu_1 = Tu_2 \Rightarrow u_1 = u_2$ .  $T$  is called *onto* if  $\text{Ran}(T) = H$ , i.e. if for every  $v \in H$  we can find a  $u \in D(T)$  such that  $Tu = v$ .

**Exercise:**  $\text{Ran}(T)$  is a subspace of  $H$ . So is  $\text{Ker}(T)$ . ◇

A linear operator  $T$  defined on the whole of  $H$ , (i.e.  $D(T) = H$ ) will be called a linear operator *on*  $H$ . Unless otherwise indicated we will assume henceforth that  $D(T) = H$ . A linear operator  $T$  on  $H$  is said to be *bounded* if

$$\sup_{0 \neq \phi \in H} \frac{\|T\phi\|}{\|\phi\|} < \infty.$$

Let  $T$  be a bounded linear operator (b.l.op.) on  $H$ . We define the *norm* of  $T$  by

$$\|T\| = \sup_{0 \neq \phi \in H} \frac{\|T\phi\|}{\|\phi\|}. \quad (1.11)$$

(It follows that  $\|T\phi\| \leq \|T\|\|\phi\| \quad \forall \phi \in H$ ).

A linear operator  $T$  on  $H$  is *continuous* if whenever  $f_n \rightarrow f$  in  $H$  then  $\|Tf_n - Tf\| \rightarrow 0$ . As in the case of bounded linear functionals we can prove (**Exercise**) that a linear operator  $T$  on  $H$  is bounded if and only if it is continuous. As in the case of bounded linear functionals we can show (**Exercise**) that

$$\|T\| = \sup_{0 \neq \phi \in H: \|\phi\| \leq 1} \|T\phi\| = \sup_{\phi \in H: \|\phi\| = 1} \|T\phi\|.$$

Now, let  $T, S, \dots$  be b.l.op.'s on  $H$ . We define their sum  $T + S$  as that (linear) operator  $W$  on  $H$ , such that  $W\phi = T\phi + S\phi, \forall \phi \in H$ . Similarly  $\lambda T = S$ , where  $S\phi = \lambda T\phi$ . It

is easy to see that with these definitions of addition and scalar multiplication, the set of b.l.op's on  $H$  forms a vector space.

**Exercise:** With the *norm* defined by (1.11) the vector space of b.l.op's on  $H$  becomes a *normed linear space*.  $\diamond$

We denote this normed linear space by  $\mathcal{B}(H)$ .

**Exercise:**  $\mathcal{B}(H)$  is a complete normed linear space.  $\diamond$

Now, let  $T, S$  be b.l.op's on  $\mathcal{B}(H)$ . Their product  $TS$  is defined as the function  $D : H \rightarrow H$  which maps the element  $u \in H$  on the element  $T(Su)$ . It is easily seen that  $(TS)u = T(Su)$  defines a linear operator on  $H$ . Moreover  $T(S_1 + S_2) = TS_1 + TS_2$  etc, while in general  $TS \neq ST$ . Since  $\|TSu\| = \|T(Su)\| \leq \|T\|\|Su\| \leq \|T\|\|S\|\|u\|$ , we see that  $\|TS\| \leq \|T\|\|S\|$ , i.e.  $TS \in \mathcal{B}(H)$ .

Referring to the Projection Theorem 1.1, set  $g = Ph$ , where  $g$  is the orthogonal projection (best approximation) of  $h$  on a closed subspace  $G$  of  $H$ .  $P$  is called the *projection operator onto  $G$* .

**Exercise:** Show that

- (i)  $P$  is a linear operator on  $H$ .
- (ii)  $P$  is a bounded linear operator on  $H$ .
- (iii)  $\text{Ran}P = G$ ,  $\text{Ker}P = G^\perp$ ,  $\text{Ran}(I - P) = G^\perp$ ,  $\text{Ker}(I - P) = G$ , where  $I$  is the identity operator  $Iu = u, \forall u \in H$  (obviously  $I \in \mathcal{B}(H)$  with  $\|I\| = 1$ ).
- (iv)  $P^2 = P, \|P\| = 1$ .
- (v)  $I - P$  is the projection operator onto  $G^\perp$ .  $\diamond$

## 1.9 The Lax–Milgram and the Galerkin theorems

Henceforth we will usually consider *real Hilbert spaces*, i.e. complete inner product spaces over the real numbers with  $(\lambda f, \mu g) = \lambda\mu(f, g), \forall \lambda, \mu$  real,  $f, g \in H$ , and  $(f, g) = (g, f), \forall f, g \in H$ .

A (real) *bilinear form* on a real Hilbert space  $H$  is a map from  $H \times H$  into  $\mathbb{R}$  denoted



by  $B(f, g)$  for  $f, g \in H$ , which satisfies:

$$\begin{aligned} B(\lambda_1 f_1 + \lambda_2 f_2, g) &= \lambda_1 B(f_1, g) + \lambda_2 B(f_2, g) \\ B(f, \mu_1 g_1 + \mu_2 g_2) &= \mu_1 B(f, g_1) + \mu_2 B(f, g_2) \end{aligned}$$

for  $f_i, f, g_i, g \in H$ ,  $\mu_i, \lambda_i \in \mathbb{R}$ . In general,  $B(f, g) \neq B(g, f)$ , i.e.  $B$  is not symmetric.

The following theorem will be central in the sequel:

**Theorem 1.3** (Lax–Milgram Theorem). *Let  $H$  be a (real) Hilbert space and let  $B(., .) : H \times H \rightarrow \mathbb{R}$  be a bilinear form on  $H$  which satisfies:*

$$\begin{aligned} \text{(i)} \quad & |B(\phi, \psi)| \leq c_1 \|\phi\| \|\psi\| \quad \forall \phi, \psi \in H \\ \text{(ii)} \quad & B(\phi, \phi) \geq c_2 \|\phi\|^2 \quad \forall \phi \in H, \end{aligned}$$

where  $c_1, c_2$  are positive constants independent of  $\phi, \psi \in H$ .

Let  $F : H \rightarrow \mathbb{R}$  be a given (real valued) bounded linear functional on  $H$ . Then there exists a unique  $u \in H$  satisfying

$$B(u, v) = F(v) \quad \text{for all } v \in H.$$

Moreover,

$$\|u\| \leq \frac{1}{c_2} \|F\|.$$

**Proof.** Let  $\phi \in H$  be fixed. Then  $\Phi : H \rightarrow \mathbb{R}$ , defined for every  $v \in H$  by  $\Phi(v) = B(\phi, v)$ , defines a continuous linear functional on  $H$ . (Linearity follows from the fact that  $B$  is a bilinear form. For boundedness observe that for each  $v \in H$ :

$$|\Phi(v)| = |B(\phi, v)| \leq c_1 \|\phi\| \|v\|.$$

Hence  $\|\Phi\| \leq c_1 \|\phi\| < \infty$ .

By the Riesz Representation Theorem (1.2) therefore, there exists a unique element  $\tilde{\phi} \in H$  such that

$$\Phi(v) = B(\phi, v) = (v, \tilde{\phi}) \quad \text{for every } v \in H. \quad (1.12)$$

Hence for every  $\phi \in H$ , we define a  $\tilde{\phi} \in H$  by (1.12) and denote the correspondence  $\phi \mapsto \tilde{\phi}$  by  $\tilde{\phi} = A\phi$ , i.e.

$$B(\phi, v) = (v, A\phi), \quad \forall \phi \in H, \quad \forall v \in H. \quad (1.13)$$

Now  $A$  is a linear operator defined on  $H$ . To show linearity, observe that, given  $\phi, \psi \in H$  for every  $v \in H$  and  $\lambda, \mu$  real we have that

$$\begin{aligned} (v, A(\lambda\phi + \mu\psi)) &= B(\lambda\phi + \mu\psi, v) = \lambda B(\phi, v) + \mu B(\psi, v) = \\ &= \lambda(v, A\phi) + \mu(v, A\psi) = (v, \lambda A\phi + \mu A\psi). \end{aligned}$$

Hence  $A(\lambda\phi + \mu\psi) = \lambda A\phi + \mu A\psi \iff A$  is linear.

We claim now that  $A$ , defined by (1.13) has a range  $\text{Ran}(A)$  which is a *closed* subspace of  $H$ . It is (easily) a subspace. To show that it is closed, let  $\hat{\phi}_n = A\phi_n$  be a convergent sequence, such that  $\hat{\phi}_n \rightarrow \hat{\phi}$ . Now, since  $B(\phi_n, v) = (v, A\phi_n) \forall v \in H \Rightarrow B(\phi_n - \phi_m, v) = (A\phi_n - A\phi_m, v) \forall v \in H$ . Choose  $\phi_n - \phi_m = v$  and using (ii) get  $\|\phi_n - \phi_m\| \leq \frac{1}{c_2} \|A\phi_n - A\phi_m\|$ . Hence  $\{\phi_n\}$  is a Cauchy sequence in  $H$ , i.e. there exists  $\phi \in H$  such that  $\phi_n \rightarrow \phi$ . We now show that  $\hat{\phi} = A\phi$ , thus showing that  $\hat{\phi} \in \text{Ran}(A)$ , i.e. that  $\text{Ran}(A)$  is closed.

Now  $|B(\phi_n, v) - B(\phi, v)| \leq C_1 \|\phi_n - \phi\| \|v\|$  gives that

$$\lim_{n \rightarrow \infty} B(\phi_n, v) = B(\phi, v) \quad \forall v \in H.$$

Also  $(A\phi_n, v) = (\hat{\phi}_n, v) \rightarrow (\hat{\phi}, v)$  since  $|(\hat{\phi}_n, v) - (\hat{\phi}, v)| \leq \|\hat{\phi}_n - \hat{\phi}\| \|v\|$ . Since  $B(\phi_n, v) = (A\phi_n, v) \forall v \in H \Rightarrow B(\phi, v) = (\hat{\phi}, v) \forall v \in V$ , i.e.  $\hat{\phi} = A\phi$ , by definition of  $A$ . Hence  $\text{Ran}(A)$  is closed. We now claim that  $\text{Ran}(A) = H$ . Suppose that  $\text{Ran}(A)$  is properly included in  $H$ , so that  $\exists z \neq 0 \in (\text{Ran}(A))^\perp$ . Hence  $(z, v) = 0 \forall v \in \text{Ran}(A)$ . In particular  $\forall \phi \in H$ ,  $B(\phi, z) = (A\phi, z) = 0$ . Hence for  $\phi = z$ ,  $0 = B(z, z) \geq c_2 \|z\|^2 \Rightarrow z = 0$  (contradiction). So  $\text{Ran}(A) = H$ .

Now, given  $F$ , a b.l.f. on  $H$ , by Riesz representation,  $\exists! \chi \in H$  such that  $F(v) = (\chi, v) \forall v \in H$ . Since  $\text{Ran}(A) = H$ ,  $\exists u \in H$  such that  $Au = \chi$ . Hence  $\exists u$  such that

$$F(v) = (Au, v) = B(u, v) \quad \forall v \in H$$

and we have existence of  $u$  as claimed in the statement of the theorem.

For uniqueness, suppose that  $\exists u_1 \neq u_2$  such that  $B(u_1, v) = F(v) = B(u_2, v) \forall v \in H$ . Hence

$$B(u_1 - u_2, v) = 0 \quad \forall v \in H \Rightarrow 0 = B(u_1 - u_2, u_1 - u_2) \geq c_2 \|u_1 - u_2\|^2 \Rightarrow u_1 = u_2.$$

Finally since  $B(u, u) = F(u)$ , (i), (ii) give that ( $u \neq 0$ )  $c_2 \|u\|^2 \leq |F(u)|$ , from which  $\|u\| \leq \frac{1}{c_2} \frac{|F(u)|}{\|u\|}$ . Hence

$$\|u\| \leq \sup_{v \neq 0} \frac{1}{c_2} \frac{|F(v)|}{\|v\|} = \frac{1}{c_2} \|F\|.$$

□

We finally present a basic theorem for the *Galerkin approximation* (see below for definition)  $u_h$  to the solution  $u$  of  $B(u, v) = F(v)$  guaranteed by the Lax–Milgram theorem.

**Theorem 1.4** (Galerkin). *Let  $H$  be a real Hilbert space and let  $B(.,.) : H \times H \rightarrow \mathbb{R}$  be a bilinear form on  $H$  which satisfies:*

$$(i) \quad |B(\phi, \psi)| \leq c_1 \|\phi\| \|\psi\| \quad \forall \phi, \psi \in H,$$

$$(ii) \quad B(\phi, \phi) \geq c_2 \|\phi\|^2 \quad \forall \phi \in H,$$

for some positive constants  $c_1, c_2$  independent of  $\phi, \psi \in H$ . Let  $F$  be a given real-valued b.l.f. on  $H$  and let  $u$  be the unique element of  $H$ , guaranteed by the Lax–Milgram theorem, satisfying  $B(u, v) = F(v), \forall v \in H$ .

Let  $\{S_h\}$  for  $0 < h \leq 1$  be a family of finite-dimensional subspaces of  $H$ . For every  $h$  there exists a unique  $u_h$  such that

$$B(u_h, v_h) = F(v_h) \quad \forall v_h \in S_h. \tag{1.14}$$

We call  $u_h$  the Galerkin approximation of  $u$  in  $S_h$ .

Moreover we have the error estimate

$$\|u - u_h\| \leq \frac{c_1}{c_2} \inf_{\chi \in S_h} \|u - \chi\|.$$

**Proof.** The existence–uniqueness of  $u_h \in S_h$  is guaranteed by the Lax–Milgram theorem applied to the Hilbert space  $(S_h, \|\cdot\|)$ . Alternatively, let  $\{\phi_j\}_{j=1}^m$  be a basis for  $S_h$ , where  $m = m(h) = \dim S_h$ , and try to find  $u_h \in S_h$  in the form  $u_h = \sum_{j=1}^m c_j \phi_j$ . By (1.14)  $u_h$  satisfies

$$\begin{aligned} B(u_h, \phi_i) &= F(\phi_i) \quad 1 \leq i \leq m, \quad \text{i.e.} \\ B\left(\sum_{j=1}^m c_j \phi_j, \phi_i\right) &= F(\phi_i) \quad 1 \leq i \leq m \implies \sum_{j=1}^m c_j B(\phi_j, \phi_i) = F(\phi_i), \quad 1 \leq i \leq m. \end{aligned}$$

Hence if  $A$  is the  $m \times m$  matrix given by  $A_{ij} = B(\phi_j, \phi_i)$ ,  $1 \leq i, j \leq m$ , the  $c_j$ 's are the solution of the linear system

$$\sum_{j=1}^m A_{ij} c_j = F(\phi_i), \quad 1 \leq i \leq m. \quad (1.15)$$

The associated homogeneous system  $\sum_{j=1}^m A_{ij} \tilde{c}_j = 0$ ,  $1 \leq i \leq m$ , has only the zero solution. (Since  $\sum_{j=1}^m A_{ij} \tilde{c}_j = 0 \Rightarrow B(\sum_{j=1}^m \tilde{c}_j \phi_j, \phi_i) = 0$ ,  $1 \leq i \leq m$ ,  $\Rightarrow B(v_h, v_h) = 0$ , where  $v_h = \sum_{j=1}^m \tilde{c}_j \phi_j$ . Hence, by (ii)  $v_h = 0 \Rightarrow \tilde{c}_i = 0$ .) Therefore  $A$  is invertible and (1.15) has a unique solution, i.e. (1.14) has a unique solution  $u_h \in S_h$ . (**Exercise:** Show that  $A$  is positive definite.)

For the error estimate observe that by (ii)

$$c_2 \|u - u_h\|^2 \leq B(u - u_h, u - u_h) = B(u - u_h, u) \quad (1.16)$$

(since  $B(u_h, \psi) = F(\psi) = B(u, \psi) \quad \forall \psi \in S_h \Rightarrow B(u - u_h, \psi) = 0 \quad \forall \psi \in S_h$ ). For the same reason, for any  $\chi \in S_h$

$$B(u - u_h, u) = B(u - u_h, u - \chi) \leq c_1 \|u - u_h\| \|u - \chi\|,$$

using (i).

By (1.16) we conclude therefore that  $c_2 \|u - u_h\|^2 \leq c_1 \|u - u_h\| \|u - \chi\|$ , i.e.

$$\begin{aligned} \|u - u_h\| &\leq \frac{c_1}{c_2} \|\chi - u\| \quad \forall \chi \in S_h, \quad \text{i.e.} \\ \|u - u_h\| &\leq \frac{c_1}{c_2} \inf_{\chi \in S_h} \|\chi - u\| = \frac{c_1}{c_2} \|P_h u - u\|, \end{aligned}$$

where  $P_h$  is the projection operator on  $S_h$ . □

Here is an immediate corollary to Galerkin's Theorem 1.4.

**Corollary 1.1.** *With notation introduced in Theorem 1.4, suppose that the family  $S_h$  of subspaces satisfies*

$$\lim_{h \rightarrow 0} \inf_{\chi \in S_h} \|u - \chi\| = 0.$$

Then  $\lim_{h \rightarrow 0} \|u - u_h\| = 0$ . □

Finally we mention that in the case of a *symmetric, bilinear form*  $B$ , i.e. when (in addition to (i), (ii) of Theorem 1.3)

$$(iii) \quad B(u, v) = B(v, u) \quad \forall u, v \in H$$

we can obtain a *variational formulation* of the problem of finding  $u \in H$  such that

$$B(u, v) = F(v) \quad \forall v \in H,$$

where  $F$  is a bounded linear functional on  $H$ .

For  $v \in H$  consider the following (nonlinear) functional  $J : H \rightarrow \mathbb{R}$  defined by

$$J(v) = \frac{1}{2} B(v, v) - F(v) \tag{1.17}$$

and the associated problem of finding  $z \in H$  such that

$$J(z) = \min_{v \in H} J(v). \tag{1.18}$$

We have the following theorem:

**Theorem 1.5** (Rayleigh–Ritz). *Suppose  $B$  is a symmetric, bilinear form which satisfies the hypotheses (i), (ii) of Theorem 1.3. Then, the problem (1.18) of minimizing over  $H$  the functional  $J$  defined by (1.17) has a unique solution which coincides with  $u$ , the existence–uniqueness of which was guaranteed by the Lax–Milgram Theorem 1.3.*

**Proof.** Let  $u$  be the solution of the problem  $B(u, v) = F(v) \quad \forall v \in H$ . Then  $\forall w \in H$ :

$$\begin{aligned} J(u + w) &= \frac{1}{2} B(u + w, u + w) - F(u + w) = \text{(due to the symmetry of } B) = \\ &= \left( \frac{1}{2} B(u, u) - F(u) \right) + \left( \frac{1}{2} B(w, w) \right) + (B(u, w) - F(w)) \\ &= J(u) + \frac{1}{2} B(w, w), \quad \text{since } B(u, w) = F(w) \text{ by Theorem 1.3.} \end{aligned}$$

Hence

$$J(u + w) = J(u) + \frac{1}{2} B(w, w) \geq J(u) + \frac{c_2}{2} \|w\|^2 \quad \text{by (ii).}$$

Therefore  $\forall w \in H, w \neq 0$ :  $J(u + w) > J(u)$  i.e.

$$J(u) = \min_{v \in H} J(v) \quad \text{and } J(v) > J(u) \text{ if } v \neq u.$$

□

Immediately, we have the following corollary, which is the analog of Theorem 1.4.

**Corollary 1.2** (Rayleigh–Ritz, Galerkin). *With notation introduced in Theorem 1.4 and the additional hypothesis of symmetry of  $B$ , the problem of minimizing the functional  $J$  defined by (1.17), over  $S_h$ , i.e. finding  $u_h \in S_h$  such that*

$$J(u_h) = \min_{\chi \in S_h} J(\chi)$$

*has a unique solution  $u_h$ , which coincides with the Galerkin approximation in  $S_h$  of  $u$ , constructed in Theorem 1.4.  $\square$*

# Chapter 2

## Elements of the Theory of Sobolev Spaces and Variational Formulation of Boundary–Value Problems in One Dimension

This chapter (and chapter 4) follows closely the analogous material in H. Brezis, *Analyse fonctionnelle, théorie et applications*, Masson, Paris, 1983. (For the translation in Greek and the new English edition, see the References)

### 2.1 Motivation

We consider the following “two–point” boundary–value problem in one dimension. Find a real–valued function  $u(x)$ , defined for  $x \in [a, b]$  and satisfying

$$(*) \begin{cases} -(p(x)u'(x))' + q(x)u(x) = f(x), & a \leq x \leq b. \\ u(a) = u(b) = 0. \end{cases}$$

Here  $p(x), q(x), f(x)$  are real–valued functions defined on  $[a, b]$  such that  $p \in C^1([a, b])$ ,  $p(x) \geq \alpha > 0$  for  $x \in [a, b]$ ,  $q \in C([a, b])$ ,  $q(x) \geq 0 \forall x \in [a, b]$ ,  $f \in C([a, b])$ . A *classical* (or *strong*) solution of the boundary–value problem (b.v.p.)  $(*)$  is a function  $u$  of class  $C^2([a, b])$  which satisfies  $(*)$  in the usual sense.

If we multiply the equation in (\*) by a function  $\phi \in C^1([a, b])$ , such that  $\phi(a) = \phi(b) = 0$  and integrate by parts we obtain

$$(**) \int_a^b pu'\phi' dx + \int_a^b qu\phi dx = \int_a^b f\phi dx, \quad \forall \phi \in C^1([a, b]), \phi(a) = \phi(b) = 0.$$

Note that (\*\*) makes sense for  $u \in C^1([a, b])$  e.g. (as opposed to (\*) which requires  $u \in C^2([a, b])$ ). In fact (\*\*) just requires that  $u, u'$  be integrable functions. One may say (vaguely) that a solution  $u \in C^1([a, b])$  (such that  $u(a) = u(b) = 0$ ) of (\*\*) is (one kind of) a *weak* or *generalized* solution of (\*).

The *variational method* for solving (i.e. proving existence and uniqueness of solutions) problems such as (\*) – and also boundary–value problems for *partial differential equations* proceeds roughly as follows:

- (i) We define precisely what we mean by a *weak solution* of (\*). Typically it will be the solution of a *weak* (or *variational*) form of the problem (\*), such as (\*\*), or, equivalently the solution of an appropriate minimization problem. Here the *Sobolev spaces* will play a central role.
- (ii) We show existence and uniqueness of the weak solution, for example by the *Lax–Milgram theorem*; note that (\*\*) suggests the variational problem  $B(u, \phi) \equiv \int_a^b (pu'\phi' + qu\phi) = F(\phi) \equiv \int_a^b f\phi, \quad \forall \phi \in C^1([a, b]),$  such that  $\phi(a) = \phi(b) = 0$ .
- (iii) We then prove that the weak solution is sufficiently *regular*. For example here we must prove that the weak solution is in  $C^2([a, b])$ .
- (iv) We finally prove that a *weak solution*, which is in  $C^2([a, b])$ , is a *strong* (*classical*) solution of (\*).

Note that a weak formulation of the problem provides us also with a method (Galerkin) for approximating its weak solution in a suitably chosen finite–dimensional subspace of functions with good approximation properties, that are also suitable for numerical computations.



## 2.2 Notation and preliminaries

We now introduce some notation on function spaces that will be used in the sequel and collect (without proof) some useful facts about  $L^2$ .

We let  $\Omega$  denote an open subset of  $\mathbb{R}^N$ ; for this part of the notes take, for example,  $\Omega$  to be an open interval  $(a, b)$  in  $\mathbb{R}$ . For simplicity we shall consider only *real-valued functions* defined on  $\Omega$  or  $\bar{\Omega}$ . We define:

$C(\Omega)$  = space of continuous functions on  $\Omega$ .

$C^k(\Omega)$  = space of  $k$ -times differentiable functions on  $\Omega$ , i.e the space of those functions  $f(x)$ ,  $x \in \Omega$  such that  $\frac{\partial^{\alpha_1+\dots+\alpha_N} f(x)}{\partial x_1^{\alpha_1} \dots \partial x_N^{\alpha_N}}$  are continuous functions on  $\Omega$  for all integers  $0 \leq \alpha_i \leq k$ ,  $1 \leq i \leq N$  such that  $\alpha_1 + \alpha_2 + \dots + \alpha_N \leq k$ . Put  $C^0(\Omega) \equiv C(\Omega)$ .

$C^\infty(\Omega) = \bigcap_{k \geq 0} C^k(\Omega)$ .

$C_c(\Omega)$  = space of functions in  $C(\Omega)$  whose support is a compact set included in  $\Omega$ . (If  $f \in C(\Omega)$ , support of  $f = \text{supp} f = \overline{\{x \in \Omega : f(x) \neq 0\}}$ ). Hence these functions vanish outside a compact set (strictly) included in  $\Omega$ .

$C_c^k(\Omega) = C^k(\Omega) \cap C_c(\Omega)$ .

$C_c^\infty(\Omega) = C^\infty(\Omega) \cap C_c(\Omega)$ . Often the notation  $C_0^\infty(\Omega)$  is used instead of  $C_c^\infty(\Omega)$ .

We recall a few facts about the spaces  $L^p(\Omega)$ . Let  $dx$  denote the Lebesgue measure in  $\mathbb{R}^n$ . By  $L^1(\Omega)$  we denote the space of (Lebesgue) integrable functions  $f$  on  $\Omega$ , i.e. the functions for which

$$\|f\|_{L^1} = \|f\|_{L^1(\Omega)} = \int_{\Omega} |f(x)| dx < \infty.$$

(We denote usually  $\int_{\Omega} f = \int_{\Omega} f(x) dx$ ).

Let  $1 \leq p < \infty$ . Then

$$L^p(\Omega) = \{f : \Omega \rightarrow \mathbb{R} ; |f|^p \in L^1(\Omega)\}.$$

We put

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

For  $p = \infty$  we define  $L^\infty(\Omega) = \{f : \Omega \rightarrow \mathbb{R}, f \text{ measurable such that there exists a constant } C < \infty \text{ such that } |f(x)| \leq C \text{ a.e. (almost everywhere) in } \Omega, \text{ i.e. such that } |f(x)| \leq C \text{ for all } x \in \Omega \text{ except possibly for some } x \text{ belonging to a subset of } \Omega \text{ of Lebesgue measure zero}\}$ . We put

$$\|f\|_{L^\infty} = \|f\|_{L^\infty(\Omega)} = \inf\{C : |f(x)| \leq C \text{ a.e. in } \Omega\}$$

and note that  $|f(x)| \leq \|f\|_{L^\infty}$  for every  $x \in \Omega - O$ , where  $O$  has measure 0. (The quantities  $\|f\|_{L^p}$ ,  $1 \leq p \leq \infty$  are *norms* on the respective spaces  $L^p$ . We have already introduced the Hilbert space  $L^2 = L^2(\Omega)$ ).

We shall mainly be concerned with  $L^2(\Omega)$ . We would like to emphasize that the “functions”  $f(x) \in L^2(\Omega)$  are not really functions but *equivalence classes* of functions, where the equivalence  $f \sim g$  holds if and only if  $f(x) = g(x)$  a.e. in  $\Omega$ . For example, when we say that  $f = 0$  as an element of  $L^2(\Omega)$ , we mean that  $f(x) = 0$  for all  $x$  outside a set of measure zero in  $\Omega$ , i.e. that  $f(x) = 0$  a.e. in  $\Omega$ ; contrast with the situation  $f \in C(\Omega)$ ,  $f = 0 \Rightarrow f(x) = 0 \forall x \in \Omega$ .

The following results (see Brezis for proofs) will be used in sequel. ( $L^1_{loc}(\Omega)$  will denote the functions  $f$  on  $\Omega$  for which  $\int_K |f(x)| dx < \infty$  for every *compact* set  $K \subset \Omega$ . For example  $f(x) = \frac{1}{x} \in L^1_{loc}((0, 1))$  but  $f \notin L^1((0, 1))$ ).

**Lemma 2.1.** *If  $f \in L^1_{loc}(\Omega)$  such that*

$$\int_{\Omega} f u = 0 \quad \forall u \in C_c(\Omega),$$

*then  $f = 0$  a.e. in  $\Omega$ .*

**Lemma 2.2.** *The space  $C_c(\Omega)$  is dense in  $L^2(\Omega)$ , i.e.*

$$\forall f \in L^2(\Omega), \forall \epsilon > 0 \exists \tilde{f} \in C_c(\Omega) : \|f - \tilde{f}\|_{L^2} < \epsilon.$$

**Definition 2.1.** *A regularizing sequence (or a sequence of mollifiers) is a sequence of functions  $\{\rho_n\}$ ,  $n = 1, 2, \dots$  such that:*

$$\rho_n \in C_c^\infty(\mathbb{R}^N),$$

$$\rho_n \geq 0 \text{ on } \mathbb{R}^N,$$

$$\text{supp } \rho_n \subset B(0, \frac{1}{n}) \equiv \{x \in \mathbb{R}^N : |x| \equiv \left(\sum_{i=1}^N x_i^2\right)^{\frac{1}{2}} \leq \frac{1}{n}\},$$

$$\int_{\mathbb{R}^N} \rho_n dx = 1.$$

Such functions clearly exist. E.g. in  $\mathbb{R}$ , let

$$\rho(x) = \begin{cases} e^{\frac{1}{x^2-1}} & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1. \end{cases}$$

Clearly  $\rho(x)$  is continuous on  $\mathbb{R}$  and

$$\rho^{(j)}(x) = \frac{\Pi_j(x)}{(1-x^2)^{2j}} e^{-\frac{1}{1-x^2}}, \quad |x| < 1,$$

where  $\Pi_j(x)$  are polynomials. Since  $y^j e^{-y} \rightarrow 0$  as  $y \rightarrow +\infty$  we see that  $\rho(x) \in C_c^\infty(\mathbb{R})$ .

Of course  $\text{supp} \rho = [-1, 1]$ . Let

$$C = \left( \int_{-\infty}^{\infty} \rho(x) dx \right)^{-1}.$$

Define

$$\rho_n(x) = C n \rho(nx).$$

Then

$$\rho_n \in C_c^\infty(\mathbb{R}), \quad \rho_n(x) \geq 0 \text{ on } \mathbb{R}, \quad \text{supp} \rho_n = \left[-\frac{1}{n}, \frac{1}{n}\right], \quad \int_{-\infty}^{\infty} \rho_n dx = 1.$$

In  $\mathbb{R}^N$  define

$$\rho(x) = \begin{cases} e^{\frac{1}{|x|^2-1}} & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1. \end{cases}$$

(where  $|x| = (\sum_{i=1}^N x_i^2)^{1/2}$ ) and

$$\rho_n(x) = C n^N \rho(nx), \quad C = \left( \int_{\mathbb{R}^N} \rho(x) dx \right)^{-1}.$$

Denote the *convolution* of two functions  $f(x), g(x)$  defined on  $\mathbb{R}^N$  as the function

$$(f * g)(x) = \int_{\mathbb{R}^N} f(x-y) g(y) dy$$

(provided the integrals exist).

**Lemma 2.3.** (i) Let  $f \in C_c(\Omega)$ . Extend  $f$  by zero on the whole of  $\mathbb{R}^N$ . Then, for sufficiently large  $n$ ,

$$\rho_n * f \in C_c^\infty(\Omega)$$

$$\text{and} \quad \sup_{x \in \Omega} |f(x) - (\rho_n * f)(x)| \rightarrow 0, \quad n \rightarrow \infty.$$

(ii) Let  $f \in C(\mathbb{R}^N)$ . Then  $\rho_n * f \in C^\infty(\mathbb{R}^N)$  and  $\rho_n * f \rightarrow f$  uniformly, on every compact set  $K \subset \mathbb{R}^N$ , i.e.

$$\sup_{x \in K} |f(x) - (\rho_n * f)(x)| \rightarrow 0, \quad n \rightarrow \infty$$

$\forall K$  compact  $\subset \mathbb{R}^N$ .

(iii) Let  $f \in L^2(\mathbb{R}^N)$ . Then  $\rho_n * f \in C^\infty(\mathbb{R}^N) \cap L^2(\mathbb{R}^N)$  and

$$\|\rho_n * f - f\|_{L^2(\mathbb{R}^N)} \rightarrow 0, \quad n \rightarrow \infty.$$

Finally we mention:

**Lemma 2.4.**  $C_c^\infty(\Omega)$  is dense in  $L^2(\Omega)$ , i.e.

$$\forall f \in L^2(\Omega), \forall \epsilon > 0 \exists \tilde{f} \in C_c^\infty(\Omega) : \|f - \tilde{f}\|_{L^2} < \epsilon.$$

## 2.3 The Sobolev space $H^1(I)$

Let  $I = (a, b)$  be an open interval in  $\mathbb{R}$ . (We will mainly use a bounded interval  $(a, b)$  in the applications but here we may suppose that  $I$  could be unbounded, i.e. that possibly  $a = -\infty$  and/or  $b = \infty$ ).

**Definition 2.2.** The Sobolev space  $H^1(I)$  is defined by

$$H^1(I) = \{u \in L^2(I) : \exists g \in L^2(I) \text{ such that } \int_I u\phi' = - \int_I g\phi, \forall \phi \in C_c^1(I)\}.$$

For  $u \in H^1(I)$  we denote  $g = u'$  and call  $g$  the weak (generalized) derivative of  $u$  (in the  $L^2$  sense).

**Remarks.**

- (i) When there is no reason for confusion we shall denote  $H^1 = H^1(I)$ ,  $L^2 = L^2(I)$ , etc.
- (ii) It is clear that the generalized derivative  $g$  in the above definition is *unique*. For suppose  $\exists g_1, g_2 \in L^2(I)$  such that  $\int_I (g_1 - g_2)\phi = 0, \forall \phi \in C_c^1(I)$ . Since  $C_c^\infty(I) \subset C_c^1(I) \subset L^2(I)$  and since (by Lemma 2.4)  $C_c^\infty(I)$  is dense in  $L^2(I)$ ,

it follows that  $C_c^1(I)$  is dense in  $L^2(I)$ . It follows that  $g_1 - g_2 = 0$  in  $L^2(I)$ . (N.B. In general in a Hilbert space  $H$ , where  $D \subset H$  dense in  $H$ , we prove that  $(g, \phi) = 0 \forall \phi \in D \Rightarrow g = 0$ , since  $\exists \phi_i \in D$  such that  $\phi_i \rightarrow g$ ,  $i \rightarrow \infty$  in  $H$ . Therefore  $0 = (g, \phi_i) \rightarrow (g, g) \Rightarrow g = 0$ ). We emphasize again that  $g_1 = g_2$  in  $L^2$  means that  $g_1(x) = g_2(x)$  a.e. in  $I$ .

- (iii) The functions  $\phi \in C_c^1(I)$  in the definition of  $g$  are called *test functions*. One could take  $C_c^\infty(I)$  to be the set of test functions instead of  $C_c^1(I)$ . (The only thing to show is that if  $\int_I u\phi' = -\int_I g\phi$ , for  $u, g \in L^2(I)$ , holds for every  $\phi \in C_c^\infty(I)$ , then it will hold for every  $\phi \in C_c^1(I)$ . This follows from the Cauchy–Schwarz inequality and the facts that  $\phi \in C_c^1(I) \Rightarrow \rho_n * \phi \in C_c^\infty(I)$  and  $\rho_n * \phi \rightarrow \phi$ , e.g. in  $L^2(I)$ , and also that  $(\rho_n * \phi)' = \rho_n * \phi' \in C_c^\infty(I)$  and  $\rho_n * \phi' \rightarrow \phi'$  in  $L^2(I)$ ).
- (iv) It is clear that if  $u \in C^1(I) \cap L^2(I)$  and if the (classical) derivative  $u'$  of  $u$  belongs to  $L^2(I)$ , then integration by parts gives that  $\int_I u\phi' = -\int_I u'\phi \forall \phi \in C_c^1(I)$ , i.e. that  $u'$  is the weak derivative of  $u$ , i.e. that  $u \in H^1(I)$ . Of course, if  $I$  is bounded, then  $u \in C^1(\bar{I}) \Rightarrow u, u' \in L^2(I)$  and we have  $C^1(\bar{I}) \subset H^1(I)$ .
- (v) There are other ways of defining the Sobolev space  $H^1$ . Using e.g. the *theory of distributions* we may conclude that every  $u \in L^2(I)$  has a distributional derivative  $u'$ . We say that  $u \in H^1(I)$  if  $u'$  coincides as a distribution with a function  $u' \in L^2(I)$ . If  $I = \mathbb{R}$  we may also define  $H^1$  using Fourier transforms.

### Examples.

- (i) Consider  $u(x) = |x|$  on  $I = (-1, 1)$ . Clearly  $u \in C(\bar{I})$ ,  $u \in L^2(I)$ , but  $u$  fails to have a classical derivative at  $x = 0$ . Consider the function

$$g(x) = \begin{cases} -1 & \text{if } -1 < x \leq 0 \\ 1 & \text{if } 0 < x < 1. \end{cases}$$

Clearly,  $g \in L^2(I)$ . In addition for each  $\phi \in C_c^1(I)$ ,

$$\begin{aligned}
-\int_{-1}^1 g(x)\phi(x) dx &= -\int_{-1}^0 (-1)\phi(x) dx - \int_0^1 1\phi(x) dx \\
&= -\int_{-1}^0 \phi(x) d(-x) - \int_0^1 \phi(x) dx = -[(-x)\phi(x)]_{-1}^0 \\
&\quad - [x\phi(x)]_0^1 + \int_{-1}^0 (-x)\phi'(x) dx + \int_0^1 x\phi'(x) dx \\
&= \int_{-1}^1 |x|\phi'(x) dx = \int_{-1}^1 u(x)\phi'(x) dx.
\end{aligned}$$

It follows that  $u(x) = |x| \in H^1((-1, 1))$  and  $u' = g$  is the weak derivative of  $u$ .

(ii) More generally, if  $I$  is a bounded interval and  $u \in C(\bar{I})$  with  $u'$  (classical derivative) piecewise continuous on  $\bar{I}$  (as would be the case e.g. if  $u$  is a piecewise polynomial, continuous function on  $\bar{I}$ ), then  $u \in H^1(I)$  and its weak derivative coincides with the classical derivative a.e. in  $I$ .

(iii) As in (i) the function

$$u(x) = \frac{1}{2}(|x| + x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

on  $I = (-1, 1)$  belongs to  $H^1$  and its weak derivative  $u'$  is the function

$$H(x) = \begin{cases} 0 & \text{if } -1 < x \leq 0 \\ 1 & \text{if } 0 < x < 1, \end{cases}$$

which is called *Heaviside's function*. Clearly  $H \in L^2(I)$ . Does  $H$  belong to  $H^1(I)$ ? The answer is no: Suppose that  $H \in H^1(I)$ . Then there must exist  $v(x) \in L^2(I)$  such that  $\int_I H\phi' = -\int_I v\phi$ ,  $\forall \phi \in C_c^1(I)$ , i.e. a  $v \in L^2(I)$  such that  $\int_{-1}^1 v\phi = -\int_{-1}^1 H\phi' = -\int_0^1 \phi' = -\phi(1) + \phi(0) = \phi(0) \forall \phi \in C_c^1(I)$ . Take then such a  $\phi$  with support in the interval  $(-1, 0)$ . It follows that  $\int_{-1}^0 v\phi = \int_{-1}^1 v(x)\phi(x) = 0 \forall \phi \in C_c^1((-1, 0))$ . Since  $C_c^1((-1, 0))$  is dense in  $L^2((-1, 0))$ , as in Remark (ii), p. 30, we see that  $v(x) = 0$  a.e. in  $(-1, 0)$ . Analogously, taking  $\phi \in C_c^1(0, 1)$  we prove that  $v(x) = 0$  a.e. in  $(0, 1)$ . We conclude therefore that  $v = 0$  a.e. in  $(-1, 1)$ . But that would contradict  $\int_{-1}^1 v\phi = \phi(0) \forall \phi \in C_c^1(I)$ . (Of course, as a *distribution*,  $H(x)$  has a distributional derivative which coincides with the  $\delta$ -“function”,  $H' = \delta_0$ . We just proved that  $\delta_0 \notin L^2(I)$ ).

It is clear that  $H^1(I)$  is a linear subspace of  $L^2(I)$ , since if  $u, v \in H^1(I)$  and  $u', v'$  are their weak derivatives, then  $\lambda u' + \mu v'$  is the weak derivative of  $\lambda u + \mu v$ . Hence  $(\lambda u + \mu v) \in H^1(I)$ , and  $(\lambda u + \mu v)' = \lambda u' + \mu v'$  for  $\lambda, \mu \in \mathbb{R}$ . We denote by  $(\cdot, \cdot)$ ,  $\|\cdot\|$ , respectively, the inner product, norm of  $L^2 = L^2(I)$ , i.e. we let, for  $u, v \in L^2(I)$ ,  $(u, v) = \int_I u(x)v(x) dx$ ,  $\|u\| = (u, u)^{\frac{1}{2}}$ . Then, for  $u, v \in H^1(I)$  we define

$$\begin{aligned}(u, v)_1 &\equiv (u, v) + (u', v'), \\ \|u\|_1 &\equiv (\|u\|^2 + \|u'\|^2)^{\frac{1}{2}} = (u, u)_1^{\frac{1}{2}}.\end{aligned}$$

It is clear that  $(\cdot, \cdot)_1$  defines an *inner product* on  $H^1 = H^1(I)$  and  $\|\cdot\|_1$  the induced *norm* on  $H^1(I)$ . (To be precise, sometimes we shall denote  $\|\cdot\|_1 = \|\cdot\|_{H^1(I)}$  etc.). Hence  $H^1(I)$  becomes an inner product space.

**Theorem 2.1.** *The space  $(H^1, \|\cdot\|_1)$  is a Hilbert space.*

**Proof.** We only need to show that  $H^1(I)$  is complete in the norm  $\|\cdot\|_1$ . Let  $\{u_n\}_{n=1,2,\dots} \in H^1(I)$  be a Cauchy sequence in the norm  $\|\cdot\|_1$ , i.e. let

$$\lim_{m,n \rightarrow \infty} \|u_m - u_n\|_1 = 0.$$

By the definition of  $\|\cdot\|_1$  it follows that  $\{u_n\}$  and  $\{u'_n\}$  are Cauchy sequences in  $L^2$ . Since  $L^2$  is complete it follows that  $\exists u, g \in L^2(I)$  such that  $u_n \rightarrow u$  in  $L^2$ ,  $u'_n \rightarrow g$  in  $L^2$ . Now, by definition,  $(u_n, \phi') = -(u'_n, \phi) \forall \phi \in C_c^1(I)$  for  $n = 1, 2, 3, \dots$

Since  $\forall \phi \in C_c^1(I)$

$$|(u_n, \phi') - (u, \phi')| \leq \|u_n - u\| \|\phi'\| \rightarrow 0, \quad n \rightarrow \infty$$

and

$$|(u'_n, \phi) - (g, \phi)| \leq \|u'_n - g\| \|\phi\| \rightarrow 0, \quad n \rightarrow \infty,$$

it follows that  $(u, \phi') = -(g, \phi) \forall \phi \in C_c^1(I)$ , i.e. that  $u \in H^1(I)$  and  $u' = g$ .

It remains to show that  $u_n \rightarrow u$  as  $n \rightarrow \infty$  in  $H^1$ . But this follows from

$$\|u_n - u\|_1^2 = \|u_n - u\|^2 + \|u'_n - u'\|^2 = \|u_n - u\|^2 + \|u'_n - g\|^2 \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Hence  $\exists u \in H^1(I)$  such that  $u_n \rightarrow u$  in  $H^1$ , i.e.  $H^1(I)$  is complete.  $\square$

**Remark:** Consider the map  $T : H^1 \rightarrow L^2 \times L^2$  given by  $Tu = [u, u']$ ,  $u \in H^1$ . Equipping  $L^2 \times L^2$  with the norm  $((u, u) + (v, v))^{1/2}$  we see that  $T$  is an isometry of  $H^1$  onto a closed subspace of  $L^2 \times L^2$ . It follows that  $H^1$  is *separable*, since  $L^2$  is.

The following theorem will be very important in sequel:

**Theorem 2.2.** *If  $u \in H^1(I)$ , then  $\exists \tilde{u} \in C(\bar{I})$  such that  $u = \tilde{u}$  a.e. in  $I$  and*

$$\tilde{u}(x) - \tilde{u}(y) = \int_y^x u'(t) dt, \quad \forall x, y \in \bar{I}.$$

Before proving the theorem we make some comments on its content. Note first that if  $u \in H^1(I)$  and  $u = v$  a.e. on  $I$ , then  $v \in H^1(I)$ . Then Theorem 2.2 tells us that in the equivalence class of an element  $u \in H^1(I)$  there is one (and *only* one since  $u, v \in C(\bar{I})$ ,  $u = v$  a.e. on  $I \Rightarrow u(x) = v(x) \forall x \in I$ ) continuous “representative” of  $u$ , denoted in the theorem by  $\tilde{u}$ . Hence, when there is need to do so, we shall use instead of  $u$  its continuous representative  $\tilde{u}$ . For example as the *value*  $u(x)$  for some  $x \in \bar{I}$  (not well-defined if  $u \in L^2$ ) we mean the *value* of  $\tilde{u}$  at that  $x$ . Sometimes we shall replace  $u$  by  $\tilde{u}$  with no special mention or by just noting that  $u$  is continuous, “upon modification on a set of measure zero in  $I$ ”. We emphasize that the statement “ $\exists \tilde{u} \in C(\bar{I})$  such that  $u = \tilde{u}$  a.e. in  $I$ ” is different from the statement that “ $u$  is continuous a.e. in  $I$ ”.

For the proof of Theorem 2.2 we shall need two lemmata.

**Lemma 2.5.** *Let  $f \in L^1_{loc}(I)$  such that*

$$\int_I f \phi' = 0 \quad \forall \phi \in C_c^1(I).$$

*Then, there exists a constant  $C$  such that  $f = C$  a.e. in  $I$ .*

**Proof.** Let  $\psi$  be a fixed function in  $C_c(I)$  such that  $\int_I \psi = 1$ . We shall show that, given  $w \in C_c(I)$ , there exists  $\phi \in C_c^1(I)$ , such that  $\phi' = w - (\int_I w) \psi$ . Indeed, given  $w \in C_c(I)$ , consider  $h(x) = w(x) - (\int_I w) \psi(x)$ . Clearly  $h \in C_c(I)$ . Put  $\phi(x) = \int_a^x h(x) dx$ . Let  $\text{supp} h \subset [c, d] \subset (a, b) = I$ . Clearly, for  $a < y < c$ ,  $\phi(y) = \int_a^y h(x) dx = 0$ , and for  $d < z < b$

$$\begin{aligned} \phi(z) &= \int_a^z h(x) dx = \int_a^b h(x) dx = \int_I h = \int_I \left( w - \left( \int_I w \right) \psi \right) \\ &= \int_I w - \left( \int_I w \right) \left( \int_I \psi \right) = \int_I w - \int_I w = 0. \end{aligned}$$



It follows that  $\phi \in C_c(I)$ . Also  $\phi'(x) = h(x) \in C_c(I)$ , i.e.  $\phi \in C_c(I)$ , and  $\phi' = h = w - (\int_I w)\psi$ . Now, by hypothesis,  $\int_I f\phi' = 0 \forall \phi \in C_c^1(I)$ . In particular, for each  $w \in C_c(I)$ ,

$$\begin{aligned} \int_I f \left( w - \left( \int_I w \right) \psi \right) &= 0 \Rightarrow \int_I fw - \int_I f\psi \int_I w = 0 \Rightarrow \int_I fw - \int_I \left( \int_I f\psi \right) w = 0 \\ &\Rightarrow \int_I \left( f - \int_I f\psi \right) w = 0. \end{aligned}$$

By Lemma 2.1 we conclude that  $f(x) = \int_I f\psi$  a.e. on  $I$ . i.e.  $f(x) = C \equiv \int_I f\psi$  a.e. on  $I$ .  $\square$

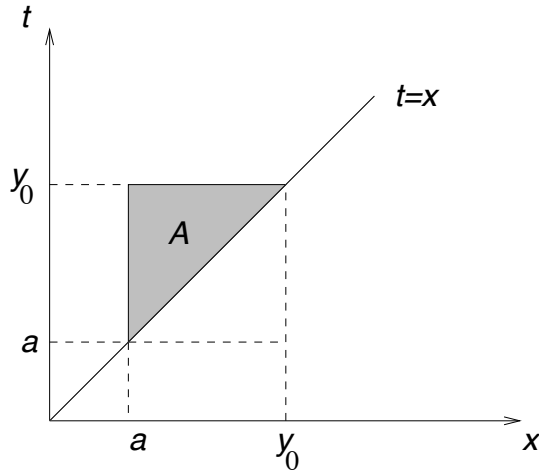
**Lemma 2.6.** Let  $g \in L_{loc}^1(I)$ . For  $y_0 \in I$  fixed, put

$$v(x) = \int_{y_0}^x g(t) dt, \quad x \in I.$$

Then  $v \in C(I)$  and

$$\int_I v\phi' = - \int_I g\phi \quad \forall \phi \in C_c^1(I).$$

**Proof.** That  $v \in C(I)$  when  $g \in L_{loc}^1(I)$ , is a well-known fact from measure theory.



We now have for  $\phi \in C_c^1(I)$  that

$$\begin{aligned} \int_I v\phi' &= \int_I \left( \int_{y_0}^x g(t) dt \right) \phi'(x) dx \\ &= - \int_a^{y_0} dx \left( \int_x^{y_0} g(t) dt \right) \phi'(x) + \int_{y_0}^b dx \left( \int_{y_0}^x g(t) dt \right) \phi'(x). \end{aligned}$$

Now

$$\begin{aligned}
-\int_a^{y_0} dx \left( \int_x^{y_0} g(t) dt \right) \phi'(x) &= -\int_a^{y_0} dx \int_x^{y_0} dt (g(t)\phi'(x)) \\
&= \text{(since } g(t)\phi'(x) \text{ is integrable on } A) \\
&= -\int_A g(t)\phi'(x) dx dt = -\int_a^{y_0} dt \int_a^t dx g(t)\phi'(x) \\
&= -\int_a^{y_0} g(t) dt \int_a^t \phi'(x) dx = -\int_a^{y_0} g(t)\phi(t) dt.
\end{aligned}$$

Similarly we may prove that

$$\int_{y_0}^b dx \left( \int_{y_0}^x g(t) dt \right) \phi'(x) = -\int_{y_0}^b g(t)\phi(t) dt.$$

We conclude that

$$\int_I v\phi' = -\int_a^{y_0} g\phi - \int_{y_0}^b g\phi = -\int_I g\phi \quad \forall \phi \in C_c^1(I).$$

□

**Proof of Theorem 2.2.** Fix  $y_0 \in I$ . Note that since  $u \in H^1(I) \Rightarrow u' \in L^2(I) \Rightarrow u' \in L_{loc}^1(I)$ . Put

$$\bar{u}(x) = \int_{y_0}^x u'(t) dt.$$

By Lemma 2.6  $\bar{u} \in C(I)$  and  $\int_I \bar{u}\phi' = -\int_I u'\phi \quad \forall \phi \in C_c^1(I)$ . But by definition of  $u'$ ,  $-\int_I u'\phi = \int_I u\phi' \quad \forall \phi \in C_c^1(I)$ . Therefore

$$\int_I (\bar{u} - u)\phi' = 0 \quad \forall \phi \in C_c^1(I).$$

By Lemma 2.5, we conclude that there exists a constant  $C$  such that  $\bar{u} - u = C$  a.e. on  $I$ . Define now  $\tilde{u}(x) = \bar{u}(x) - C$ . It follows that  $\tilde{u} \in C(\bar{I})$  and  $\tilde{u} = u$  a.e. on  $I$ . Moreover for  $x, y \in \bar{I}$ ,

$$\begin{aligned}
\tilde{u}(x) - \tilde{u}(y) &= \bar{u}(x) - \bar{u}(y) = \int_{y_0}^x u' - \int_{y_0}^y u' \\
&= \int_{y_0}^x u' + \int_y^{y_0} u' = \int_y^x u'(t) dt
\end{aligned}$$

□

**Remark:** Lemma 2.6 gives in particular that the primitive (antiderivative)  $v$  of a function  $g \in L^2(I)$  is in  $H^1(I)$  provided  $v \in L^2(I)$ . (The latter fact is always true if  $I$  is bounded).

The following theorem gives a technical tool that will be often used in sequel.

**Theorem 2.3** (Extension operator). *There exists an extension operator*

$$E : H^1(I) \rightarrow H^1(\mathbb{R}),$$

*linear and continuous, such that*

$$(i) \quad Eu|_I = u \quad \forall u \in H^1(I), \quad (f|_I \text{ denotes the restriction of } f \text{ to } I).$$

$$(ii) \quad \|Eu\|_{L^2(\mathbb{R})} \leq C\|u\|_{L^2(I)} \quad \forall u \in H^1(I).$$

$$(iii) \quad \|Eu\|_{H^1(\mathbb{R})} \leq C\|u\|_{H^1(I)} \quad \forall u \in H^1(I).$$

(In (ii) we can take  $C = 2\sqrt{2}$  and in (iii)  $C = C_0(1 + 1/\mu(I))$ , where  $C_0$  some constant, independent of  $u$  and  $I$ , and  $\mu(I)$  the length of  $I$  – possibly  $\mu(I) = \infty$ ).

**Proof.** We begin with the case  $I = (0, \infty)$ . We will show that the extension operator defined by *even reflection* about  $x = 0$ , i.e. by

$$(Eu)(x) \equiv u^*(x) = \begin{cases} u(x) & \text{if } x \geq 0 \\ u(-x) & \text{if } x < 0, \end{cases}$$

$u \in H^1(I)$ , solves the problem. Indeed

$$\|u^*\|_{L^2(\mathbb{R})}^2 = \int_{-\infty}^0 (u(-x))^2 dx + \int_0^{\infty} (u(x))^2 dx = 2\|u\|_{L^2(I)}^2.$$

So (ii) is satisfied. (Obviously  $E$  is linear and satisfies (i)). Now put

$$v(x) = \begin{cases} u'(x) & \text{if } x > 0 \\ -u'(-x) & \text{if } x < 0. \end{cases}$$

Clearly  $v \in L^2(\mathbb{R})$  since  $\|v\|_{L^2(\mathbb{R})}^2 = 2\|u'\|_{L^2(I)}^2$ . By Theorem 2.2 we also have that

$$u^*(x) - u(0) = \int_0^x u'(t) dt = \int_0^x v(t) dt \quad \text{for } x \geq 0.$$

Also, for  $x < 0$ ,

$$u^*(x) - u(0) = \int_0^{-x} u'(t) dt = \int_0^x -u'(-t) dt = \int_0^x v(t) dt.$$

Hence,  $u^*(x) - u(0) = \int_0^x v(t) dt$ ,  $x \in \mathbb{R}$ . Since  $u^* \in L^2(\mathbb{R})$  and  $v \in L^2(\mathbb{R})$ , it follows by Lemma 2.6 (see remark after end of proof of Theorem 2.2) that  $u^* \in H^1(\mathbb{R})$  and  $(u^*)' = v$ . Hence

$$\|u^*\|_{H^1(\mathbb{R})}^2 = \|u^*\|_{L^2(\mathbb{R})}^2 + \|v\|_{L^2(\mathbb{R})}^2 = 2 \left( \|u\|_{L^2(I)}^2 + \|u'\|_{L^2(I)}^2 \right) = 2\|u\|_{H^1(I)}^2.$$

Hence in the case  $I = (0, \infty)$ , with  $Eu = u^*$ , (ii) and (iii) are satisfied (as equalities) with  $C = \sqrt{2}$ . (The proof holds for any unbounded interval of the form  $(a, \infty)$  or  $(-\infty, a)$ ,  $a \in \mathbb{R}$ . For example, for  $u \in H^1((a, \infty))$ , define  $Eu$  by reflection evenly about  $x = a$ , i.e. as

$$(Eu)(x) = \begin{cases} u(x) & \text{if } x > a \\ u(2a - x) & \text{if } x \leq a, \end{cases}$$

and the proof follows – with the same constants  $C$  – mutatis mutandis).

We now go to the case of a bounded interval. It suffices to consider the case of  $I = (0, 1)$ . Consider a fixed function  $\eta \in C^1(\mathbb{R})$ ,  $0 \leq \eta(x) \leq 1 \forall x \in \mathbb{R}$  such that

$$\eta(x) = \begin{cases} 1 & \text{if } x < \frac{1}{4} \\ 0 & \text{if } x > \frac{3}{4}, \end{cases}$$

and for every  $f$  defined on  $(0, 1)$  denote by  $\tilde{f}$  its extension by zero to  $(0, \infty)$ , i.e. put

$$\tilde{f}(x) = \begin{cases} f(x) & \text{if } x \in (0, 1) \\ 0 & \text{if } x \geq 1. \end{cases}$$

Now if  $u \in H^1(I)$  it follows that  $\eta\tilde{u} \in H^1((0, \infty))$  and that  $(\eta\tilde{u})' = \eta'\tilde{u} + \eta\tilde{u}'$ , where by  $\tilde{u}'$  we mean the extension by zero to  $(0, \infty)$  of  $u' \in L^2((0, 1))$ . To see this, note first that  $\eta\tilde{u} \in L^2((0, \infty))$  since

$$\int_0^\infty \eta^2(\tilde{u})^2 \leq \int_0^{\frac{3}{4}} u^2 \leq \|u\|_{L^2((0,1))}^2.$$

Moreover, for any  $\phi \in C_c^1((0, \infty))$  we have that

$$\begin{aligned} \int_0^\infty \eta\tilde{u}\phi' &= \int_0^1 u\eta\phi' = \int_0^1 u((\eta\phi)' - \eta'\phi) = \int_0^1 u(\eta\phi)' - \int_0^1 u\eta'\phi \\ &= (\text{since } \phi \in C_c^1((0, \infty)) \Rightarrow \eta\phi \in C_c^1((0, 1)), \text{ and } u \in H^1((0, 1))) \\ &= - \int_0^1 u'(\eta\phi) - \int_0^1 u\eta'\phi = - \int_0^1 (u'\eta + u\eta')\phi = \int_0^\infty g\phi, \end{aligned}$$

where

$$g(x) = \begin{cases} u'\eta + u\eta' & \text{if } x \in (0, 1) \\ 0 & \text{if } x \geq 1. \end{cases}$$

Now  $g \in L^2((0, \infty))$  since

$$\begin{aligned} \|g\|_{L^2((0,\infty))}^2 &= \int_0^1 (u'\eta + u\eta')^2 \leq 2 \left( \int_0^1 \eta^2(u')^2 + \int_0^1 (\eta')^2 u^2 \right) \\ &\leq 2 \left( \int_0^1 (u')^2 + \max_{0 \leq x \leq 1} |\eta'(x)|^2 \int_0^1 u^2 \right) \leq C_1 \|u\|_{H^1(I)}^2. \end{aligned}$$

(Note that we can easily arrange that  $\max_{0 \leq x \leq 1} |\eta'(x)|$  be equal to e.g. 2.5). Moreover  $g = \eta'\tilde{u} + \eta\tilde{u}'$ . It follows that  $\eta\tilde{u} \in H^1((0, \infty))$  and  $(\eta\tilde{u})' = g = \eta'\tilde{u} + \eta\tilde{u}'$ . Returning to the proof of the theorem, for  $u \in H^1(I)$ ,  $I = (0, 1)$ , write  $u$  as

$$u = \eta u + (1 - \eta)u, \quad \eta \text{ as above.}$$

The function  $\eta u$  can be extended to  $(0, \infty)$  by  $\eta\tilde{u}$  as before. Clearly  $\eta\tilde{u} \in H^1((0, \infty))$  and  $\|\eta\tilde{u}\|_{L^2((0, \infty))}^2 \leq \|u\|_{L^2((0, 1))}^2$ . Also, as above

$$\begin{aligned} \|(\eta\tilde{u})'\|_{L^2((0, \infty))}^2 &= \int_0^\infty g^2 \leq 2 \left( \|u'\|_{L^2(I)}^2 + \max_{0 \leq x \leq 1} |\eta'(x)|^2 \|u\|_{L^2((0, 1))}^2 \right) \\ &\leq C_1 \|u\|_{H^1(I)}^2. \end{aligned}$$

It follows that

$$\|\eta\tilde{u}\|_{H^1((0, \infty))}^2 \leq C_2 \|u\|_{H^1(I)}^2.$$

Now, extend  $\eta\tilde{u} \in H^1((0, \infty))$  as in the first part of the proof to a function  $v_1(x) \in H^1(\mathbb{R})$  by even reflection about  $x = 0$ . It follows that

$$\|v_1\|_{L^2(\mathbb{R})} = \sqrt{2} \|\eta\tilde{u}\|_{L^2((0, \infty))} \leq \sqrt{2} \|u\|_{L^2(I)}$$

and that

$$\|v_1\|_{H^1(\mathbb{R})} = \sqrt{2} \|\eta\tilde{u}\|_{H^1((0, \infty))} \leq \sqrt{2} C_2 \|u\|_{H^1(I)}.$$

(It is clear that  $v_1|_I = \eta u$  and that the operation  $\eta u \mapsto v_1$  is linear in  $u$ ).

Analogously the function  $(1 - \eta)u$  (for which  $(1 - \eta)u = 0$  for  $0 \leq x \leq 1/4$ ), can be extended to  $(-\infty, 1)$  by  $(1 - \eta)\tilde{\tilde{u}}$  where

$$\tilde{\tilde{u}}(x) = \begin{cases} u(x) & \text{if } 0 < x < 1 \\ 0 & \text{if } -\infty < x \leq 0. \end{cases}$$

We obtain again  $(1 - \eta)\tilde{\tilde{u}} \in H^1((-\infty, 1))$  with

$$\|(1 - \eta)\tilde{\tilde{u}}\|_{L^2((-\infty, 1))} \leq \|u\|_{L^2(I)}$$

and

$$\|(1 - \eta)\tilde{\tilde{u}}\|_{H^1((-\infty, 1))} \leq C_2 \|u\|_{H^1(I)}.$$

Extend now  $(1 - \eta)\tilde{\tilde{u}}$  to a function  $v_2 \in H^1(\mathbb{R})$  by (even) reflection about  $x = 1$ . It follows that

$$\|v_2\|_{L^2(\mathbb{R})} = \sqrt{2} \|(1 - \eta)\tilde{\tilde{u}}\|_{L^2((-\infty, 1))} \leq \sqrt{2} \|u\|_{L^2(I)}$$

and that

$$\|v_2\|_{H^1(\mathbb{R})} = \sqrt{2}\|(1-\eta)\tilde{u}\|_{H^1((-\infty,1))} \leq \sqrt{2}C_2\|u\|_{H^1(I)},$$

that  $v_2|_I = (1-\eta)u$  and that  $(1-\eta)u \mapsto v_2$  is linear.

We define now the operator  $E$  as  $Eu = v_1 + v_2$ . Clearly  $E$  satisfies (i) and is linear; (ii) and (iii) follow by the above and the triangle inequality. (Note that when  $I = (a, b)$ ,  $\eta$  must be redefined as

$$\eta_{a,b}(x) = \eta\left(\frac{x-a}{x-b}\right),$$

so that

$$\eta'_{a,b}(x) = \frac{1}{b-a} \eta'\left(\frac{x-a}{x-b}\right).$$

□

The following result is a basic *density theorem* for  $H^1(I)$  and will be used very often in sequel.

**Theorem 2.4.** *Let  $u \in H^1(I)$ . There exists a sequence  $\{u_n\}_{n=1,2,\dots}$  of functions in  $C_c^\infty(\mathbb{R})$  such that*

$$u_n|_I \rightarrow u \text{ in } H^1(I), \quad n \rightarrow \infty.$$

**Comment:** The theorem asserts that if  $I = \mathbb{R}$ , then  $C_c^\infty(\mathbb{R})$  is dense in  $H^1(\mathbb{R})$ . Otherwise,  $C_c^\infty(I)$  is not dense in  $H^1(I)$  – in fact we shall see later that the closure of  $C_c^\infty(I)$  in  $H^1(I)$  is the space  $\overset{\circ}{H}^1$  consisting of those functions of  $H^1(I)$  which are zero at the boundary of  $I$ . If  $I$  is bounded, Theorem 2.4 asserts that there is a sequence of functions  $u_n \in C^\infty(\bar{I})$  such that  $u_n \rightarrow u$  in  $H^1(I)$ .

**Proof of Theorem 2.4.** First note that it suffices to consider the case  $I = \mathbb{R}$ . For suppose that the result holds for  $\mathbb{R}$ . If  $I \subset \mathbb{R}$  extend  $u$  to  $Eu$  in  $H^1(\mathbb{R})$  as in Theorem 2.3. Then there exists a sequence  $u_n \in C_c^\infty(\mathbb{R})$  such that  $\|u_n - Eu\|_{H^1(\mathbb{R})} \rightarrow 0, n \rightarrow \infty$ . But then

$$\|u_n|_I - u\|_{H^1(I)} = \|u_n - Eu\|_{H^1(I)} \leq \|u_n - Eu\|_{H^1(\mathbb{R})} \rightarrow 0, \quad n \rightarrow \infty,$$

i.e. the result holds for  $I$ .

Hence consider the case  $I = \mathbb{R}$ . The approximating sequence is constructed by *regularization* and *truncation* as  $u_n = \zeta_n(\rho_n * u)$ . Here  $\{\rho_n\}, n = 1, 2, \dots$  is the

regularizing sequence defined in Def. 2.1 and  $\zeta_n = \zeta_n(x) \in C_c^\infty(\mathbb{R})$  is a truncation function, defined for  $n = 1, 2, \dots$  by

$$\zeta_n(x) = \zeta\left(\frac{x}{n}\right),$$

where  $\zeta(x)$  is a fixed function in  $C_c^\infty(\mathbb{R})$  such that

$$\zeta(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| \geq 2. \end{cases}$$

Hence  $\zeta_n(x) = 1$  for  $|x| \leq n$  and  $\zeta_n(x) = 0$  for  $|x| \geq 2n$ . Moreover,

$$|\zeta_n'(x)| = \frac{1}{n} |\zeta'\left(\frac{x}{n}\right)| \leq \frac{C}{n}, \text{ where } C = \max_{x \in \mathbb{R}} |\zeta'(x)|.$$

Note, by Lebesgue's dominated convergence theorem, that we have  $\zeta_n f \rightarrow f$  as  $n \rightarrow \infty$  in  $L^2$  for every  $f \in L^2(\mathbb{R})$ . Let now  $u_n = \zeta_n(\rho_n * u)$ . Clearly  $u_n \in C_c^\infty(\mathbb{R})$  since  $\zeta_n \in C_c^\infty(\mathbb{R})$  and  $\rho_n * u \in C^\infty(\mathbb{R})$ , cf. Lemma 2.3. We have

$$u_n - u = \zeta_n(\rho_n * u) - u = \zeta_n[(\rho_n * u) - u] + (\zeta_n u - u).$$

It follows that

$$\begin{aligned} \|u_n - u\|_{L^2(\mathbb{R})} &\leq \|\zeta_n[(\rho_n * u) - u]\|_{L^2(\mathbb{R})} + \|\zeta_n u - u\|_{L^2(\mathbb{R})} \\ &\leq \|\rho_n * u - u\|_{L^2(\mathbb{R})} + \|\zeta_n u - u\|_{L^2(\mathbb{R})} \rightarrow 0, \quad n \rightarrow \infty \end{aligned}$$

by the above and Lemma 2.3(iii). Now  $u_n' = \zeta_n(\rho_n * u)' + \zeta_n'(\rho_n * u)$ . (Note that for  $u \in H^1(\mathbb{R})$ ,  $\rho_n * u \in H^1(\mathbb{R})$  and  $(\rho_n * u)' = \rho_n * u'$ : The interested reader may verify that for  $\phi \in C_c^1(\mathbb{R})$  the following equalities hold

$$\begin{aligned} \int_{-\infty}^{\infty} (\rho_n * u)\phi' &= \int_{-\infty}^{\infty} u(\rho_n(-x) * \phi') = \int_{-\infty}^{\infty} u(\rho_n(-x) * \phi)' = - \int_{-\infty}^{\infty} u'(\rho_n(-x) * \phi) \\ &= - \int_{-\infty}^{\infty} (\rho_n * u')\phi. \end{aligned}$$

Hence, since  $u_n' - u' = \zeta_n(\rho_n * u') + \zeta_n'(\rho_n * u) - u'$ , we have

$$\begin{aligned} \|u_n' - u'\|_{L^2(\mathbb{R})} &\leq \|\zeta_n'(\rho_n * u)\|_{L^2(\mathbb{R})} + \|\zeta_n[(\rho_n * u') - u']\|_{L^2(\mathbb{R})} \\ &\quad + \|\zeta_n u' - u'\|_{L^2(\mathbb{R})} \leq \max_{x \in \mathbb{R}} |\zeta_n'(x)| \|\rho_n * u\|_{L^2(\mathbb{R})} \\ &\quad + \|\rho_n * u' - u'\|_{L^2(\mathbb{R})} + \|\zeta_n u' - u'\|_{L^2(\mathbb{R})} \rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

by Lemma 2.3, since  $\|\rho_n * u\|_{L^2(\mathbb{R})}$ ,  $n = 1, 2, 3, \dots$  is bounded. □

We are able now to prove *Sobolev's imbedding theorem* for  $H^1(I)$ .

**Theorem 2.5.** *There exists a constant  $C$  (depending only on  $\mu(I) \leq \infty$ ) such that*

$$\|u\|_{L^\infty(I)} \leq C \|u\|_{H^1(I)} \quad \forall u \in H^1(I). \quad (2.1)$$

(We say that  $H^1(I) \subset L^\infty(I)$ , i.e. that  $H^1(I) \subset C(\bar{I})$  – in view of Theorem 2.2 – if  $I$  bounded, with continuous imbedding).

**Proof.** Again it suffices to prove the result for  $I = \mathbb{R}$ . (For suppose it holds for  $\mathbb{R}$  and let  $u \in H^1(I)$ . Extend  $u$  to  $Eu$  in  $H^1(\mathbb{R})$  as in Theorem 2.3. Then

$$\|u\|_{L^\infty(I)} \leq \|Eu\|_{L^\infty(\mathbb{R})} \leq C \|Eu\|_{H^1(\mathbb{R})} \leq C' \|u\|_{H^1(I)},$$

using (iii) in Theorem 2.3 and (2.1) for  $I = \mathbb{R}$ ). Suppose first that  $v \in C_c^\infty(\mathbb{R})$ . Then for every  $x \in \mathbb{R}$ :

$$\begin{aligned} v^2(x) &= \int_{-\infty}^x (v^2)' = 2 \int_{-\infty}^x vv' \leq 2\|v\|_{L^2(\mathbb{R})}\|v'\|_{L^2(\mathbb{R})} \\ &\leq \|v\|_{L^2(\mathbb{R})}^2 + \|v'\|_{L^2(\mathbb{R})}^2 = \|v\|_{H^1(\mathbb{R})}^2. \end{aligned}$$

Hence  $\|v\|_{L^\infty(\mathbb{R})} \leq \|v\|_{H^1(\mathbb{R})} \quad \forall v \in C_c^\infty(\mathbb{R})$ , i.e. (2.1) holds on  $C_c^\infty(\mathbb{R})$ . Now given  $u \in H^1(\mathbb{R})$ , since  $C_c^\infty(\mathbb{R})$  is dense in  $H^1(\mathbb{R})$  (Theorem 2.4), we can find a sequence  $\{u_n\}_{n=1,2,\dots}$  in  $C_c^\infty(\mathbb{R})$  such that  $u_n \rightarrow u$  in  $H^1(\mathbb{R})$ . It follows that  $u_n \rightarrow u$  in  $L^2(\mathbb{R})$ . Therefore there is a subsequence of  $u_n$  (denote it again by  $u_n$ , i.e. consider that subsequence to be the original sequence) such that  $u_n(x) \rightarrow u(x)$  a.e. on  $\mathbb{R}$  as  $n \rightarrow \infty$ . By (2.1), which was established on  $C_c^\infty(\mathbb{R})$ , we have that  $u_n$  is Cauchy in  $L^\infty(\mathbb{R})$  (since it is Cauchy in  $H^1(\mathbb{R})$ ). Therefore it converges in  $L^\infty$  to some element  $\tilde{u} \in L^\infty(\mathbb{R})$ . It follows that  $u = \tilde{u}$  a.e. on  $\mathbb{R}$ , i.e. that  $u \in L^\infty(\mathbb{R})$ . Taking limits in  $\|u_n\|_{L^\infty(\mathbb{R})} \leq \|u_n\|_{H^1(\mathbb{R})}$  we obtain (2.1).  $\square$

**Remarks.**

- (i) If  $I$  is bounded, the imbedding  $H^1(I) \subset C(\bar{I})$  is *compact*. This follows from Theorem 2.2: If  $u \in \mathcal{N}$  (=the unit ball in  $H^1(I)$  with center zero), we have

$$|u(x) - u(y)| = \left| \int_y^x u'(t) dt \right| \leq \|u'\|_{L^2(I)} |x - y|^{\frac{1}{2}} \quad \forall x, y \in I.$$

Hence  $\forall u \in \mathcal{N}$ ,  $|u(x) - u(y)| \leq |x - y|^{1/2}$  and the conclusion follows from the Arzela–Ascoli theorem.



(ii) (2.1) for  $I = \mathbb{R}$  implies that if  $u \in H^1(\mathbb{R})$ , then

$$\lim_{|x| \rightarrow \infty} u(x) = 0.$$

For if  $C_c^\infty(\mathbb{R}) \ni u_n \rightarrow u$  in  $H^1(\mathbb{R})$ , then  $\|u_n - u\|_{L^\infty(\mathbb{R})} \rightarrow 0$ ,  $n \rightarrow \infty$ . Hence  $\forall \epsilon > 0 \exists N$  such that  $\|u_N - u\|_{L^\infty(\mathbb{R})} < \epsilon \Rightarrow |u(x)| < \epsilon$  for  $|x|$  sufficiently large, since  $u_N \in C_c^\infty(\mathbb{R})$ , i.e.  $\lim_{|x| \rightarrow \infty} u(x) = 0$ .

The following two propositions follow from Theorem 2.5 and they are useful in the applications.

**Proposition 2.1.** *Let  $u, v \in H^1(I)$ . Then  $uv \in H^1(I)$  and  $(uv)' = u'v + uv'$ . Moreover we can integrate by parts:*

$$\int_y^x u'v = u(x)v(x) - u(y)v(y) - \int_y^x uv' \quad \forall x, y \in \bar{I}.$$

**Proof.** Since  $u \in H^1(I) \xrightarrow{\text{Th.2.5}} u \in L^\infty(I)$ . Hence  $v \in L^2(I) \Rightarrow uv \in L^2(I)$ . Now let  $u_n, v_n$ ,  $n = 1, 2, \dots$  be sequences in  $C_c^\infty(\mathbb{R})$  such that  $u_n|_I \rightarrow u$  in  $H^1(I)$  and  $v_n|_I \rightarrow v$  in  $H^1(I)$  as  $n \rightarrow \infty$  (Theorem 2.4). It follows by Theorem 2.5 that  $u_n|_I \rightarrow u$  in  $L^\infty(I)$  and  $v_n|_I \rightarrow v$  in  $L^\infty(I)$ . Now

$$\begin{aligned} \|u_n v_n - uv\|_{L^2(I)} &\leq \|u_n v_n - u_n v\|_{L^2(I)} + \|u_n v - uv\|_{L^2(I)} \\ &\leq \|u_n\|_{L^\infty(I)} \|v_n - v\|_{L^2(I)} + \|v\|_{L^\infty(I)} \|u_n - u\|_{L^2(I)} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Hence  $u_n v_n|_I \rightarrow uv$  in  $L^2(I)$ .

In addition  $(u_n v_n)' = u_n' v_n + u_n v_n' \rightarrow u'v + uv'$  ( $\in L^2(I)$ ) in  $L^2(I)$ . (To see this note e.g. that

$$\begin{aligned} \|u_n' v_n - u'v\|_{L^2(I)} &\leq \|u_n' v_n - u'v_n\|_{L^2(I)} + \|u'v_n - u'v\|_{L^2(I)} \\ &\leq \|v_n\|_{L^\infty(I)} \|u_n' - u'\|_{L^2(I)} + \|u'\|_{L^2(I)} \|v_n - v\|_{L^\infty(I)} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Similarly  $u_n v_n' \rightarrow uv'$  in  $L^2(I)$ ).

We now have a sequence  $\phi_n \equiv u_n v_n|_I$  in  $H^1(I)$  such that  $\phi_n \xrightarrow{L^2} \phi \equiv uv$ , and such that  $\phi_n' \xrightarrow{L^2} \psi \equiv uv' + u'v$ ,  $\psi \in L^2(I)$ . Hence  $\phi_n$  is Cauchy in  $L^2$  and  $\phi_n'$  is Cauchy in  $L^2 \Rightarrow \phi_n$  is Cauchy in  $H^1 \Rightarrow \phi_n \rightarrow w$  in  $H^1$ . Hence  $\phi_n \rightarrow w$  in  $L^2 \Rightarrow w = \phi$  and  $\phi_n' \rightarrow w'$  in  $L^2 \Rightarrow w' = \psi$ ; thus  $\phi' = (uv)' = \psi = u'v + uv'$ . The integration by parts formula

follows by integrating both members of  $(uv)' = u'v + uv'$  and using, since  $uv \in H^1(I)$ , Theorem 2.2.  $\square$

**Remark:** It is well known that  $u, v \in L^2(I) \not\Rightarrow uv \in L^2(I)$ . (Take e.g.  $I = (0, 1)$ ,  $u = v = x^{-\frac{1}{4}}$ ). Hence  $\{L^2(I), \|\cdot\|_{L^2(I)}\}$  is not a Banach algebra whereas  $\{H^1(I), \|\cdot\|_{H^1(I)}\}$  is.

**Proposition 2.2.** *Let  $G \in C^1(\mathbb{R})$  such that  $G(0) = 0$  and let  $u \in H^1(I)$ . Then  $G(u) \in H^1(I)$  and  $(G(u(x)))' = G'(u(x))u'(x)$ .*

**Proof.** Since  $u \in H^1(I) \Rightarrow u \in L^\infty(I)$ . Let  $M = \|u\|_{L^\infty(I)}$ . Since  $G \in C^1(\mathbb{R})$  and  $G(0) = 0$ , by the mean value theorem, given  $s$  there exists  $\theta : G(s) = G'(\theta)s$ . Hence, given  $\delta > 0$ , there exists a constant  $C = C(M, G, \delta)$  such that

$$|G(s)| \leq C|s| \text{ for } s \in [-M - \delta, M + \delta]. \quad (2.2)$$

Since  $|u(x)| \leq \|u\|_{L^\infty(I)}$  a.e. on  $I$ , it follows that  $-M \leq u(x) \leq M$  a.e. on  $I$ , i.e. that  $|G(u(x))| \leq C|u(x)|$  a.e. on  $I$ . Since  $u \in L^2(I)$  it follows that  $G(u) \in L^2(I)$ . Also by (2.2) we have that  $|G'(u(x))| \leq C$  a.e. on  $I$ . It follows that  $|G'(u)||u'| \leq C|u'|$  a.e.  $\Rightarrow G'(u)u' \in L^2(I)$  since  $u' \in L^2(I)$ . It remains to show that

$$\int_I G(u)\phi' = - \int_I G'(u)u'\phi \quad \forall \phi \in C_c^1(I).$$

Since  $u \in H^1(I)$ , it follows by Theorem 2.4 that there exists a sequence  $\{u_n\} \in C_c^\infty(\mathbb{R})$  such that  $u_n \rightarrow u$  in  $H^1(I)$ . Moreover, by Theorem 2.5 we have that  $u_n \rightarrow u$  in  $L^\infty(I)$ . By continuity,  $G(u_n) \rightarrow G(u)$  in  $L^\infty(I)$ . Since  $\|u_n\|_{L^\infty(I)} \rightarrow \|u\|_{L^\infty(I)}$  it follows that for  $n$  large enough,  $\|u_n\| \leq M + \delta$ . Hence (2.2) gives that for  $n$  large enough  $|G(u_n)| \leq C|u_n| \Rightarrow G(u_n) \in L^2(I)$  and  $\|G(u_n)\|_{L^2(I)} \leq C'\|u\|_{L^2(I)}$ . By the dominated convergence theorem it follows that  $G(u_n) \rightarrow G(u)$  in  $L^2(I)$ . Hence  $\forall \phi \in C_c^1(I)$   $\int_I G(u_n)\phi' \rightarrow \int_I G(u)\phi'$ , as  $n \rightarrow \infty$ . Now

$$\begin{aligned} \|G'(u_n)u'_n - G'(u)u'\|_{L^2(I)} &\leq \|G'(u_n)u'_n - G'(u_n)u'\|_{L^2(I)} + \\ + \|G'(u_n)u' - G'(u)u'\|_{L^2(I)} &\leq \|G'(u_n)\|_{L^\infty(I)}\|u'_n - u'\|_{L^2(I)} + \\ &+ \|u'\|_{L^2(I)}\|G'(u_n) - G'(u)\|_{L^\infty(I)}. \end{aligned}$$

Now,  $\|u'_n - u'\|_{L^2(I)} \rightarrow 0$  since  $u_n \rightarrow u$  in  $H^1(I)$ . Also, for  $n$  large enough  $\|u_n\|_{L^\infty(I)} \leq M + \frac{\delta}{2}$  and therefore  $\|G'(u_n)\|_{L^\infty(I)} \leq C$ . Since  $u_n \rightarrow u$  in  $L^\infty(I)$ , by continuity we

have that  $G'(u_n) \rightarrow G'(u)$  in  $L^\infty(I)$ . It follows that  $G'(u_n)u'_n \rightarrow G'(u)u'$  in  $L^2(I)$ ; hence  $-\int_I G'(u_n)u'_n\phi \rightarrow -\int_I G'(u)u'\phi \forall \phi \in C_c^1(I)$ . Since  $\phi \in C_c^1(I)$ ,  $u_n \in C_c^\infty(\mathbb{R})$  we have that

$$\int_I G(u_n)\phi' = -\int_I G'(u_n)u'_n\phi, \quad n = 1, 2, 3, \dots$$

Letting  $n \rightarrow \infty$  we obtain the desired equality

$$\int_I G(u)\phi' = -\int_I G'(u)u'\phi.$$

□

## 2.4 The Sobolev spaces $H^m(I)$ , $m = 2, 3, 4, \dots$

In analogy to  $H^1(I)$  we define for  $m \geq 2$  integer the space

$$H^m(I) = \{u \in L^2(I) : \exists g_i \in L^2(I), i = 1, \dots, m \text{ such that} \\ \int_I u\phi^{(i)} = (-1)^i \int_I g_i\phi, 1 \leq i \leq m \quad \forall \phi \in C_c^\infty(I)\}.$$

Here

$$\phi^{(i)} = \left(\frac{d}{dx}\right)^i \phi.$$

It follows easily that  $H^m(I) \subset H^1(I)$  and that  $g_1 = u'$ , that  $u' \in H^1(I)$ , and that  $(u')' = g_2$  etc..., and that finally

$$u^{(m-1)} = \underbrace{(((u')')' \dots)'}_{m-1 \text{ times}} \in H^1(I) \text{ and that } u^{(m)} = (u^{(m-1)})' = g_m.$$

We call  $g_i$  the (uniqueness easy) *weak (generalized) derivative of order  $i$*  (in the  $L^2$  sense) of  $u \in H^m(I)$  and define

$$D^i u \equiv u^{(i)} \equiv g_i, \quad 1 \leq i \leq m.$$

It follows easily that for  $m \geq 1$

$$H^m(I) = \{u \in H^{m-1}(I) : u' \in H^{m-1}(I)\}.$$

Here  $H^0(I) \equiv L^2(I)$ . We can easily construct examples of functions in  $H^m(I)$ . For example, on a bounded interval if  $u \in C^1(\bar{I})$  with  $u''$  (classical derivative) piecewise continuous on  $\bar{I}$ , then  $u \in H^2(I)$  and its weak second derivative coincides a.e. with  $u''$ .

We equip  $H^m(I)$  with the inner product  $(\cdot, \cdot)_m$ , where

$$(u, v)_m = \sum_{j=0}^m (D^j u, D^j v), \quad (u^{(0)} \equiv D^0 u = u), \quad \forall u, v \in H^m(I).$$

This inner product induces the norm

$$\|u\|_m = (u, u)_m^{\frac{1}{2}} = \left( \sum_{i=0}^m \|D^i u\|^2 \right)^{\frac{1}{2}}.$$

An obvious modification of Theorem 2.1 shows that  $\{H^m(I), (\cdot, \cdot)_m\}$  is a Hilbert space. By definition and Theorem 2.2, it follows that if  $u \in H^m(I)$ , then there exists  $\tilde{u} \in C^{m-1}(\bar{I})$  such that  $u = \tilde{u}$  a.e. on  $I$  and

$$D^i \tilde{u}(x) - D^i \tilde{u}(y) = \int_y^x D^{i+1} u(t) dt, \quad \forall x, y \in \bar{I}, \quad i = 0, 1, 2, \dots, m-1.$$

Also, given  $u \in H^m(I)$ , there exists a sequence  $\{u_n\} \in C_c^\infty(\mathbb{R})$  such that  $u_n|_I \rightarrow u$  in  $H^m(I)$  (density) and that  $H^m(I) \subset C^{m-1}(I)$  with

$$\sum_{j=0}^{m-1} \|D^j u\|_{L^\infty(I)} \leq C \|u\|_{H^m(I)}, \quad \forall u \in H^m(I).$$

If  $u, v \in H^m(I)$ , then  $uv \in H^m(I)$  and

$$D^m(uv) = \sum_{j=0}^m \binom{m}{j} D^j u D^{m-j} v, \quad m \geq 1 \text{ (Leibniz's rule)}.$$

These results follow easily with techniques similar to the ones used in their  $H^1$  counterparts.

We finally mention without proof the following *interpolation* result. If  $1 \leq j \leq m-1$ , then  $\forall \epsilon > 0 \exists C_\epsilon = C(\epsilon, \mu(I) \leq \infty)$  such that

$$\|D^j u\| \leq \epsilon \|D^m u\| + C_\epsilon \|u\| \quad \forall u \in H^m(I).$$

It follows that the quantity  $\|u\| + \|D^m u\|$  for  $u \in H^m(I)$  is a *norm* on  $H^m(I)$ , equivalent to  $\|u\|_m$ .

## 2.5 The space $\overset{\circ}{H}^1(I)$

**Definition** We define  $\overset{\circ}{H}^1(I)$  to be the closure of  $C_c^1(I)$  in  $H^1(I)$ , i.e. the (closed) subspace of  $H^1(I)$  whose elements are limits in  $H^1(I)$  of sequences of functions in  $C_c^1(I)$ .

It follows that  $\{\overset{\circ}{H}^1(I), (\cdot, \cdot)_1\}$  is a complete Hilbert space (separable).

**Remarks.**

- (i) If  $I = \mathbb{R}$ , since  $C_c^\infty(\mathbb{R}) \subset C_c^1(\mathbb{R}) \subset H^1(\mathbb{R})$  and  $C_c^\infty(\mathbb{R})$  is dense in  $H^1(\mathbb{R})$  (cf. Theorem 2.4) it follows that  $C_c^1(\mathbb{R})$  is dense in  $H^1(\mathbb{R}) \Rightarrow \overset{\circ}{H}^1(\mathbb{R}) = H^1(\mathbb{R})$ . However if  $I \neq \mathbb{R}$ , then  $\overset{\circ}{H}^1(I) \subset H^1(I)$ . For example, on a bounded interval  $I$  consider  $u(x) = c_1 e^x + c_2 e^{-x} \in C^\infty(\bar{I})$  for which  $u'' = u$ . Hence

$$0 = (u'' - u, \phi) = \int_I (u'' - u)\phi = - \int_I u'\phi' - \int_I u\phi = -(u, \phi)_1 \quad \forall \phi \in C_c^1(I).$$

Hence  $u$  is orthogonal in  $H^1(I)$  to  $C_c^1(I)$ , which therefore cannot be dense in  $H^1(I)$ .

- (ii) In fact  $C_c^\infty(I)$  is dense in  $\overset{\circ}{H}^1(I)$ . To see this, let, for  $u \in \overset{\circ}{H}^1(I)$ ,  $\epsilon > 0$ ,  $\phi \in C_c^1(I)$  be such that  $\|u - \phi\|_1 \leq \frac{\epsilon}{2}$ . Now extending  $\phi \in C_c^1(I)$  by zero outside its support to the whole of  $\mathbb{R}$ , we have  $\phi \in C_c^1(\mathbb{R})$  and  $\rho_n * \phi \in C_c^\infty(I)$  for sufficiently large  $n$  (cf. Lemma 2.3). Moreover, (cf. Remark (iii), p. 30)  $\|\phi - \rho_n * \phi\|_{H^1(I)} \rightarrow 0$  as  $n \rightarrow \infty$ . Therefore choose  $n$  so that  $\|\phi - \rho_n * \phi\|_1 \leq \frac{\epsilon}{2}$ , from which it follows that  $\|u - \rho_n * \phi\|_1 \leq \epsilon$ , i.e. that  $C_c^\infty(I)$  is dense in  $\overset{\circ}{H}^1(I)$ .

- (iii) Let us also remark that  $u \in H^1(I) \cap C_c(I) \Rightarrow u \in \overset{\circ}{H}^1(I)$ . In fact, if  $u \in H^1(I) \cap C_c(I)$ , extending  $u$  by zero outside  $I$  to the whole of  $\mathbb{R}$ , we have, by Lemma 2.3, that, for  $n$  sufficiently large,  $\rho_n * u \in C_c^\infty(I)$  and (cf. Proof of Theorem 2.4),  $\|u - \rho_n * u\|_1 \rightarrow 0$  as  $n \rightarrow \infty$ . Hence by (ii) above,  $u \in \overset{\circ}{H}^1(I)$ .

These remarks have prepared the ground for the following theorem which characterizes the functions in  $\overset{\circ}{H}^1(I)$  in a very useful way.

**Theorem 2.6.** *Let  $u \in H^1(I)$ . Then  $u \in \overset{\circ}{H}^1(I)$  if and only if  $u = 0$  on  $\partial I$  (= the boundary of  $I$ ).*

**Comments:** Again for  $u \in H^1(I)$  the statement “ $u = 0$  on  $\partial I$ ” is well understood, in the sense of Theorem 2.2 – see the remarks there. Theorem 2.6 makes then  $\overset{\circ}{H}^1(I)$  a very useful space, in which homogeneous (zero) boundary conditions are automatically satisfied for  $u|_{\partial I}$ , i.e. a space in which “weak” solutions of boundary-value problems such as (\*), p. 25 may be naturally sought. Again assume  $I \neq \mathbb{R}$ , since  $H^1(\mathbb{R}) = \overset{\circ}{H}^1(\mathbb{R})$ .

**Proof.** If  $u \in \overset{0}{H}^1(I)$ , then there exists a sequence  $\{u_n\} \in C_c^1(I)$  such that  $u_n \rightarrow u$  in  $H^1(I)$ . By Sobolev's Theorem 2.5  $u_n \rightarrow u$  in  $L^\infty(I)$ , i.e., if  $\tilde{u}(x) \in C(\bar{I})$ ,  $\tilde{u} = u$  a.e. on  $I \Rightarrow \sup_{x \in I} |u_n(x) - \tilde{u}(x)| \rightarrow 0$ ,  $n \rightarrow \infty$ . Since  $u_n|_{\partial I} = 0 \Rightarrow \tilde{u}|_{\partial I} = 0 \Leftrightarrow u|_{\partial I} = 0$ .

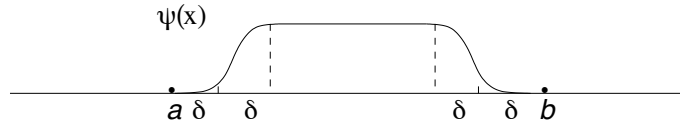
Suppose now that  $u \in H^1(I)$  and  $u|_{\partial I} = 0$ . We shall show that  $u \in \overset{0}{H}^1(I)$ .

First, let us examine the case of a bounded interval and take, with no loss of generality,  $I = (0, 1)$ . Consider the intervals

$$K_1 = \left(-\frac{1}{3}, \frac{1}{3}\right), \quad I = (0, 1), \quad K_2 = \left(\frac{2}{3}, \frac{4}{3}\right).$$

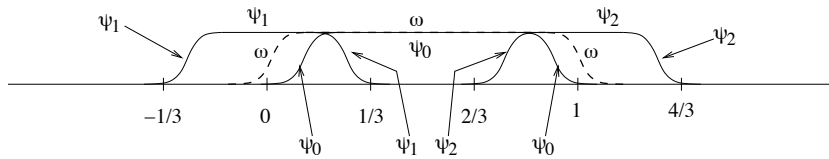
Then, we may find functions  $\phi_1 \in C_c^\infty(K_1)$ ,  $\phi_0 \in C_c^\infty(I)$ ,  $\phi_2 \in C_c^\infty(K_2)$  such that  $0 \leq \phi_i \leq 1$  and (extending them by zero outside their intervals of definition)  $\phi_1(x) + \phi_0(x) + \phi_2(x) = 1$ ,  $\forall x \in I$ . (The functions  $\phi_i$  form a *partition of unity* corresponding to the open cover  $\{K_1, I, K_2\}$  of  $I$ , and can be constructed e.g. as follows:

It is clear that given any interval  $(a, b)$  we can find  $\psi \in C_c^\infty(a, b)$  such that (for  $\delta > 0$



small enough)  $\psi(x) = 1$  for  $a + 2\delta \leq x \leq b - 2\delta$  and  $\psi(x) = 0$  for  $x \in (a, a + \delta]$  and  $x \in [b - \delta, b)$ . With the same  $\delta$  (take any  $0 < \delta < \frac{1}{12}$ ) construct such functions  $\psi_1(x)$  for  $(-\frac{1}{3}, \frac{1}{3})$ ,  $\psi_0(x)$  for  $(0, 1)$ ,  $\psi_2(x)$  for  $(\frac{2}{3}, \frac{4}{3})$ . It is clear then that  $\psi_0(x) + \psi_1(x) + \psi_2(x) \geq 1$ ,  $x \in [-\frac{1}{3} + 2\delta, \frac{4}{3} - 2\delta]$ .

Let  $\omega(x)$  be such a function (with the same e.g.  $\delta$ ) for the interval  $[-2\delta, 1 + 2\delta]$ . Then



let

$$\phi_i(x) = \frac{\psi_i(x)}{\sum \psi_i(x)} \omega(x), \quad i = 0, 1, 2.$$

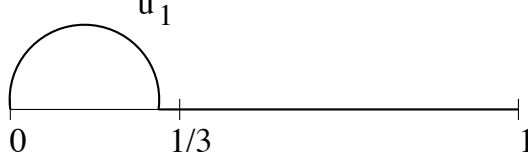
It is easily seen that the  $\phi_i$ 's satisfy the desired properties).

Let now  $u_i(x) = u(x)\phi_i(x)$ ,  $i = 0, 1, 2$ , so that

$$u(x) = \sum_{i=0}^2 u_i(x), \quad x \in I.$$

Now  $u_0(x) = u(x)\phi_0(x)$ . Since  $u \in H^1(I)$ ,  $\phi_0 \in C_c^\infty(I) \Rightarrow u_0 \in H^1(I) \cap C_c(I)$  (identifying  $u_0$  with its continuous representative). By Remark (iii) p. 46,  $u_0 \in \mathring{H}^1(I)$ .

We consider now  $u_1(x) = u(x)\phi_1(x)$ . Extending  $\phi_1$  by zero to the whole of  $I$ , we see that  $u \in H^1(I)$ ,  $\phi_1 \in C_c^\infty(-\frac{1}{3}, 1) \Rightarrow u_1 \in H^1(I)$ . By hypothesis we also have that



$u_1(0) = u(0)\phi_1(0) = 0 \cdot \phi_1(0) = 0$ . Also  $\text{supp}u_1 \subset [0, \frac{1}{3}]$ . Extend now  $u_1$  by zero to  $\mathbb{R}$ , i.e. consider the function

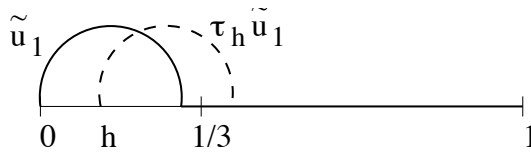
$$\tilde{u}_1(x) = \begin{cases} u_1(x) & \text{if } x \in I \\ 0 & \text{if } x \notin I. \end{cases}$$

This function belongs to  $H^1(\mathbb{R})$ , since, for any  $\phi \in C_c^\infty(\mathbb{R})$ ,

$$\begin{aligned} \int_{-\infty}^{\infty} \tilde{u}_1 \phi' &= \int_0^1 u_1 \phi' \stackrel{\text{Prop. 2.1}}{=} \underbrace{u_1(1)}_{=0} \phi(1) - \underbrace{u_1(0)}_{=0} \phi(0) - \int_0^1 u_1' \phi \\ &= - \int_{-\infty}^{\infty} \tilde{u}_1' \phi, \end{aligned}$$

where  $\tilde{u}_1'$ , the extension by zero outside  $I$  of  $u_1'$ , is in  $L^2(\mathbb{R})$  since  $u_1 \in H^1(I)$ .

Now, for any function  $f$  on  $\mathbb{R}$  let for  $h > 0$   $\tau_h f$  denote the right  $h$ -translate of  $f$ ,



$$(\tau_h f)(x) \equiv f(x - h).$$

Since  $\tilde{u}_1 \in H^1(\mathbb{R})$  it follows that  $\tau_h \tilde{u}_1 \in H^1(\mathbb{R})$  and

$$\lim_{h \rightarrow 0} \|\tau_h \tilde{u}_1 - \tilde{u}_1\|_{H^1(\mathbb{R})} = 0.$$

(To see this, given  $\epsilon > 0$  let  $\phi \in C_c^\infty(\mathbb{R})$  be such that  $\|\tilde{u}_1 - \phi\|_{H^1(\mathbb{R})} < \frac{\epsilon}{3}$ . It follows, for any  $h > 0$ , that

$$\|\tau_h \tilde{u}_1 - \tau_h \phi\|_{H^1(\mathbb{R})} = \|\tilde{u}_1 - \phi\|_{H^1(\mathbb{R})} < \frac{\epsilon}{3}.$$

But it is obvious, since  $\phi, \tau_h \phi \in C_c^\infty(\mathbb{R})$ , that there exists  $h_0$  such that

$$0 \leq h < h_0 \Rightarrow \|\tau_h \phi - \phi\|_{H^1(\mathbb{R})} \leq \frac{\epsilon}{3}.$$

Hence, given  $\epsilon > 0 \exists h_0$  such that

$$0 \leq h < h_0 \Rightarrow \|\tau_h \tilde{u}_1 - \tilde{u}_1\|_{H^1(\mathbb{R})} < \epsilon$$

by the triangle inequality). Now, the restriction  $\tau_h \tilde{u}_1|_I$ , for  $h$  sufficiently small, belongs to  $H^1(I) \cap C_c(I)$  (possibly upon modification it on a set of measure zero). Hence, by Remark (iii), p. 46,  $\tau_h \tilde{u}_1|_I \in \mathring{H}^1(I)$ . Therefore, given  $\epsilon > 0$  we have that  $\exists \phi \in C_c^\infty(I)$  such that  $\|\tau_h \tilde{u}_1 - \phi\|_{H^1(I)} < \frac{\epsilon}{2}$ . Since  $\lim_{h \rightarrow 0} \|\tau_h \tilde{u}_1 - \tilde{u}_1\|_{H^1(\mathbb{R})} = 0 \Rightarrow \exists h$  such that

$$\|\tau_h \tilde{u}_1 - \tilde{u}_1\|_{H^1(\mathbb{R})} = \|\tau_h \tilde{u}_1 - u_1\|_{H^1(I)} < \frac{\epsilon}{2}.$$

It follows that  $\|u_1 - \phi\|_{H^1(I)} < \epsilon$ , i.e. that  $u_1 \in \mathring{H}^1(I)$ .

Entirely analogous considerations show that  $u_2 \in \mathring{H}^1(I)$ . Since  $u = u_0 + u_1 + u_2$  it follows that  $u \in \mathring{H}^1(I)$  QED.

For a semi-infinite interval the proof follows in the analogous manner. Let  $I = (0, \infty)$ , with no loss of generality. Construct  $\phi_1 \in C_c^\infty(-\frac{1}{3}, \frac{1}{3})$ ,  $\phi_0 \in C_c^\infty(0, \infty)$  with  $\text{supp} \phi_0 \subset [\alpha, \infty)$ ,  $\alpha > 0$  sufficiently small, so that  $0 \leq \phi_i \leq 1$  and so that (extending  $\phi_1$  by zero to  $[\frac{1}{3}, \infty)$ )  $\phi_1(x) + \phi_0(x) = 1$  for  $x \in [0, \infty)$ . (This can be achieved as on p. 47, by taking  $\psi_1$  as before and extending  $\psi_0(x)$  and  $w(x)$  by setting them equal to 1 for  $x \geq \frac{1}{2}$ ). Again with  $u_i = u\phi_i$  we have as before that, since  $u_0 = 0$ ,  $u_1 \in \mathring{H}^1(0, \infty)$ . Consider  $u_0 = u\phi_0$ . Extend it by zero to the whole of  $\mathbb{R}$ , i.e. put

$$\tilde{u}_0(x) = \begin{cases} u_0(x) & \text{if } 0 \leq x < \infty \\ 0 & \text{if } -\infty < x < 0. \end{cases}$$

It is clear that  $\tilde{u}_0 \in H^1(\mathbb{R})$  (since  $u_0 \in H^1(I)$ ). Since  $u_0$  it has support in  $[\alpha, \infty)$ ,  $\alpha > 0$ , it is not hard to see that for  $n$  sufficiently large,  $\rho_n * \tilde{u}_0$ , has support in  $[\alpha', \infty)$ ,  $\alpha' > 0$ , belongs to  $C^\infty(\mathbb{R}) \cap H^1(\mathbb{R})$  and, of course,

$$\|\tilde{u}_0 - \rho_n * \tilde{u}_0\|_{H^1(\mathbb{R})} = \|u_0 - (\rho_n * \tilde{u}_0)|_I\|_{H^1(I)} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Consider now the functions

$$u_{0,n} = \zeta_n|_I(\rho_n * \tilde{u}_0)|_I,$$



where  $\zeta_n(x)|_I$  is the restriction to  $I = [0, \infty)$  of the truncation function  $\zeta_n(x)$  introduced in the proof of Theorem 2.4. It follows that for  $n$  sufficiently large,  $u_{0,n} \in C_c^\infty(I)$ . A similar calculation to the one used in the proof of Theorem 2.4 shows finally that

$$\|u_{0,n} - u_0\|_{H^1(I)} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Hence  $u_0 \in \overset{\circ}{H}^1(I)$ . □

**Remark:** Essentially the proof above shows the following characterization of  $\overset{\circ}{H}^1(I)$  which is of interest by itself. For  $u \in L^2(I)$  let  $\tilde{u}(x)$  be its extension by zero to the whole real line, i.e. let

$$\tilde{u}(x) = \begin{cases} u(x) & \text{if } x \in I \\ 0 & \text{if } x \in \mathbb{R} - I. \end{cases}$$

Then,  $u \in \overset{\circ}{H}^1(I)$  if and only if  $\tilde{u} \in H^1(\mathbb{R})$ .

We finally mention a result which will be very useful in the existence – uniqueness theory of weak solutions of boundary–value problems.

**Proposition 2.3** (Inequality of Poincaré–Friedrichs). *Suppose that  $I$  is a bounded interval. Then, there exists a constant  $C_*$  (depending on  $\mu(I)$ ) such that*

$$\|u\|_1 \leq C_* \|u'\| \quad \forall u \in \overset{\circ}{H}^1(I), \tag{2.3}$$

*in other words, the quantity  $\|u'\|$  is a norm on  $\overset{\circ}{H}^1(I)$ , equivalent to  $\|u\|_1$ .*

**Proof.** If  $u \in \overset{\circ}{H}^1(I) \equiv \overset{\circ}{H}^1(a, b)$  we have, by Theorems 2.2 and 2.6 that for  $x \in I$

$$|u(x)| = \left| \int_a^x u'(t) dt \right| \leq \int_a^b |u'(t)| dt \leq (b-a)^{\frac{1}{2}} \|u'\|.$$

Hence

$$\|u\|^2 = \int_a^b u^2(t) dt \leq (b-a)^2 \|u'\|^2,$$

and (2.3) holds with  $C_* = (1 + (b-a)^2)^{1/2}$ . □

**Remarks:**

- (i) It follows that on  $\overset{\circ}{H}^1(I)$ ,  $I$  bounded, the expression  $(u', v')$  defines an inner product.

(ii) Entirely analogously one may define, for  $m \geq 2$  integer, the spaces  $\overset{0}{H}{}^m(I)$  as completions of  $C_c^\infty(I)$  in  $H^m(I)$ . One may show that

$$\overset{0}{H}{}^m(I) = \{u \in H^m(I) : u = Du = \dots = D^{m-1}u = 0 \text{ on } \partial I\}.$$

We should keep in mind the distinction between e.g.

$$\overset{0}{H}{}^2(I) = \{u \in H^2(I) : u = Du = 0 \text{ on } \partial I\}$$

and

$$H^2(I) \cap \overset{0}{H}{}^1(I) = \{u \in H^2(I) : u = 0 \text{ on } \partial I\}.$$

## 2.6 Two–point boundary–value problems

We return now to the two–point boundary–value problem (\*) (p. 25) (also called a *Sturm–Liouville* problem). We shall mainly discuss homogeneous (zero) Dirichlet and Neumann boundary conditions and follow the “variational method” outlined in 2.1.

### 2.6.1 Zero Dirichlet boundary conditions.

We let  $I = (a, b)$  be a bounded interval. We consider the problem of finding  $u(x)$ ,  $x \in [a, b]$  such that

$$-(pu')' + qu = f \quad \text{in } (a, b), \tag{2.4}$$

$$u(a) = u(b) = 0, \tag{2.5}$$

where  $p, q, f$  are given functions on  $I$  such that  $p \in C^1(\bar{I})$ ,  $p(x) \geq \alpha > 0 \quad \forall x \in \bar{I}$ ,  $q \in C(\bar{I})$  such that  $q(x) \geq 0 \quad \forall x \in \bar{I}$  and where we shall assume  $f \in C(\bar{I})$  (sometimes just  $f \in L^2(I)$ ). Based on the discussion on p. 26 we make the following

**Definition** *Let  $f \in C(\bar{I})$ . Then a classical solution of (2.4), (2.5), is a function  $u(x) \in C^2(\bar{I})$  which satisfies the D.E. (2.4) in the usual sense for each  $x \in I$  and which also satisfies the b.c. (2.5) in the usual sense. (We shall say that “ $u$  satisfies (2.4) and (2.5) in the usual sense”). If  $f \in L^2(I)$ , then a weak solution of (2.4), (2.5), is a function  $u \in \overset{0}{H}{}^1(I)$  which satisfies*

$$\int_I pu'v' + \int_I quv = \int_I fv \quad \forall v \in \overset{0}{H}{}^1(I). \tag{2.6}$$

Following the program outlined on pp. 25–26 we prove a series of results:

**(i) A classical solution of (2.4), (2.5) is a weak solution of (2.4), (2.5) as well.**

Let  $u$  satisfy (2.4), (2.5) classically. Since  $u \in C^2(\bar{I})$  and  $u(a) = u(b) = 0 \Rightarrow u \in H^2(I) \cap \overset{\circ}{H}^1(I)$ . We multiply the D.E. (2.4) by any  $v \in \overset{\circ}{H}^1(I)$  and integrate on  $I$ . By Proposition 2.1, p. 42, since  $pu' \in H^1$  we can integrate by parts and obtain

$$\begin{aligned} \int_I f v &= - \int_I (pu')' v + \int_I q u v = -pu'v|_a^b + \int_I pu'v' + \int_I q u v \\ &= \int_I pu'v' + \int_I q u v, \end{aligned}$$

since  $v \in \overset{\circ}{H}^1 = \overset{\circ}{H}^1(I)$  (we suppress  $I$  from the symbols of the function spaces). Hence  $u$  is a weak solution.

**(ii) Existence and uniqueness of the weak solution.**

Let  $f \in L^2(I)$  and let  $B(v, w) = \int_I pv'w' + \int_I qvw$ . Clearly, by our hypotheses  $B(\cdot, \cdot)$  is a bilinear, symmetric form on  $\overset{\circ}{H}^1 \times \overset{\circ}{H}^1$  (i.e. on  $H^1 \times H^1$ ). Moreover, for  $v, w \in H^1$  we have

$$\begin{aligned} |B(v, w)| &\leq \int_I |p||w'| |v'| + \int_I |q||w||v| \\ &\leq \max_{x \in \bar{I}} |p(x)| \|w'\| \|v'\| + \max_{x \in \bar{I}} |q(x)| \|w\| \|v\| \\ &\leq c_1 \|v\|_1 \|w\|_1, \end{aligned} \tag{2.7}$$

where  $c_1 = \max_{x \in \bar{I}} |p(x)| + \max_{x \in \bar{I}} |q(x)|$ . Hence  $B$  is continuous on  $H^1 \times H^1$ .

Now, for  $v \in \overset{\circ}{H}^1$  we have, by our hypotheses on  $p$  and  $q$  that

$$B(v, v) = \int_I p(v')^2 + \int_I qv^2 \geq \alpha \int_I (v')^2 \geq c_2 \|v\|_1^2, \tag{2.8}$$

where  $c_2 = \alpha/C_*^2$ , with  $C_*$  being the constant in the Poincaré–Friedrichs inequality, which was used (since  $v \in \overset{\circ}{H}^1$ ) in the last step.

Hence the bilinear form  $B$  satisfies, on the Hilbert space  $\{\overset{\circ}{H}^1, \|\cdot\|_1\}$ , the hypotheses (i) and (ii) of the Lax–Milgram theorem (p. 19). In addition it is straightforward to see that in the R.H.S. of (2.6),  $F(v) = \int_I f v$  is a continuous, linear functional on  $\overset{\circ}{H}^1$  since  $|F(v)| \leq \|f\| \|v\| \leq \|f\| \|v\|_1 \forall v \in \overset{\circ}{H}^1$ . (Note that  $\|F\| \leq \|f\|$  where by  $\|F\|$  we denote the norm of the b.l.f.  $F$  on  $\overset{\circ}{H}^1$ , i.e. the  $\sup_{0 \neq v \in \overset{\circ}{H}^1} \frac{|F(v)|}{\|v\|_1}$ ). Hence the

Lax–Milgram theorem applies and shows that the problem

$$\int_I pu'v' + quv = \int_I fv, \quad \text{i.e. } B(u, v) = F(v) \quad \forall v \in \overset{\circ}{H}^1,$$

has a unique solution  $u \in \overset{\circ}{H}^1$ . Moreover

$$\|u\|_1 \leq \frac{1}{c_2} \|f\|. \quad (2.9)$$

Since a classical solution is a weak solution and we have shown now uniqueness of the weak solution, it follows that there is at most one classical solution.

**(iii) Regularity of the weak solution.**

Let  $f \in L^2(I)$ . Then (2.6) gives that if  $u$  is the weak solution, then

$$\int_I pu'v' = \int_I (f - qu)v \quad \forall v \in \overset{\circ}{H}^1.$$

Hence

$$\int_I pu'\phi' = \int_I (f - qu)\phi \quad \forall \phi \in C_c^\infty(I),$$

which shows that  $pu' \in H^1$ , since  $pu' \in L^2$  and since there exists  $g (= qu - f)$ ,  $g \in L^2$  such that  $\int_I pu'\phi' = -\int_I g\phi \quad \forall \phi \in C_c^\infty(I)$ . It follows that  $pu' \in H^1$  and  $(pu')' = qu - f$ . i.e.  $-(pu')' + qu = f$  holds in  $L^2$ .

Now, since  $p(x) \geq \alpha > 0$ ,  $x \in \bar{I}$ ,  $p \in C^1(\bar{I})$ , it follows that  $\frac{1}{p(x)} \in C^1(\bar{I})$ . Hence  $u' = \frac{1}{p}(pu') \in H^1$ , since  $pu' \in H^1$  and  $p^{-1} \in H^1$ . It follows that the weak solution  $u$  (shown to be in  $\overset{\circ}{H}^1$  by the Lax–Milgram theorem) actually belongs to  $H^2 \cap \overset{\circ}{H}^1$ . Moreover, since  $-(pu')' + qu = f$  in  $L^2$ , we have now  $pu'' = -p'u' + qu - f$ . Hence

$$\alpha \|u''\| \leq \max_{x \in \bar{I}} |p'(x)| \|u'\| + \max_{x \in \bar{I}} |q(x)| \|u\| + \|f\|.$$

This estimate coupled with (2.9) ( $\|u\|_1 \leq c_2^{-1} \|f\|$ ), shows that there exists a constant  $c_3 = c_3(p, q, I)$ , such that for the weak solution  $u \in H^2 \cap \overset{\circ}{H}^1$  of (2.4), (2.5) we also have

$$\|u\|_2 \leq c_3 \|f\|. \quad (2.10)$$

(Estimates such as (2.10) are called “elliptic regularity” estimates ( $H^2 - L^2$ ) for equation (2.4)).

Now, let  $f \in C(\bar{I})$ . We shall show now that the weak solution  $u$  is in  $C^2(\bar{I})$  ( $u \in \overset{\circ}{H}^1(I)$  guarantees that  $u(a) = u(b) = 0$  in the sense of Sobolev’s theorem). Since

$u \in H^2(I) \Rightarrow u' \in C(\bar{I})$  (take the continuous representative of  $u'$ ). Hence  $u \in C^1(\bar{I})$ . But  $pu'' = -p'u' + qu - f$  (in  $L^2$ ). Since  $f \in C(\bar{I})$ , the R.H.S. is continuous  $\Rightarrow u'' \in C(\bar{I})$  by our hypotheses on  $p(x)$ . Hence  $u \in C^2(\bar{I})$ .

**(iv) If  $f \in C(\bar{I})$ , the weak solution is a classical solution.**

Since  $f \in C(\bar{I})$  the above shows that  $u \in C^2(\bar{I})$ . Now (2.6) gives that

$$\int_I (pu')\phi' + \int_I qu\phi = \int_I f\phi \quad \forall \phi \in C_c^\infty(I).$$

Integrating by parts, since  $u \in C^2(\bar{I})$ , we have that

$$\int_I (-(pu')' + qu - f)\phi = 0 \quad \forall \phi \in C_c^\infty(I).$$

Since  $C_c^\infty(I)$  is dense in  $L^2(I) \Rightarrow -(pu')' + qu - f = 0$  a.e. on  $I$ . (This also follows from the fact that  $-(pu')' + qu - f = 0$  in  $L^2$ ). But  $-(pu')' + qu - f \in C(\bar{I})$ . Therefore  $-(pu')' + qu = f \quad \forall x \in I$ . Since  $u(a) = u(b) = 0$  as well, it follows that for  $f \in C(\bar{I})$  the weak solution is also classical.

### Remarks

- (i) Since  $B(u, v)$  is *symmetric*, the weak solution of (2.4), (2.5), i.e. the solution of the variational problem

$$B(u, v) = (f, v) \quad \forall v \in \overset{0}{H}^1,$$

may be characterized as the (unique) solution of the minimization problem

$$J(u) = \min_{v \in \overset{0}{H}^1} J(v),$$

where  $J$  is the *energy functional*

$$J(v) = \frac{1}{2} B(v, v) - F(v) = \frac{1}{2} \int_I p(v')^2 + qv^2 - \int_I fv.$$

This follows from the Rayleigh–Ritz Theorem 1.5 and is known in the present context as *Dirichlet's principle*.

- (ii) It is evident that the weak solution of (2.4), (2.5), i.e. the solution of (2.6), exists under much weaker conditions on  $p$  and  $q$  than those assumed. For example, the two integrals appearing in the definition of  $B(v, w)$  exist if e.g.  $p, q \in L^\infty(I)$ .

Similarly (2.7) and (2.8) hold only under the additional assumptions  $p \geq \alpha > 0$ ,  $q \geq 0$  a.e. on  $I$ . Moreover the right-hand side  $F(v)$  makes sense (interpreting  $(f, v)$  properly) as a bounded linear functional on  $\overset{\circ}{H}^1$  with much more general  $f$  (than  $f \in L^2$ ). More precisely,  $f$  may belong to the *dual of  $\overset{\circ}{H}^1$* , the so called “negative” Sobolev space  $H^{-1}(I)$ ; e.g. the delta function is an  $H^{-1}$  “function” in 1 dimension. It is evident that such weak conditions guarantee only existence–uniqueness of  $u$  in  $\overset{\circ}{H}^1$ . To obtain more smoothness for  $u$ , i.e. to show that  $u \in H^2$  or that  $u \in C^2(\bar{I})$ , it is clear that one has to assume more smoothness for the coefficients  $p$ ,  $q$  and  $f$ . In the same vein of thought one may allow  $q(x)$  to take on negative values as long as  $q(x) \geq \beta$ ,  $\forall x \in \bar{I}$ , where  $\beta$  is such that in (2.8),  $B(v, v) \geq c_2 \|v\|_1^2$  for some  $c_2 > 0$ . (A lower bound on  $\beta$  may be easily found in terms of  $0 < \alpha = \min_{x \in I} p(x)$  and the constant  $\mu$  of Poincaré’s inequality  $\int_I v^2 \leq \mu \int_I (v')^2$ . It is not hard to see that the best such constant  $\mu$  is equal to  $\frac{1}{\lambda_1}$ , where  $\lambda_1$  is the smallest eigenvalue of the problem  $-u'' = \lambda u$  with zero Dirichlet b.c. at  $a$  and  $b$ , i.e.  $\lambda_1 = \frac{\pi^2}{(b-a)^2}$ ).

- (iii) Using (2.9) and Sobolev’s inequality we have that under our hypotheses the weak solution  $u$  belongs to  $L^\infty(I)$  and that  $\|u\|_{L^\infty} \leq c \|f\|$ . Using the elliptic regularity (2.10) and Sobolev’s inequality we can conclude that  $u' \in L^\infty$  and  $\|u'\|_{L^\infty} \leq c \|f\|$ . Finally, using the equation we conclude that  $\|u''\|_{L^\infty} \leq c \|f\|$ , i.e. that the weak solution  $u \in H^2 \cap \overset{\circ}{H}^1$  actually belongs to a space of functions with bounded generalized derivatives. For  $u \in C^2(\bar{I})$ , the classical solution, we then have

$$\max_{x \in \bar{I}} (|u| + |u'| + |u''|) \leq c \max_{x \in \bar{I}} |f|.$$

- (iv) Consider again the weak solution, i.e.  $u \in \overset{\circ}{H}^1$  solving  $B(u, v) = (f, v) \forall v \in \overset{\circ}{H}^1$  for  $f \in L^2$ . The map  $f \mapsto u$  is linear; let us denote it by  $Tf = u$ .  $T$ , the “solution operator” of the problem (2.6), is, by the above, a bounded linear operator from  $L^2$  into  $H^2 \cap \overset{\circ}{H}^1$ : we interpret (2.10) as  $\|Tf\|_2 \leq c_3 \|f\|$ .  $T$  is then the inverse of the “elliptic” operator  $Lu = -(pu')' + qu$  defined on  $H^2 \cap \overset{\circ}{H}^1$ . Note that  $T$  is defined by

$$B(Tf, v) = (f, v), \quad \forall v \in \overset{\circ}{H}^1$$

for each  $f \in L^2$ .  $T$  is actually a bounded, self-adjoint operator from  $L^2$  into  $L^2$ ; that it is bounded follows from (2.10). To see that it is self-adjoint, note that for  $f, g \in L^2$ ,  $Tg \in H^2 \cap \overset{\circ}{H}^1$  and

$$(f, Tg) = B(Tf, Tg) = B(Tg, Tf) = (g, Tf) = (Tf, g).$$

Note that  $TL = I$  on  $H^2 \cap \overset{\circ}{H}^1$  and  $LT = I$  on  $L^2$  ( $I = \text{identity}$ ). Note also that  $(Tf, f) = B(Tf, Tf) \geq c_2 \|Tf\|_1^2$ . If  $Tf = 0 \Rightarrow u = 0 \Rightarrow f = 0$ , i.e.  $T$  is positive definite.

- (v) The case of the b.v.p. with non homogeneous Dirichlet boundary conditions  $u(a) = a_1$ ,  $u(b) = a_2$  easily reverts to the problem (2.4), (2.5). Let  $\psi(x)$  be a linear function (or any other function in  $C^2(\bar{I})$ ) such that  $\psi(a) = a_1$ ,  $\psi(b) = a_2$ . Then if  $u$  is the solution of the nonhomogeneous b.v.p., the function  $v = u - \psi$  satisfies the homogeneous b.c.  $v(a) = v(b) = 0$  and the D.E.

$$-(pv')' + qv = g := f - L\psi = f + (p\psi)' - q\psi,$$

i.e. a D.E. of the same form with new R.H.S.  $g = f - L\psi \in C(\bar{I})$ .

## 2.6.2 Neumann boundary conditions.

We now consider the problem

$$-(pu')' + qu = f \quad \text{in } (a, b), \quad (2.11)$$

$$u'(a) = u'(b) = 0, \quad (2.12)$$

i.e. the (homogeneous) *Neumann b.c.* problem. Again we let  $p \in C^1(\bar{I})$ ,  $p(x) \geq \alpha > 0 \forall x \in \bar{I}$ ,  $f \in C(\bar{I})$  (or  $f \in L^2(I)$ ), but now we assume that  $q \in C(\bar{I})$  with  $q(x) \geq \beta > 0 \forall x \in \bar{I}$ . (Note possible nonuniqueness if  $q = 0$ : e.g. consider  $p = 1$ ,  $f = 0$ ,  $q = 0$ , i.e.  $-u'' = 0$ ,  $u'(a) = 0$ ,  $u'(b) = 0$ , which has any constant as its solution). It is not hard to motivate the following

**Definition** Let  $f \in C(\bar{I})$ . Then a *classical solution* of the Neumann b.v.p. (2.11), (2.12), is a function  $u(x) \in C^2(\bar{I})$  which satisfies (2.11), (2.12) in the usual sense. If  $f \in L^2(I)$ , then a *weak solution* of (2.11), (2.12), is a function  $u \in H^1(I)$  such that

$$\int_I pu'v' + \int_I quv = \int_I fv \quad \forall v \in H^1(I). \quad (2.13)$$

Following the steps of the Dirichlet b.c. case we have:

**(i) A classical solution of (2.11), (2.12) is a weak solution as well.**

Proof obvious as before (now multiply by  $v \in H^1$  and use the “natural” b.c.  $u'(a) = u'(b) = 0$  on  $u$ ; (2.13) follows).

**(ii) Existence and uniqueness of the weak solution.**

Let  $f \in L^2(I)$  and  $B(v, w) = \int_I pv'w' + \int_I qvw$ . By our hypotheses  $B(\cdot, \cdot)$  is a bilinear, symmetric form on  $H^1 \times H^1$ . (2.7) holds with the same constant  $c_1$  as on p. 52. For  $v \in H^1$  we have

$$B(v, v) = \int_I p(v')^2 + \int_I qv^2 \geq \alpha \int_I (v')^2 + \beta \int_I v^2 \geq c_2 \|v\|_1^2, \quad (2.14)$$

where  $c_2 = \min\{\alpha, \beta\} > 0$ . Since  $F(v) = \int_I fv$  is a b.l.f. on  $H^1$  there follows, from the Lax–Milgram theorem that there exists a weak solution  $u \in H^1$  of (2.13) and that  $\|u\|_1 \leq \frac{1}{c_2} \|f\|$ . (Note that since  $u \in H^1$  we cannot give meaning to the point values  $u'(a), u'(b)$  unless we prove that the weak solution is in  $H^2$ , something that we do in the next step).

**(iii) Regularity of the weak solution.**

For  $f \in L^2(I)$  the weak solution  $u$  satisfies

$$\int_I pu'v' = \int_I (f - qu)v \quad \forall v \in H^1.$$

Hence, a fortiori,

$$\int_I pu'\phi' = \int_I (f - qu)\phi \quad \forall \phi \in C_c^\infty(I).$$

It follows, as on p. 53, that  $pu' \in H^1$  and that  $-(pu')' + qu = f$  holds in  $L^2$ . Since  $u' = \frac{1}{p}(pu')$ , it follows, as on p. 53, that  $u \in H^2$ . Returning now to

$$\int_I pu'v' = \int_I (f - qu)v \quad \forall v \in H^1,$$

since  $pu' \in H^1$  and  $v \in H^1$ , integration by parts gives (note that we can assign meaning to the values  $u'(a), u'(b) \equiv$  the values of the continuous representative of  $u' \in H^1$  at  $a, b$ ) that

$$\int_I (-(pu')'v + quv - fv) dx + p(b)u'(b)v(b) - p(a)u'(a)v(a) = 0$$

holds, for each  $v \in H^1$ . Since we already saw that  $-(pu')' + qu = f$  in  $L^2$ , it follows that  $p(b)u'(b)v(b) - p(a)u'(a)v(a) = 0 \quad \forall v \in H^1 \Rightarrow u'(b) = 0, u'(a) = 0$ . (Choose e.g.



$v(x) = x - a$  or  $v(x) = x - b$ . Hence the weak solution (2.11), (2.12), which exists uniquely by Lax–Milgram in  $H^1$ , actually belongs to  $H^2$ , satisfies  $-(pu')' + qu = f$  in  $L^2$  and  $u'(a) = u'(b) = 0$ . Moreover, exactly as before we obtain  $\|u\|_2 \leq c\|f\|$  (“elliptic regularity”).

Assume now  $f \in C(\bar{I})$ . Then exactly as in the Dirichlet b.c. case, use of Sobolev’s theorem gives that  $u \in C^2(\bar{I})$  for the weak solution.

**(iv) If  $f \in C(\bar{I})$ , the weak solution is classical.**

The proof that  $-(pu')' + qu = f$  a.e.  $\Rightarrow -(pu')' + qu = f$  everywhere in  $I$ , i.e. that the weak solution (which is  $C^2$  if  $f \in C(\bar{I})$ ) is classical, is identical to the one of the Dirichlet b.c. case. Of course  $u'(a) = u'(b) = 0$  already for the weak solution (in  $H^2$ ).

**Remarks.**

- (i) The weak solution is characterized now (by the Rayleigh–Ritz theorem) as the solution of the minimization problem

$$J(u) = \min_{v \in H^1} J(v), \quad J(v) = \frac{1}{2} \int_I p(v')^2 + qv^2 - \int_I f v, \quad v \in H^1.$$

This follows from the symmetry of  $B$ .

- (ii) Analogous remarks to the Dirichlet b.c. remarks (ii)–(v) follow, mutatis mutandis.
- (iii) We can consider of course other homogeneous b.v. problems for the equation  $-(pu')' + qu = f$  on  $(a, b)$  as well. For example the problem with b.c. (assume Dirichlet b.c. assumptions on  $p, q$ )

$$u(a) = 0, \quad u'(b) = 0, \quad (\text{“mixed” b.c.})$$

has the following weak formulation: Let

$$\overset{0}{H}_a \equiv \{v \in H^1(I) : v(a) = 0\}.$$

Clearly  $\{\overset{0}{H}_a, \|\cdot\|_1\}$  is a Hilbert space (a closed subspace of  $H^1(I)$  that includes  $\overset{0}{H}^1$ ). We then seek  $u \in \overset{0}{H}_a$  such that

$$B(u, v) \equiv \int_I pu'v' + \int_I quv = \int_I fv \quad \text{holds } \forall v \in \overset{0}{H}_a.$$

The Lax–Milgram theorem shows existence–uniqueness of the weak solution since  $B(\cdot, \cdot)$  is bilinear, bounded and coercive on  $\overset{0}{H}_a \times \overset{0}{H}_a$ . The rest ( $f \in C(\bar{I}) \Rightarrow u$  classical etc.) follows easily.

The problem with b.c.

$$u'(a) - k u(a) = 0 \quad (k \text{ const.}), \quad u(b) = 0$$

has the following weak formulation on  $\overset{0}{H}_b \equiv \{v \in H^1(I) : v(b) = 0\}$ :

Seek  $v \in \overset{0}{H}_b$  such that

$$B_k(u, v) \equiv \int_I pu'v' + \int_I quv + p(a)k u(a)v(a) = \int_I fv \quad \forall v \in \overset{0}{H}_b.$$

It is straightforward to see that  $B_k(\cdot, \cdot)$  is bilinear, symmetric, continuous on  $\overset{0}{H}_b \times \overset{0}{H}_b$  (by Sobolev's theorem) and coercive, provided  $k \geq 0$  (or  $k < 0$  with  $|k|$  sufficiently small). Hence a weak solution  $u \in \overset{0}{H}_b$  exists and its classical analog follows.

The *periodic boundary condition* problem, i.e. the problem with b.c.'s

$$u(a) = u(b), \quad u'(a) = u'(b)$$

may be similarly treated with the following weak formulation: Seek  $u \in H_\pi^1$  (assume  $p(a) = p(b)$ )

$$\int_I (pu'v' + quv) = \int_I fv, \quad \forall v \in H_\pi^1,$$

where  $H_\pi^1 \equiv \{v \in H^1(I) : v(a) = v(b)\}$ , the so called "periodic"  $H^1$ .

Let us also remark here that with the machinery already developed in this section (and the added fact that the injection of  $H^1(I) \hookrightarrow L^2(I)$  is compact – for bounded  $I$  –) we can use the spectral theorem for the bounded, self-adjoint, positive definite operator  $T$  (cf. remark (iv), p. 55) –  $T$  is *compact*, as an operator from  $L^2$  into  $L^2$  – to prove e.g. theorems about *Sturm–Liouville eigenproblems*. For example, we may thus prove that, with  $p \in C^1(\bar{I})$ ,  $p(x) \geq \alpha > 0$ ,  $q \in C(\bar{I})$ , there exists a sequence of reals  $\{\lambda_n\}_{n=1}^\infty$  such that  $\lambda_n \rightarrow \infty$  when  $n \rightarrow \infty$ , and a complete orthonormal system  $\{\phi_n\}_{n=1}^\infty$  in  $L^2(I)$ , such that  $\phi_n \in C^2(\bar{I})$  and such that for  $n = 1, 2, 3, \dots$

$$\begin{cases} -(p\phi'_n)' + q\phi_n = \lambda_n\phi_n, & \text{in } I \\ \phi_n(a) = \phi_n(b) = 0. \end{cases}$$

The numbers  $\lambda_n$  are called the *eigenvalues* and the functions  $\phi_n(x)$  the *eigenfunctions* of the Sturm–Liouville operator  $L(\cdot) \equiv -\frac{d}{dx} \left( p \frac{d}{dx}(\cdot) \right) + q \cdot$ , with homogeneous Dirichlet boundary conditions.

# Chapter 3

## Galerkin Finite Element Methods for Two-Point Boundary-Value Problems

### 3.1 Introduction

We consider again the two-point boundary-value problem on a bounded interval  $(a, b)$  with homogeneous Dirichlet b.c.'s:

$$-(pu')' + qu = f \quad \text{in } (a, b) \equiv I, \quad (3.1)$$

$$u(a) = u(b) = 0, \quad (3.2)$$

for which we assume, as in §2.6, that  $p \in C^1(\bar{I})$ ,  $p(x) \geq \alpha > 0$  in  $\bar{I}$ ,  $q \in C(\bar{I})$ ,  $q(x) \geq 0$  on  $\bar{I}$ . We recall that if  $f \in L^2(I)$ , there exists a unique weak solution of (3.1), (3.2) satisfying

$$u \in \mathring{H}^1(I), \quad B(u, v) = (f, v) \quad \forall v \in \mathring{H}^1(I), \quad (3.3)$$

where

$$B(v, w) = \int_a^b (pv'w' + qvw), \quad v, w \in H^1(I), \quad (v, w) = \int_a^b vw.$$

In fact,  $u \in H^2 \cap \mathring{H}^1$  (suppress the  $I$  in the notation  $\mathring{H}^1(I)$  etc.) and we have the elliptic regularity estimate

$$\|u\|_2 \leq C \|f\|, \quad (3.4)$$

where  $C$  is a nonnegative constant independent of  $u$  and  $f$ . If  $f \in C(\bar{I})$ , then the (unique) weak solution is in fact in  $C^2(\bar{I})$  and solves (3.1), (3.2) in the classical sense.

The *Galerkin method* for the approximation of the weak solution, i.e. of the solution of (3.3), follows the abstract framework of §1.9. We note that  $B$  satisfies (see §2.6)

$$|B(v, w)| \leq C_1 \|v\|_1 \|w\|_1 \quad \forall v, w \in H^1(I), \quad (3.5)$$

$$B(v, v) \geq C_2 \|v\|_1^2 \quad \forall v \in \overset{\circ}{H}^1, \quad (3.6)$$

where in fact e.g.  $C_1 = \|p\|_\infty + \|q\|_\infty$ ,  $C_2 = \alpha/C_*^2$ , where  $C_*$  is the constant in the Poincaré–Friedrichs inequality. Hence, if  $S_h$ ,  $0 < h \leq 1$ , is a family of finite-dimensional subspaces of  $\overset{\circ}{H}^1$ , Galerkin's Theorem 1.4 gives that for each  $h$ , there exists a unique element  $u_h \in S_h$ , the *Galerkin approximation of  $u$  in  $S_h$*  satisfying

$$B(u_h, v_h) = (f, v_h) \quad \forall v_h \in S_h. \quad (3.7)$$

Moreover,  $u_h$  satisfies the  $H^1$ -error estimate

$$\|u - u_h\|_1 \leq \frac{C_1}{C_2} \inf_{\phi \in S_h} \|u - \phi\|_1. \quad (3.8)$$

The *discrete problem* (3.7) (discrete version of (3.3)) is equivalent to a linear system of equations. Let  $N = N_h = \dim S_h$  and  $\{\phi_i\}_{i=1}^N$  be a *basis* of  $S_h$ . Then, as we saw in Ch. 1, the coefficients  $\{c_i\}$  of  $u_h$  with respect to the basis  $\phi_i$ , i.e. the numbers  $c_i$ :

$$u_h(x) = \sum_{i=1}^N c_i \phi_i(x) \quad (3.9)$$

satisfy the linear system

$$A c = f_h \quad (3.10)$$

where  $A$  is the  $N \times N$  matrix with elements  $A_{ij} = B(\phi_j, \phi_i) = \int_a^b (p\phi_j'\phi_i' + q\phi_j\phi_i)$ ,  $1 \leq i, j \leq N$ ,  $c = [c_1, \dots, c_N]^T$ ,  $f_h = [(f, \phi_1), \dots, (f, \phi_N)]^T$ . The matrix  $A$  is symmetric and positive definite on  $\mathbb{R}^N$  due to our assumptions on  $B$  (i.e. on  $p$  and  $q$ ).

The question is, of course, how to choose the finite-dimensional subspace  $S_h$  of  $\overset{\circ}{H}^1(I)$ . The choice should be such that

- $\inf_{\phi \in S_h} \|u - \phi\|_1$  is *small* (cf. (3.8)), i.e. that we can *approximate well* elements  $u$  of  $H^2 \cap \overset{\circ}{H}^1$  by elements of  $S_h$ .

- The system (3.10) is *easy to solve*.

A classical choice (Ritz, Galerkin, “spectral methods”) is to let  $S_h \equiv S_N$  be the span of the first  $N$  eigenfunctions  $\phi_j$  of the S–L operator  $Lu = -(pu')' + qu$  with zero b.c.’s at  $x = a$  and  $x = b$ . If  $\lambda_n, \phi_n$  are the eigenvalues, resp. eigenfunctions, of this problem (cf. concluding remarks of Ch. 2), then the system (3.10) may be solved explicitly by

$$c_i = \frac{(f, \phi_i)}{\lambda_i}, \quad 1 \leq i \leq N,$$

and the Galerkin approximation  $u_h \equiv u_N = \sum_{i=1}^N c_i \phi_i(x)$  has good approximation properties; in particular

$$\inf_{\phi \in S_N} \|u - \phi\|_1 \rightarrow 0, \quad \text{as } N \rightarrow \infty \text{ for } u \in H^2 \cap \overset{\circ}{H}^1.$$

The difficulty in this approach is of course that it requires the explicit knowledge of the eigenpairs  $(\lambda_i, \phi_i)$ ,  $1 \leq i \leq N$ , which are, in general, not easy to find analytically.

Another obvious choice that will also guarantee good approximation properties is to choose  $S_h$  as the vector space of the polynomials of a fixed degree that vanish at the endpoints. The problem with this approach is that, in general, the *condition* of the linear system (3.10) will be bad. Moreover, the matrix  $A$  will be, in general, full, since the polynomial basis functions  $\phi_j$  will not be of small support in  $\bar{I}$ . Since we expect  $N$  to be large for approximability, we conclude that, in general, the system (3.10) will be very hard to solve accurately. Moreover, since  $N \sim$  degree of polynomials in  $S_h$ , we will run into problems trying to compute  $u_h$  as a polynomial function of large degree.

A *good* choice turns out to be *piecewise polynomial functions* (consisting of polynomials – of small degree – on each subinterval of a partition of  $I$ ) continuous on  $I$  and endowed with *basis functions of small support*. In this way, we obtain various *finite element* subspaces  $S_h$  of  $\overset{\circ}{H}^1(I)$ .

## 3.2 The Galerkin–finite element method with piecewise linear, continuous functions

Let  $a = x_0 < x_1 < x_2 < \dots < x_N < x_{N+1} = b$  define an arbitrary partition of  $[a, b]$  with  $N$  interior points  $x_i$ ,  $1 \leq i \leq N$ . Let  $h_i = x_{i+1} - x_i$ ,  $0 \leq i \leq N$  and put  $h = \max_i h_i$ .

(For uniform partitions  $h = h_i = (b-a)/(N+1)$ ). Let  $S_h$  be the vector space of functions defined by

$$S_h = \{ \phi : \phi \in C[a, b], \phi(a) = \phi(b) = 0, \phi \text{ is a linear polynomial on each } (x_i, x_{i+1}), 0 \leq i \leq N \} \quad (3.11)$$

(we write the last condition as  $\phi|_{(x_i, x_{i+1})} \in \mathbb{P}_1$ ).

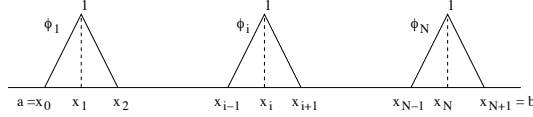
It is not hard to see that  $S_h$  is a finite dimensional subspace of  $H^1$  and that  $\dim S_h = N$ . The latter follows e.g. from the fact that the set of functions  $\{\phi_i\}_{i=1}^N$  defined by

$$\{ \phi_i \in S_h, \phi_i(x_j) = \delta_{ij} \},$$

or, explicitly, for  $1 \leq i \leq N$ , as

$$\phi_i(x) = \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}}, & \text{if } x_{i-1} \leq x \leq x_i \\ \frac{x_{i+1}-x}{x_{i+1}-x_i}, & \text{if } x_i \leq x \leq x_{i+1} \\ 0, & \text{if } x \in \bar{I} \setminus (x_{i-1}, x_{i+1}) \end{cases} \quad (3.12)$$

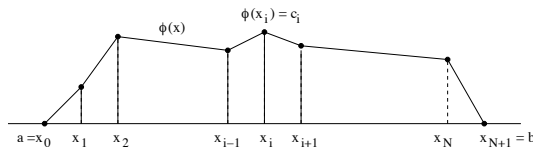
forms a *basis* of  $S_h$ : Obviously  $\phi_i \in S_h$ .



Moreover, if  $\sum_{i=1}^N c_i \phi_i(x) = 0 \forall x \in I$ , putting  $x = x_j$ ,  $1 \leq j \leq N$ , we see that  $c_j = 0$ , i.e. that  $\{\phi_i\}$  are linearly independent. In addition, since for each  $\phi \in S_h$ ,

$$\phi(x) = \sum_{i=1}^N c_i \phi_i(x), \text{ where } c_i = \phi(x_i),$$

i.e. since each  $\phi$  in  $S_h$  is uniquely determined by its *nodal values*  $\phi(x_i)$ ,  $1 \leq i \leq N$ , we conclude that  $\{\phi_i\}$  spans  $S_h$ .



Evaluating the elements  $A_{ij} = \int_a^b (p\phi_j'\phi_i' + q\phi_j\phi_i)$  of the matrix  $A$  of the system (3.10) yields that (since  $\text{supp}\phi_i = [x_{i-1}, x_{i+1}]$ )  $A_{ii} \neq 0$ ,  $1 \leq i \leq N$ ,  $A_{i,i+1} \neq 0$ ,  $A_{i+1,i} \neq 0$ ,  $1 \leq i \leq N-1$ , while all other  $A_{ij} = 0$ . Hence, due to the small support of the basis functions, the matrix  $A$  is *sparse*; in this case *tridiagonal*. Since it is also symmetric and positive definite, the system (3.10) may be efficiently solved e.g. by a *banded* (tridiagonal) *Cholesky* algorithm. In the special case  $p = q = 1$  and a uniform partition of  $I = (0, 1)$  with  $h = 1/(N+1)$ ,  $A$  is the sum of the *stiffness* matrix  $S$ , where  $S_{ij} = \int_0^1 \phi_j'\phi_i'$ , and the *mass* matrix  $M$ , where  $M_{ij} = \int_0^1 \phi_j\phi_i$ . Indeed,  $S$  and  $M$  are then the tridiagonal, symmetric and positive definite matrices

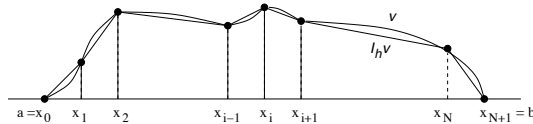
$$S = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & \mathbf{0} \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ \mathbf{0} & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}, \quad M = \frac{h}{6} \begin{pmatrix} 4 & 1 & & & \mathbf{0} \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ \mathbf{0} & & 1 & 4 & 1 \\ & & & 1 & 4 \end{pmatrix}.$$

We now return to the error estimate (3.8). As  $h \rightarrow 0$ , i.e. as  $\dim S_h \rightarrow \infty$ , we would like to prove that  $\inf_{\phi \in S_h} \|u - \phi\|_1 \rightarrow 0$ , where  $u \in H^2 \cap \overset{\circ}{H}^1$  is the weak solution of (3.1), (3.2). In fact, we shall prove that there is a constant  $C$ , independent of  $h$  and  $v$ , such that  $\inf_{\phi \in S_h} \|v - \phi\|_1 \leq Ch \|v''\|$  for any  $v \in H^2 \cap \overset{\circ}{H}^1$ .

A simple and convenient way to do this is to study the properties of the *interpolant* of  $v$  in the space  $S_h$ .

The interpolant  $(I_h v)(x)$  of a continuous function  $v(x)$  (such that  $v(a) = v(b) = 0$ ) in the space  $S_h$  is defined as the (unique) element of  $S_h$  satisfying

$$(I_h v)(x_i) = v(x_i), \quad 1 \leq i \leq N, \quad (3.13)$$



i.e. as the element  $(I_h v)(x) = \sum_{i=1}^N v(x_i)\phi_i(x)$  of  $S_h$ . Hence  $I_h : \overset{\circ}{H}^1 \rightarrow S_h$  is a linear operator on  $\overset{\circ}{H}^1$  (the *interpolation* operator onto  $S_h$ ).



**Lemma 3.1.** Let  $v \in \overset{\circ}{H}^1$ . Then  $I_h v$  satisfies

$$(i) \quad ((I_h v - v)', \varphi') = 0 \quad \forall \varphi \in S_h, \quad (3.14)$$

$$(ii) \quad \|v - I_h v\| \leq C h \|(v - I_h v)'\| \quad (3.15)$$

for some constant  $C$  independent of  $h$  and  $v$ .

**Proof.** (i) Let  $\varphi \in S_h$ . Then

$$\begin{aligned} ((I_h v - v)', \varphi') &= \int_a^b (I_h v - v)' \varphi' dx = \sum_{i=0}^N \int_{x_i}^{x_{i+1}} (I_h v - v)' \varphi' dx = \\ &= \sum_{i=0}^N \left\{ [(I_h v - v) \varphi']_{x=x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} (I_h v - v) \varphi'' dx \right\} = 0, \end{aligned}$$

since  $(I_h v)(x_i) = v(x_i)$ ,  $0 \leq i \leq N + 1$ , and  $\varphi''|_{[x_i, x_{i+1}]} = 0$ .

(ii) Since  $(I_h v - v)(x_i) = 0$ ,  $0 \leq i \leq N + 1$ ,  $I_h v - v \in \overset{\circ}{H}^1(x_i, x_{i+1})$  for each  $i = 0, 1, \dots, N$ . By the proof of Proposition 2.3 (the Poincaré–Friedrichs inequality), we conclude that there is a constant  $C$ , independent of the  $x_i$  or  $v$  such that

$$\|I_h v - v\|_{L^2(x_i, x_{i+1})} \leq C (x_{i+1} - x_i) \|(I_h v - v)'\|_{L^2(x_i, x_{i+1})}$$

for all  $i$  (in fact the best value of  $C$  is  $1/\pi$ ). We conclude that

$$\begin{aligned} \|I_h v - v\|^2 &= \sum_{i=0}^N \|I_h v - v\|_{L^2(x_i, x_{i+1})}^2 \leq C^2 h^2 \sum_{i=0}^N \|(I_h v - v)'\|_{L^2(x_i, x_{i+1})}^2 = \\ &= C^2 h^2 \|(I_h v - v)'\|^2 \quad \text{q.e.d.} \end{aligned}$$

□

Using Lemma 3.1 we may now prove the following approximation properties of the interpolant  $I_h v$ :

**Proposition 3.1.** There is a constant  $C$  independent of  $h$  such that

$$(i) \quad \|v - I_h v\| + h\|(v - I_h v)'\| \leq C h \|v'\| \quad \forall v \in \overset{\circ}{H}^1, \quad (3.16)$$

$$(ii) \quad \|v - I_h v\| + h\|(v - I_h v)'\| \leq C h^2 \|v''\| \quad \forall v \in H^2 \cap \overset{\circ}{H}^1. \quad (3.17)$$

**Proof:** (i) For  $v \in \overset{\circ}{H}^1$ ,  $\|(v - I_h v)'\| \leq \|v'\| + \|(I_h v)'\|$ . But, by (3.14),  $((I_h v)', \varphi') = (v', \varphi') \quad \forall \varphi \in S_h$ . Put  $\varphi = I_h v$  and obtain

$$\|(I_h v)'\|^2 = (v', (I_h v)') \leq \|v'\| \|(I_h v)'\|, \quad \text{i.e.} \quad \|(I_h v)'\| \leq \|v'\|.$$

We conclude that  $\|(v - I_h v)'\| \leq 2\|v'\|$ . Hence, by (3.15)

$$\|v - I_h v\| \leq C h \|(v - I_h v)'\| \leq 2 C h \|v'\|,$$

proving (3.16).

(ii) Let now  $v \in H^2 \cap \mathring{H}^1$ . Then,

$$\|(v - I_h v)'\| \leq C h \|v''\| \quad (3.18)$$

for some constant  $C$  independent of  $h$  and  $v$ . To see this, we have

$$\begin{aligned} \|(v - I_h v)'\|^2 &= ((v - I_h v)', (v - I_h v)') = (v', (v - I_h v)') \quad (\text{why?}) \\ &= \int_a^b v' (v - I_h v)' = [v' (v - I_h v)]_a^b - \int_a^b v'' (v - I_h v) \\ &\quad (\text{since } v' \in H^1, v - I_h v \in H^1) \\ &= - \int_a^b v'' (v - I_h v) \leq \|v''\| \|v - I_h v\| \stackrel{\text{by(3.15)}}{\leq} C h \|v''\| \|(v - I_h v)'\|. \end{aligned}$$

We conclude therefore that  $\|(v - I_h v)'\| \leq C h \|v''\|$ . This gives, in view of (3.15),  $\|v - I_h v\| \leq C h^2 \|v''\|$ . Hence (3.17) holds q.e.d.  $\square$

In view of Proposition 3.1, we now have, since

$$\inf_{\phi \in S_h} (\|v - \phi\| + h \|(v - \phi)'\|) \leq \|v - I_h v\| + h \|(v - I_h v)'\|,$$

that

$$\inf_{\phi \in S_h} (\|v - \phi\| + h \|(v - \phi)'\|) \leq C_k h^k \|D^k v\| \quad (3.19)$$

for  $k = 1, 2$  if  $v \in H^k \cap \mathring{H}^1$ .

Hence,

$$\inf_{\phi \in S_h} \|v - \phi\|_1 \leq C h \|v''\| \quad (3.20)$$

for  $v \in H^2 \cap \mathring{H}^1$ .

We are now ready to prove *error estimates* in the  $H^1$  and the  $L^2$  norms for the error  $u - u_h$  of the Galerkin approximation  $u_h \in S_h$  of  $u$ , the weak solution of (3.1), (3.2).

**Theorem 3.1.** *Let  $u$  be the weak solution of (3.1), (3.2) and  $u_h \in S_h$  its Galerkin approximation in  $S_h$ . Then,*

$$\|u - u_h\|_1 \leq C h \|u''\| \quad (3.21)$$

$$\text{and} \quad \|u - u_h\| \leq C h^2 \|u''\| \quad (3.22)$$

for some constant  $C$  independent of  $h$  and  $u$ .

**Proof:** (3.21) follows from (3.8) and (3.20). The proof of (3.22) follows from the following *duality* argument (Nitsche trick). Set  $e = u - u_h \in \overset{\circ}{H}^1$ . Let  $\psi$  be the solution of the problem

$$B(\psi, v) = (e, v) \quad \forall v \in \overset{\circ}{H}^1, \quad (3.23)$$

i.e. the weak solution of the problem  $-(p\psi)' + q\psi = e$  in  $(a, b)$ ,  $\psi(a) = \psi(b) = 0$ . We know that  $\psi$  exists uniquely in  $\overset{\circ}{H}^1$  and in fact  $\psi \in H^2 \cap \overset{\circ}{H}^1$  and that it satisfies (elliptic regularity)

$$\|\psi\|_2 \leq C \|e\|. \quad (3.24)$$

Put  $v = e$  in (3.23); then

$$\|e\|^2 = (e, e) = B(\psi, e) = B(e, \psi) = B(u - u_h, \psi) = B(e, \psi - \chi)$$

for **any**  $\chi \in S_h$  (why?). Take  $\chi = I_h\psi$ . Then

$$\begin{aligned} \|e\|^2 &= B(e, \psi - I_h\psi) \leq C_1 \|e\|_1 \|\psi - I_h\psi\|_1 \stackrel{\text{by(3.17)}}{\leq} \\ &C_1 \|e\|_1 C h \|\psi''\| \leq C' h \|e\|_1 \|e\|, \quad (\text{by (3.24)}). \end{aligned}$$

Hence,  $\|e\| \leq C' h \|e\|_1$  and (3.22) follows from (3.21).  $\square$

**Remarks:**

**1.** Property (3.14) characterizes uniquely the interpolant as an element of  $S_h$ . I.e. the equation

$$((v_h - v)', \varphi') = 0 \quad \forall \varphi \in S_h \quad (3.25)$$

(given  $v \in \overset{\circ}{H}^1$ ), has a unique solution  $v_h \in S_h$ , which, by Lemma 3.1 (i), coincides with the interpolant  $I_h v$  of  $v$ . To prove uniqueness, note that (3.25) may be written as  $(v_h', \varphi') = (v', \varphi') \quad \forall \varphi \in S_h$ . Hence, if there existed two such elements  $v_h^i$ , we would have

$$((v_h^1 - v_h^2)', \varphi') = 0 \quad \forall \varphi \in S_h \stackrel{\varphi=v_h^1-v_h^2}{\implies} \|(v_h^1 - v_h^2)'\| = 0 \implies v_h^1 = v_h^2$$

by the Poincaré–Friedrichs inequality.

Equation (3.25) also states that in our case,  $I_h v$  is the *projection* of  $v$  on  $S_h$  with respect to the inner product  $(u', v')$  on  $\overset{\circ}{H}^1$ . We conclude that

$$\|(I_h v - v)'\| = \inf_{\varphi \in S_h} \|v' - \varphi'\|.$$

Hence, by Poincaré–Friedrichs

$$\|I_h v - v\|_1 \leq C \|(I_h v - v)'\| \leq C \inf_{\varphi \in S_h} \|v - \varphi\|_1.$$

Since, obviously,  $\inf_{\varphi \in S_h} \|v - \varphi\|_1 \leq \|I_h v - v\|_1$ , we finally obtain that

$$\inf_{\varphi \in S_h} \|v - \varphi\|_1 \leq \|I_h v - v\|_1 \leq C \inf_{\varphi \in S_h} \|v - \varphi\|_1,$$

i.e. that  $I_h v$  is a *quasi-optimal* approximation of  $v$  in  $S_h$  (just as the Galerkin solution  $u_h$  is a quasi-optimal approximation of  $u$  in  $S_h$ ). Hence, nothing essentially was lost by using the upper bound  $\|I_h u - u\|_1$  of  $\inf_{\varphi \in S_h} \|u - \varphi\|_1$  in (3.8) – which led to (3.21).

**2.** It may be proved that the exponents of  $h$  in the estimates (right hand sides)

$$\begin{aligned} \|I_h u - u\|_1 &\leq C_1 h \|u''\| \\ \|I_h u - u\| &\leq C_2 h^2 \|u''\| \\ \|u_h - u\|_1 &\leq C'_1 h \|u''\| \\ \|u_h - u\| &\leq C'_2 h^2 \|u''\|, \end{aligned}$$

i.e. the powers 1 (for  $H^1$  norms) and 2 (for  $L^2$  norms) in the errors of  $I_h u$  and  $u_h$  (viewed as approximations of  $u \in H^2 \cap \overset{0}{H}{}^1$ ) are *optimal*, i.e. they cannot be increased in general.

We may also derive error estimates in *other norms*. For example, if the solution  $u$  of (3.1), (3.2) belongs to  $C^2[a, b]$  (if it is a classical solution, that is), we may prove that there exists a constant  $C$ , independent of  $h$  and  $u$  such that

$$\|u_h - u\|_\infty + h \|u'_h - u'\|_\infty \leq C h^2 \|u''\|_\infty.$$

(Again, the interpolant satisfies a similar estimate. E.g. it is obvious (Lagrange interpolation) that  $\|I_h u - u\|_\infty \leq (h^2/8) \|u''\|_\infty$ ).

**3. Other boundary conditions.** Let us consider for example (3.1) with the *Neumann* b.c.

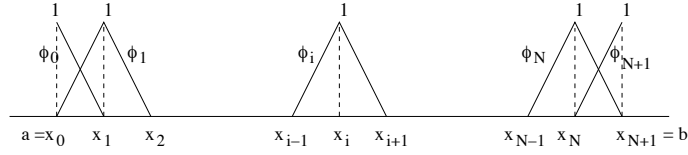
$$u'(a) = u'(b) = 0, \tag{3.26}$$

where we assume now  $q(x) \geq \beta > 0$ ,  $x \in [a, b]$  for uniqueness. As the space in which the weak solution lies now is  $H^1$ , we consider subspaces of  $H^1$  in which we seek the Galerkin approximation  $u_h$  of  $u$ .

Let  $a = x_0 < x_1 < \dots < x_{N+1} = b$  be an arbitrary partition of  $[a, b]$  and let  $h = \max_i(x_{i+1} - x_i)$  as before. The finite-dimensional space

$$\tilde{S}_h = \{\phi : \phi \in C[a, b], \phi|_{(x_i, x_{i+1})} \in \mathbb{P}_1\}$$

is a subspace of  $H^1$  and has the basis  $\{\phi_i\}_{i=0}^{N+1}$  consisting of the “hat” functions  $\phi_i$ ,  $1 \leq i \leq N$ , that were defined by (3.12), plus the end-point “half-hat” functions  $\phi_0$ ,  $\phi_{N+1}$



defined as follows:

$$\phi_0 \in \tilde{S}_h, \phi_0(x_0) = 1, \phi_0(x_j) = 0, j \neq 0, \phi_{N+1}(x_{N+1}) = 1, \phi_{N+1}(x_j) = 0, j \neq N + 1.$$

Hence,  $\dim \tilde{S}_h = N + 2$ .

Defining again the interpolant  $\tilde{I}_h v$  of a continuous function  $v$  on  $[a, b]$  by the formula

$$(\tilde{I}_h v)(x) = \sum_{i=0}^{N+1} v(x_i) \phi_i(x),$$

i.e. as the unique element of  $\tilde{S}_h$  that satisfies

$$(\tilde{I}_h v)(x_i) = v(x_i), 0 \leq i \leq N + 1,$$

we may prove again that:

$$(i) \quad ((\tilde{I}_h v)', \phi') = (v', \phi') \quad \forall v \in H^1, \phi \in \tilde{S}_h. \quad (3.27)$$

(The proof is identical to that of Lemma 3.1 (i))

$$(ii) \quad \|(\tilde{I}_h v)'\|^2 + \|(v - \tilde{I}_h v)'\|^2 = \|v'\|^2 \quad \forall v \in H^1. \quad (3.28)$$

(This follows from (i) by putting  $\phi = \tilde{I}_h v$ )

With these in mind we may now see that

$$\|(v - \tilde{I}_h v)'\| \leq \|v'\|, \quad \forall v \in H^1. \quad (3.29)$$

(A simple consequence of (3.28)). Since for each  $i$ ,  $v - \tilde{I}_h v \in \overset{\circ}{H}^1((x_i, x_{i+1}))$ , we have again, as in Lemma 3.1, that

$$\|v - \tilde{I}_h v\| \leq h \|(v - \tilde{I}_h v)'\| \quad \forall v \in H^1. \quad (3.30)$$

As a consequence of (3.29) and (3.30),

$$\|v - \tilde{I}_h v\| \leq h \|v'\| \quad \forall v \in H^1. \quad (3.31)$$

Now, we have that for  $v \in H^2$ ,

$$\|(v - \tilde{I}_h v)'\|^2 = -(v - \tilde{I}_h v, v''). \quad (3.32)$$

The proof of (3.32) is identical to that given for  $I_h$  after the estimate (3.18).

As a consequence of (3.32), we have

$$\|v' - (\tilde{I}_h v)'\|^2 \leq \|v - \tilde{I}_h v\| \|v''\|,$$

which, when combined with (3.30) yields

$$\|v' - (\tilde{I}_h v)'\| \leq h \|v''\| \quad \forall v \in H^2. \quad (3.33)$$

Using (3.33) in (3.30) we see that for  $v \in H^2$

$$\|v - \tilde{I}_h v\| \leq h^2 \|v''\|. \quad (3.34)$$

We conclude therefore that

$$\|v - \tilde{I}_h v\| + h \|(v - \tilde{I}_h v)'\| \leq C_1 h \|v'\| \quad \forall v \in H^1 \quad (3.35)$$

and

$$\|v - \tilde{I}_h v\| + h \|(v - \tilde{I}_h v)'\| \leq C_2 h^2 \|v''\| \quad \forall v \in H^2. \quad (3.36)$$

These inequalities are the analogs of (3.16) and (3.17) and lead e.g. to the estimate

$$\inf_{\phi \in \tilde{S}_h} \|v - \phi\|_1 \leq \|v - \tilde{I}_h v\|_1 \leq C h \|v''\|, \quad v \in H^2. \quad (3.37)$$

The Galerkin approximation  $u_h$  of  $u$  in  $\tilde{S}_h$  is defined again as the (unique) element of  $\tilde{S}_h$  that satisfies

$$B(u_h, \phi) = (f, \phi) \quad \forall \phi \in \tilde{S}_h,$$

and may be found again in the form  $u_h = \sum_{i=0}^{N+1} c_i \phi_i$ , where  $c = [c_0, \dots, c_{N+1}]^T$  is the solution of the linear system  $A c = f_h$ , where the  $(N+2) \times (N+2)$  (symmetric, positive definite) tridiagonal matrix  $A$  is given by  $A_{ij} = B(\phi_j, \phi_i)$ , and  $f_h$  is the vector  $[(f, \phi_0), \dots, (f, \phi_{N+1})]^T$ . As before, if  $u$  is the weak solution of (3.1), (3.26), then

$$\|u - u_h\|_1 \leq C \inf_{\phi \in S_h} \|u - \phi\|_1,$$

which yields, in view of (3.37) the  $H^1$ -error estimate

$$\|u - u_h\|_1 \leq C h \|u''\|, \quad (3.38)$$

since  $u \in H^2$ . The error bound ( $O(h)$ ) is of optimal rate.

The  $L^2$  error estimate (of optimal rate as well)

$$\|u - u_h\| \leq C h^2 \|u''\| \quad (3.39)$$

follows again by a duality argument, exactly as in the proof of (3.22) in Theorem 3.1. (Use is made of the elliptic regularity inequality for the Neumann problem).

Analogously, we may approximate the solution of other two-point boundary-value problems with (3.1) as the underlying differential equation. For example, with the boundary conditions  $u(a) = 0$ ,  $u'(b) = 0$  we use the finite element space (defined on our usual partition):

$$\{\phi \in C[a, b], \phi(a) = 0, \phi|_{(x_i, x_{i+1})} \in \mathbb{P}_1\},$$

which is a finite-dimensional subspace of the Hilbert space  $\{v : v \in H^1(a, b), v(a) = 0\}$  etc.

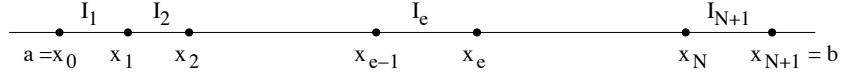
#### 4. A note on implementation. (From Brenner–Scott, pp. 10-12)

A basic computational problem in the finite element method is the evaluation, or *assembly*, of the matrix and the right-hand side of the system  $A c = f_h$  of the Galerkin equations, and in general, of the inner product  $(f, v)$  and the bilinear form

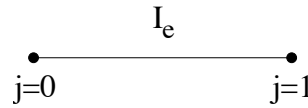
$$B(v, w) = \int_a^b (p(x)v'(x)w'(x) + q(x)v(x)w(x)) dx$$

given  $v, w \in S_h$ . We will discuss this problem in its present simple 1-dimensional context; the strategy extends in a natural way to the multidimensional case.

Given the partition  $a = x_0 < x_1 < x_2 < \dots < x_N < x_{N+1} = b$  of  $[a, b]$  we will simply refer to a subinterval  $I_k = [x_{k-1}, x_k]$  as an *element*. Hence in our problem we have  $N + 1$  elements,



the subintervals  $I_e = [x_{e-1}, x_e]$ ,  $e = 1, 2, \dots, N + 1$ . Each element  $I_e$



has two *nodes*, its endpoints, corresponding to the values of the *local node index*  $j = 0$  (for the left endpoint  $x_{e-1}$ ) and  $j = 1$  (for the right endpoint  $x_e$ ). The *element number*  $e$  and the local index  $j$  determine the *global index*  $i$  for the node  $x_i$ . In our case we have

$$i = i(e, j) = e + j - 1, \quad e = 1, 2, \dots, N + 1, \quad j = 0, 1.$$

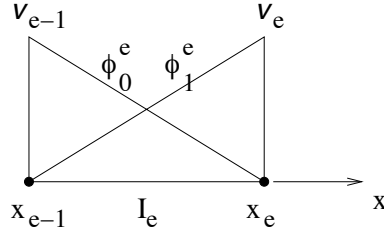
Thus the left endpoint of the element  $I_{15}$  has global index  $i = 15 + 0 - 1 = 14$ , i.e. it is the node  $x_{14}$ , which of course coincides with the right endpoint of the element  $I_{14}$  (given by  $e=14, j=1$ ).

Let us consider the basis  $\{\varphi_i\}_{i=0}^{N+1}$  of “hat” functions, introduced as basis of the subspace  $\tilde{S}_h$  used for the Neumann problem, i.e. the basis for the finite element space of piecewise linear, continuous functions with no boundary conditions imposed. Each function  $v(x)$  in this subspace may be written as

$$v(x) = \sum_{i=0}^{N+1} v(x_i) \varphi_i(x) = \sum_{i=0}^{N+1} v_i \varphi_i(x), \quad v_i \equiv v(x_i).$$

There is a particularly effective way of describing  $v(x)$  (for the purposes of efficient assembly of  $(f, v)$  and  $B(v, w)$ ) in terms of its nodal values  $v(x_i)$  and the “*local*” *basis functions*  $\varphi_j^e$ ,  $j = 0, 1$ . The latter are simply the restrictions on  $[x_{e-1}, x_e]$  of the basis functions  $\varphi_{e-1}(x)$ ,  $\varphi_e(x)$ , respectively, and may be described in terms of two fixed





functions  $\Phi_0, \Phi_1$  on the “reference” element  $[0,1]$ , as follows: First define  $\Phi_0, \Phi_1$  as

$$\Phi_0(y) = \begin{cases} 1 - y & \text{if } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}, \quad \Phi_1(y) = \begin{cases} y & \text{if } 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Associated with  $I_e = [x_{e-1}, x_e]$  (defined by the *affine* transformation  $x = h_e y + x_{e-1}$ ,  $0 \leq y \leq 1$ , where  $h_e = x_e - x_{e-1}$ , that maps the reference element  $[0,1]$  onto  $I_e$ ), define the local basis functions  $\varphi_0^e, \varphi_1^e$  as

$$\varphi_0^e(x) = \Phi_0(y), \quad \varphi_1^e(x) = \Phi_1(y), \quad \text{whenever } x = h_e y + x_{e-1},$$

i.e. as

$$\varphi_j^e(x) = \Phi_j\left(\frac{x - x_{e-1}}{h_e}\right), \quad j = 0, 1.$$

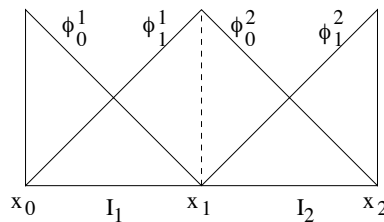
(Consequently,  $\varphi_j^e(x) = 0$  for  $x \notin I_e$ ).

A function  $v(x)$  in the finite element subspace may be described now as

$$v(x) = \sum_{e=1}^{N+1} \sum_{j=0}^1 v_{i(e,j)} \varphi_j^e(x). \quad (3.40)$$

If  $x \in (x_{e-1}, x_e)$  (3.40) reduces to

$$v(x) = v_{e-1} \varphi_0^e(x) + v_e \varphi_1^e(x),$$



which is the correct description of  $v(x)$  restricted to  $I_e$ . However, (3.40) is not a correct expression at the nodes. For example, at  $x = x_1$ , the right-hand side of (3.40) is equal to

$$v_{i(1,1)}\varphi_1^1(x_1) + v_{i(2,0)}\varphi_0^2(x_1) = v_1 + v_1 = 2v_1,$$

instead of correct value  $v_1 = v(x_1)$ . This inconsistency will not affect the assembly of  $B(v, w)$  or  $(f, v)$ , which involve *integrals* over the  $I_e$ 's.

The assembly e.g. of  $B(v, w)$  (for  $v, w$  in the finite element subspace) is now done as follows. We write

$$B(v, w) = \sum_e B_e(v, w), \quad (3.41)$$

where  $B_e(v, w)$ , is a locally defined bilinear expression given by

$$B_e(v, w) = \int_{I_e} pv'w' + qvw = \int_{x_{e-1}}^{x_e} (p(x)v'(x)w'(x) + q(x)v(x)w(x)) dx$$

where  $' = \frac{d}{dx}$ .

Using (3.40) we write

$$\begin{aligned} B_e(v, w) &= \int_{x_{e-1}}^{x_e} p(x) \left( \sum_j v_{i(e,j)} \varphi_j^e(x) \right)' \left( \sum_j w_{i(e,j)} \varphi_j^e(x) \right)' dx \\ &+ \int_{x_{e-1}}^{x_e} q(x) \left( \sum_j v_{i(e,j)} \varphi_j^e(x) \right) \left( \sum_j w_{i(e,j)} \varphi_j^e(x) \right) dx. \end{aligned}$$

Using the map  $x \mapsto y = (x - x_{e-1})/h_e$  we may transform the integrals above to integrals over the reference element  $[0,1]$ :

$$\begin{aligned} B_e(v, w) &= \frac{1}{h_e} \int_0^1 p(h_e y + x_{e-1}) \left( \sum_j v_{i(e,j)} \Phi_j(y) \right)' \left( \sum_j w_{i(e,j)} \Phi_j(y) \right)' dy \\ &+ h_e \int_0^1 q(h_e y + x_{e-1}) \left( \sum_j v_{i(e,j)} \Phi_j(y) \right) \left( \sum_j w_{i(e,j)} \Phi_j(y) \right) dy, \end{aligned}$$

where now the prime  $'$  denotes differentiation with respect to  $y$ . Letting  $\tilde{p}_e(y) = p(h_e y + x_{e-1})$ ,  $\tilde{q}_e(y) = q(h_e y + x_{e-1})$  ( $\tilde{p}_e$  and  $\tilde{q}_e$  depend on  $e$ ), we may rewrite the above in matrix-vector form

$$B_e(v, w) = \frac{1}{h_e} (v_{i(e,0)}, v_{i(e,1)}) S_e \begin{pmatrix} w_{i(e,0)} \\ w_{i(e,1)} \end{pmatrix} + h_e (v_{i(e,0)}, v_{i(e,1)}) M_e \begin{pmatrix} w_{i(e,0)} \\ w_{i(e,1)} \end{pmatrix}, \quad (3.42)$$

where  $S_e$  is the  $2 \times 2$  *local stiffness matrix* given by

$$S_e = \begin{pmatrix} \int_0^1 \tilde{p}_e(\Phi'_0)^2 & \int_0^1 \tilde{p}_e \Phi'_0 \Phi'_1 \\ \int_0^1 \tilde{p}_e \Phi'_1 \Phi'_0 & \int_0^1 \tilde{p}_e(\Phi'_1)^2 \end{pmatrix} = \int_0^1 \tilde{p}_e(y) dy \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

and  $M_e$  is the  $2 \times 2$  *local mass matrix* defined as

$$M_e = \begin{pmatrix} \int_0^1 \tilde{q}_e \Phi_0^2 & \int_0^1 \tilde{q}_e \Phi_0 \Phi_1 \\ \int_0^1 \tilde{q}_e \Phi_1 \Phi_0 & \int_0^1 \tilde{q}_e \Phi_1^2 \end{pmatrix}.$$

The formula (3.42) is quite effective for computing the local bilinear form  $B_e(v, w)$ . It reduces to a few matrix – vector operations and the evaluation of the integrals in the entries of  $S_e$  and  $M_e$ . These integrals are evaluated on the reference finite element  $[0, 1]$ , in practice by some *numerical integration rule*. For example, use of the *trapezoidal rule*

$$\int_0^1 f(y) dy \approx \frac{f(0) + f(1)}{2}$$

yields an approximate local stiffness matrix  $\tilde{S}_e$  given by

$$\tilde{S}_e = \frac{p(x_{e-1}) + p(x_e)}{2} \cdot \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

and a *diagonal* (“lumped”) local mass matrix  $\tilde{M}_e$  (why?)

$$\tilde{M}_e = \frac{1}{2} \cdot \begin{pmatrix} q(x_{e-1}) & 0 \\ 0 & q(x_e) \end{pmatrix}.$$

(It may be proved that using e.g. the trapezoidal rule in computing the elements of the matrix and the right–hand side of the Galerkin system  $A c = f_h$ , yields a new Galerkin approximation  $\tilde{u}_h$  that satisfies again  $\|\tilde{u}_h - u\|_1 = O(h)$ ,  $\|\tilde{u}_h - u\| = O(h^2)$ , i.e. the same type optimal–rate error estimates).

### 3.3 An indefinite problem

(After Schatz, Math. Comp. 28 (1974), 959-962).

Let  $k > 0$  be a constant and  $f \in L^2(a, b)$ . Consider the model two–point boundary–value problem for the “one–dimensional” Helmholtz equation, given by

$$-u'' - k^2 u = f(x), \quad a < x < b, \quad (3.43)$$

$$u(a) = u(b) = 0. \quad (3.44)$$

The weak formulation of this problem is to find  $u \in \overset{\circ}{H}^1$  so that

$$(u', v') - k^2(u, v) = (f, v) \quad \forall v \in \overset{\circ}{H}^1 \quad (3.45)$$

is satisfied. The corresponding Galerkin–finite element approximation is to seek  $u_h \in S_h$ , where  $S_h$  a finite–dimensional subspace of  $\overset{\circ}{H}^1$  (for definiteness let us take  $S_h$  as the space of continuous, piecewise linear functions on the arbitrary partition  $a = x_0 < x_1 < x_2 < \dots < x_{N+1} = b$ , that vanish at  $x = a$  and  $b$ ), satisfying

$$(u'_h, \phi') - k^2(u_h, \phi) = (f, \phi) \quad \forall \phi \in S_h. \quad (3.46)$$

The bilinear form  $B(u, v) \equiv (u', v') - k^2(u, v)$  corresponding to this problem is still continuous on  $H^1 \times H^1$  but since  $B(v, v) = \|v'\|^2 - k^2 \|v\|^2$ , it is not coercive (elliptic) on  $H^1$ . Hence the Lax–Milgram theorem for the existence–uniqueness of the weak solution of (3.43)–(3.44) cannot be used. In fact, a unique solution may even fail to exist. This happens e.g. if  $f = 0$  and  $k^2$  is one of the eigenvalues of  $-\frac{d^2}{dx^2}$  with zero b.c. at  $x = a$  and  $x = b$ , i.e. one of the numbers  $\lambda_n \equiv \frac{n^2\pi^2}{(b-a)^2}$ ,  $n = 1, 2, 3, \dots$ . However if  $k^2 \neq \lambda_n$  (equivalently if  $f = 0 \Rightarrow u = 0$ ) we can easily see, by solving (3.43)–(3.44) using the Green’s function integral representation of the solution, that a unique classical solution exists. More generally, we can state the following lemma:

**Lemma 3.2.** *Suppose that for  $f = 0$ , (3.45) has only the trivial solution  $u = 0$  in  $\overset{\circ}{H}^1$ . Then for each  $f \in L^2$ , there exists a unique solution  $u \in \overset{\circ}{H}^1$  of (3.45), which, actually, belongs to  $H^2 \cap \overset{\circ}{H}^1$  and satisfies*

$$\|u\|_2 \leq C_1(k) \|f\|. \quad (3.47)$$

□

(In the sequel we shall denote constants that depend on  $k$  by  $C_1(k)$ ,  $C_2(k)$ ,  $\dots$ . Constants independent of  $k$  will be generically denoted by  $C$ .)

We proceed now to analyze the Galerkin approximation  $u_h \in S_h$  of  $u$ . Because the bilinear form  $B(\cdot, \cdot)$  of the problem is indefinite, the bilinear system that must be solved to determine  $u_h$  in terms of its coefficients with respect to the usual hat function basis  $\{\phi_i\}_{i=1}^N$  may not have a unique solution (its matrix is symmetric but indefinite). We shall show however that if  $h$  is sufficiently small, then, provided that  $k^2 \neq \lambda_n$ ,

(3.46) has a unique solution  $u_h \in S_h$  which converges to  $u$  at the optimal rates in  $\overset{\circ}{H}^1$  and  $L^2$ , as  $h \rightarrow 0$ . We will prove the following result:

**Theorem 3.2.** *Suppose that  $f$  and  $u$  satisfy the conditions of Lemma 3.2. Then, there exists a constant  $C_2(k)$  such that if  $C_2(k)h < 1$ , (3.46) has a unique solution  $u_h \in S_h$ . Moreover, for constants  $C_3(k)$ ,  $C_4(k)$  we have*

$$\|u - u_h\|_1 \leq C_3(k)h \|u''\| \quad (3.48)$$

$$\text{and} \quad \|u - u_h\| \leq C_4(k)h^2 \|u''\|. \quad (3.49)$$

**Proof.** Under our hypothesis, (3.45) has a unique solution  $u$ . Suppose for the moment that (3.46) has a solution  $u_h$ . Set  $e = u - u_h \in \overset{\circ}{H}^1$ . Then, subtracting (3.45) and (3.46) yields

$$(e', \phi') - k^2(e, \phi) = 0 \quad \forall \phi \in S_h. \quad (3.50)$$

Now

$$\begin{aligned} \|e'\|^2 - k^2\|e\|^2 &= (e', e') - k^2(e, e) = (e', u' - u'_h) - k^2(e, u - u_h) = \text{(by (3.50))} \\ &= (e', u') - k^2(e, u). \end{aligned}$$

Hence, by the Cauchy–Schwarz inequality we have

$$\|e'\|^2 - k^2\|e\|^2 \leq \|e'\| \|u'\| + k^2\|e\| \|u\|.$$

Applying in the right hand side of the above, the elementary inequality  $\alpha\beta \leq \frac{1}{2}\alpha^2 + \frac{1}{2}\beta^2$ , we see that

$$\|e'\|^2 - k^2\|e\|^2 \leq \frac{1}{2}\|e'\|^2 + \frac{1}{2}\|u'\|^2 + \frac{k^2}{2}\|e\|^2 + \frac{k^2}{2}\|u\|^2,$$

from which we obtain that

$$\|e'\|^2 - 3k^2\|e\|^2 \leq \|u'\|^2 + k^2\|u\|^2. \quad (3.51)$$

Now, still under the hypothesis that  $u_h$  exists, by a duality argument we may prove that there is a constant  $\tilde{C}(k)$  such that

$$\|e\| \leq \tilde{C}(k)h \|e'\|. \quad (3.52)$$

Indeed, let  $\psi \in \overset{\circ}{H}^1$  be the (unique) solution of the problem

$$(\psi', v') - k^2(\psi, v) = (e, v) \quad \forall v \in \overset{\circ}{H}^1. \quad (3.53)$$

By Lemma 3.2  $\psi \in H^2 \cap \overset{\circ}{H}^1$  and

$$\|\psi\|_2 \leq C_1(k)\|e\|. \quad (3.54)$$

Putting  $v = e$  in (3.53) we see that

$$\begin{aligned} (e, e) &= (\psi', e') - k^2(\psi, e) = (e', \psi') - k^2(e, \psi) = \text{(see (3.50))} \\ &= (e', \psi' - (I_h\psi)') - k^2(e, \psi - I_h\psi), \end{aligned}$$

where  $I_h\psi$  is the  $S_h$ -interpolant of  $\psi$ . By the estimates of §3.2 we see that

$$\|\psi - I_h\psi\| \leq C h^2 \|\psi''\|, \quad \|\psi' - (I_h\psi)'\| \leq C h \|\psi''\|,$$

for a constant  $C$  independent of  $h$  and  $\psi$  (and  $k$ ). Therefore, by the above

$$\|e\|^2 \leq C h \|e'\| \|\psi''\| + k^2 \|e\| C h^2 \|\psi''\| \leq C h \|e'\| \|\psi''\| (1 + C k^2 h),$$

where use was made of the Poincaré inequality. Hence by (3.54), assuming  $h < 1$ , we have e.g.

$$\|e\|^2 \leq C C_1(k) \|e\| (1 + C k^2) \|e'\| h,$$

whence  $\|e\| \leq \tilde{C}(k) h \|e'\|$ , i.e. that (3.52) holds with  $\tilde{C}(k) = C(1 + C k^2)C_1(k)$ .

We combine now (3.51) and (3.52) and obtain

$$\|e'\|^2 \leq \|u'\|^2 + k^2 \|u\|^2 + 3k^2 (\tilde{C}(k))^2 h^2 \|e'\|^2,$$

i.e.

$$\left(1 - 3k^2 \tilde{C}(k)^2 h^2\right) \|e'\|^2 \leq \|u'\|^2 + k^2 \|u\|^2. \quad (3.55)$$

Let now  $C'_2(k) = \sqrt{3} k \tilde{C}(k)$ . It follows from (3.55) that if  $h$  is sufficiently small, so that  $C'_2(k) h < 1$ , then

$$\|e'\| \leq \frac{1}{\beta'} (\|u'\|^2 + k^2 \|u\|^2), \quad (3.56)$$

where  $0 < \beta' \equiv 1 - (C'_2(k) h)^2$ .

After all these preliminary estimates, we now prove that the linear system represented by the equations (3.46) has a unique solution. For this purpose, consider the homogeneous system associated with (3.46), i.e. the equations

$$(u'_h, \phi') - k^2(u_h, \phi) = 0 \quad \forall \phi \in S_h, \quad (3.57)$$

and suppose that (3.57) has a solution  $u_h \in S_h$ . Let  $f = 0$ . Then, by our hypothesis that (3.45) has a unique solution,  $u = 0$ . Let  $h$  be sufficiently small so that  $C'_2(k)h < 1$ , i.e.  $\beta' > 0$ . Then, since  $u_h$ , a solution of (3.57), may be considered as a Galerkin approximation of  $u$ , all estimates leading to (3.56) hold, and (3.56) gives  $\|e'\| \leq 0$ , which implies that  $e = 0$ , i.e.  $u_h = 0$ , since  $u = 0$ . We conclude that under the hypothesis  $C'_2(k)h < 1$ , the homogeneous system of equations represented by (3.57) has only the trivial solution. Therefore the nonhomogeneous system (3.46) has, for each  $f$ , a unique solution  $u_h$ , the nodal values of which may be computed e.g. by Gauss elimination with pivoting of the matrix–vector form of (3.46).

We now turn to the proof of the error estimates (3.48) and (3.49). Note that (3.48) and (3.52) imply (3.49) with  $C_4(k) = C_3(k)\tilde{C}(k)$ . Hence, it suffices to prove (3.48).

We have

$$\|e'\|^2 = \|(u - u_h)'\|^2 = \|u' - u'_I + u'_I - u'_h\|^2 \leq 2\|u' - u'_I\|^2 + 2\|u'_I - u'_h\|^2, \quad (3.58)$$

where  $u_I$  denotes the  $S_h$ -interpolant of  $u$ , and where use was made of the elementary inequality  $(\alpha + \beta)^2 \leq 2\alpha^2 + 2\beta^2$ .

To estimate the second term in the right-hand side of (3.58), note that (3.50) with  $\phi = u_I - u_h$  gives

$$\begin{aligned} (u' - u'_h, u'_I - u'_h) - k^2(u - u_h, u_I - u_h) &= 0 \Rightarrow \\ ((u' - u'_I) + (u'_I - u'_h), u'_I - u'_h) - k^2((u - u_I) + (u_I - u_h), u_I - u_h) &= 0 \Rightarrow \\ (u' - u'_I, u'_I - u'_h) + \|u'_I - u'_h\|^2 - k^2(u - u_I, u_I - u_h) - k^2\|u_I - u_h\|^2 &= 0. \end{aligned}$$

Hence

$$\begin{aligned} \|u'_I - u'_h\|^2 - k^2\|u_I - u_h\|^2 &= -(u' - u'_I, u'_I - u'_h) + k^2(u - u_I, u_I - u_h) \\ &\leq \|u' - u'_I\| \|u'_I - u'_h\| + k^2\|u - u_I\| \|u_I - u_h\| \\ &\leq \frac{1}{2} \|u' - u'_I\|^2 + \frac{1}{2} \|u'_I - u'_h\|^2 + \\ &+ \frac{k^2}{2} \|u - u_I\|^2 + \frac{k^2}{2} \|u_I - u_h\|^2. \end{aligned}$$

Hence, we obtain an “analog” of (3.51), namely that

$$\|u'_I - u'_h\|^2 - 3k^2 \|u_I - u_h\|^2 \leq \|u' - u'_I\|^2 + k^2 \|u - u_I\|^2.$$

Hence,

$$\begin{aligned} \|u'_I - u'_h\|^2 &\leq 3k^2 \|u_I - u_h\|^2 + \|u' - u'_I\|^2 + k^2 \|u - u_I\|^2 \\ &\leq 3k^2 \|u_I - u + u - u_h\|^2 + \|u' - u'_I\|^2 + k^2 \|u - u_I\|^2 \\ &\leq 6k^2 \overbrace{\|u - u_h\|^2}^e + \|u' - u'_I\|^2 + 7k^2 \|u - u_I\|^2 \\ &\stackrel{(3.52)}{\leq} 6(\tilde{C}(k))^2 k^2 h^2 \|e'\|^2 + \|u' - u'_I\|^2 + 7k^2 \|u - u_I\|^2. \end{aligned} \quad (3.59)$$

We use now (3.59) in (3.58), obtaining:

$$\|e'\|^2 \leq 12k^2 (\tilde{C}(k))^2 h^2 \|e'\|^2 + 4\|u' - u'_I\|^2 + 14k^2 \|u - u_I\|^2 \Rightarrow$$

$$\begin{aligned} \left(1 - 12k^2 (\tilde{C}(k))^2 h^2\right) \|e'\|^2 &\leq 4\|u' - u'_I\|^2 + 14k^2 \|u - u_I\|^2 \\ &\leq C h^2 \|u''\|^2 + C k^2 h^4 \|u''\|^2 \quad (\text{error of the interpolant}) \\ &\leq C (1 + k^2) h^2 \|u''\|^2 \quad (\text{if } h < 1); \end{aligned}$$

where  $C$  is independent of  $h$  and  $u$ .

Letting now  $C_2(k) = 2\sqrt{3} k \tilde{C}(k)$  (note that  $C_2 = 2C'_2$ ), and supposing that  $C_2(k)h < 1$  – which is slightly stronger than  $C'_2(k)h < 1$ , the inequality needed for existence of  $u_h$  – we have that

$$\|e'\|^2 \leq \frac{1}{\beta} C (1 + k^2) h^2 \|u''\|^2,$$

where  $\beta = 1 - 12k^2 (\tilde{C}(k))^2 h^2 > 0$ , i.e.

$$\|e'\| \leq C_3(k) h \|u''\|,$$

where  $C_3 = (C (1 + k^2)/\beta)^{1/2}$ .

Hence (3.48) is proved. □



### 3.4 Approximation by Hermite cubic functions and cubic splines

In this section we will construct two examples of finite dimensional subspaces of  $H^1(I)$  (or  $\overset{\circ}{H}^1(I)$ ),  $I = (a, b)$ , consisting of piecewise *cubic* polynomials. The Galerkin approximation  $u_h$  of  $u$ , the solution of (3.1),(3.2) or of (3.1),(3.24), will have higher order of accuracy. For example, it's  $L^2$  error  $\|u - u_h\|$  will have an  $O(h^4)$  bound, provided  $u$  is in  $H^4(I)$ .

#### 3.4.1 Hermite, piecewise cubic functions

On  $I = [a, b]$  we consider, for simplicity, the uniform partition of meshlength  $h = \frac{b-a}{N+1}$ , defined by  $x_i = a + ih$ ,  $i = 0, \dots, N + 1$ , and the associated vector space of functions

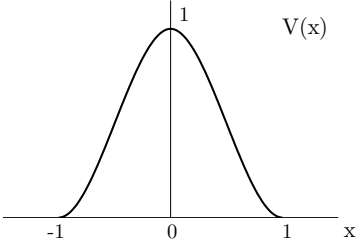
$$\mathcal{H} = \{\phi : \phi \in C^1[a, b], \phi \text{ is a cubic polynomial on each } (x_i, x_{i+1}), 0 \leq i \leq N\}.$$

The space  $\mathcal{H}$  is called the vector space of *Hermite, piecewise cubic functions* on  $I$ , relative to the partition  $\{x_i\}$ .

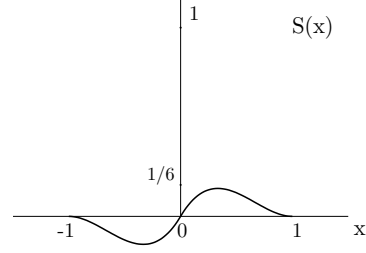
It is not hard to see that  $\mathcal{H}$  is a  $(2N + 4)$ -dimensional subspace of  $H^2(I)$ . To construct a suitable *basis* for  $\mathcal{H}$  we define two functions,  $V$  and  $S$  on  $[-1, 1]$  as follows:

$$\begin{aligned} V &\in C^1[-1, 1], & S &\in C^1[-1, 1], \\ V|_{[-1,0]}, V|_{[0,1]} &\in \mathbb{P}_3, & S|_{[-1,0]}, S|_{[0,1]} &\in \mathbb{P}_3, \\ V(-1) = 0, V'(-1) = 0, V(0) = 1, & & S(-1) = 0, S'(-1) = 0, S(0) = 0, \\ V'(0) = 0, V(1) = 0, V'(1) = 0, & & S'(0) = 1, S(1) = 0, S'(1) = 0. \end{aligned}$$

Since a cubic polynomial is uniquely defined by its values and the values of its derivative at two points, it follows that the functions  $V$  and  $S$  exist uniquely. In fact, they are given by the following formulas:

$$V(x) = \begin{cases} (x+1)^2(-2x+1) & -1 \leq x \leq 0, \\ (x-1)^2(2x+1) & 0 \leq x \leq 1, \end{cases}$$


$$S(x) = \begin{cases} x(x+1)^2 & -1 \leq x \leq 0, \\ x(x-1)^2 & 0 \leq x \leq 1, \end{cases}$$



Using  $V$  and  $S$  we define the functions  $\{V_j, S_j\}$ ,  $0 \leq j \leq N+1$ , on  $\bar{I}$  as follows:

For  $j = 1, \dots, N$  we let:

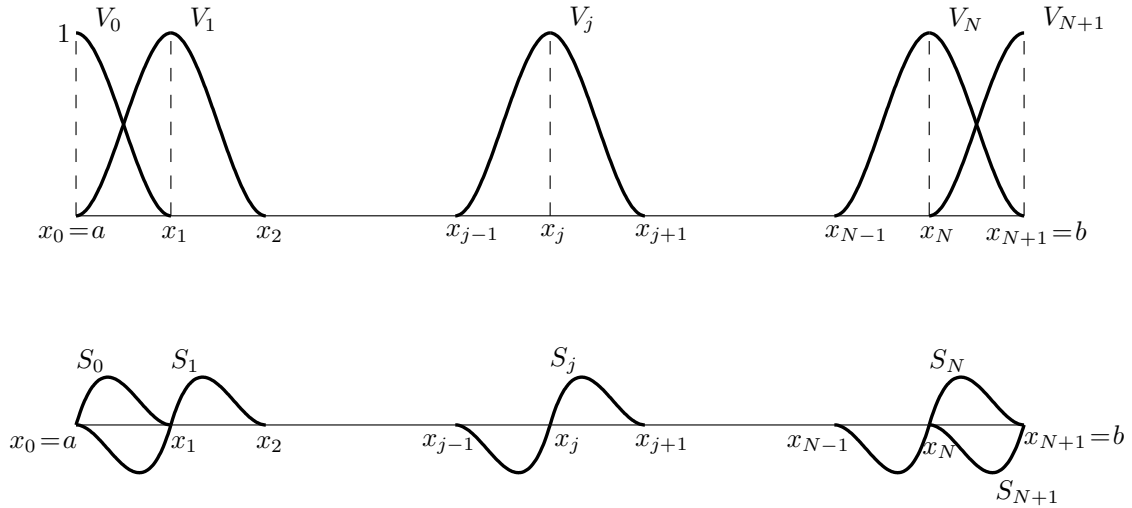
$$V_j(x) = \begin{cases} V\left(\frac{x-x_j}{h}\right) & x_{j-1} \leq x \leq x_{j+1} \\ 0 & \text{otherwise} \end{cases}, \quad S_j(x) = \begin{cases} hS\left(\frac{x-x_j}{h}\right) & x_{j-1} \leq x \leq x_{j+1} \\ 0 & \text{otherwise} \end{cases}.$$

We also define  $V_0, V_{N+1}, S_0, S_{N+1}$  as

$$V_0(x) = \begin{cases} V\left(\frac{x-a}{h}\right) & a \leq x \leq x_1 \\ 0 & \text{otherwise} \end{cases}, \quad S_0(x) = \begin{cases} hS\left(\frac{x-a}{h}\right) & a \leq x \leq x_1 \\ 0 & \text{otherwise} \end{cases},$$

$$V_{N+1}(x) = \begin{cases} V\left(\frac{x-b}{h}\right) & x_N \leq x \leq b \\ 0 & \text{otherwise} \end{cases}, \quad S_{N+1}(x) = \begin{cases} hS\left(\frac{x-b}{h}\right) & x_N \leq x \leq b \\ 0 & \text{otherwise} \end{cases},$$

We may easily verify that  $V_j, S_j \in \mathcal{H}$ ,  $0 \leq j \leq N+1$ , and that  $V_j(x_k) = \delta_{jk}$ ,  $V_j'(x_k) = 0$ ,  $S_j(x_k) = 0$ ,  $S_j'(x_k) = \delta_{jk}$ ,  $0 \leq j, k \leq N+1$ .



The  $2N+4$  functions  $\{V_j, S_j\}$ ,  $0 \leq j \leq N+1$ , form a *basis* of  $\mathcal{H}$ . Indeed, any function  $\phi \in \mathcal{H}$  may be written in the form

$$\phi(x) = \sum_{j=0}^{N+1} \phi(x_j) V_j(x) + \phi'(x_j) S_j(x), \quad x \in \bar{I}, \quad (3.60)$$

since, on the interval  $[x_k, x_{k+1}]$  the cubic polynomial  $\phi|_{[x_k, x_{k+1}]}$  is uniquely determined by the values  $\phi(x_k)$ ,  $\phi'(x_k)$ ,  $\phi(x_{k+1})$  and  $\phi'(x_{k+1})$ . On the other hand, the right-hand side of (3.60) is a cubic polynomial on  $[x_k, x_{k+1}]$ , whose values at  $x_k$ ,  $x_{k+1}$  are  $\phi(x_k)$ ,  $\phi(x_{k+1})$ , respectively, and whose derivative values are  $\phi'(x_k)$  and  $\phi'(x_{k+1})$  as well. In addition, since the relation

$$\sum_{j=0}^{N+1} c_j V_j(x) + d_j S_j(x) = 0,$$

implies (set  $x = x_k$ ,  $0 \leq k \leq N + 1$ ) that  $c_k = 0$  all  $k$ , and that

$$\sum_{j=0}^{N+1} c_j V_j'(x) + d_j S_j'(x) = 0.$$

(from which, similarly,  $d_j = 0$  all  $j$ ), we conclude that the set  $\{V_j, S_j\}$  is linearly independent.

Note that  $\text{supp}(V_j) = \text{supp}(S_j) = [x_{j-1}, x_{j+1}]$  for  $1 \leq j \leq N$ . Hence, the functions  $V_j, S_j$  have the minimal support possible in  $\mathcal{H}$ : A function in  $\mathcal{H}$  with support in *one* interval  $[x_k, x_{k+1}]$  is identically zero.

If we are given now a function  $f \in C^1[a, b]$ , there is a uniquely determined element  $\tilde{f} \in \mathcal{H}$  such that

$$f(x_j) = \tilde{f}(x_j), \quad f'(x_j) = \tilde{f}'(x_j), \quad 0 \leq j \leq N + 1. \quad (3.61)$$

$\tilde{f}$  is called the *cubic Hermite interpolant* of  $f$  (relative to the partition  $\{x_i\}$  of  $I$ ) and is given by the formula

$$\tilde{f}(x) = \sum_{j=0}^{N+1} f(x_j) V_j(x) + f'(x_j) S_j(x), \quad x \in \bar{I}. \quad (3.62)$$

We shall study the approximation properties of the Hermite interpolant. Denote by  $\|\cdot\|_k$  the norm on the Sobolev space  $H^k = H^k(I)$ . Then, the following theorem holds:

**Theorem 3.3.** *There exists  $C > 0$  (independent of  $h$ ), such that for each  $f \in H^m$ ,  $m = 2, 3$  or  $4$ , we have*

$$\|f - \tilde{f}\|_k \leq C h^{m-k} \|f^{(m)}\|_k \quad k = 0, 1, 2. \quad \square \quad (3.63)$$

Before proving it, let us remark that  $m \geq 2$  since, otherwise,  $f'(x_k)$  may not have a meaning. In addition, since  $\tilde{f} \in H^2$  but  $\tilde{f} \notin H^3$  in general, we take  $k \leq 2$ . If  $f$  is taken in  $H^4$ , then we have the best order of accuracy,  $O(h^{4-k})$  in  $H^k$ ,  $k = 0, 1, 2$ . This rate cannot be improved even if  $f$  is smoother.

Theorem 3.3 follows from two lemmas:

**Lemma 3.3.** *Let  $f \in H^2$  and  $k = 0$  or  $1$ . Then we have for  $1 \leq j \leq N + 1$*

$$\|D^k(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \leq h \|D^{k+1}(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \quad (3.64)$$

**Proof:** Since for  $k = 0$  or  $1$ ,  $D^k(f - \tilde{f})(x_j) = 0$ , all  $j$ , we have  $D^k(f - \tilde{f})(x) = \int_{x_{j-1}}^x (D^{k+1}(f - \tilde{f}))(t) dt$ ,  $x_{j-1} \leq x \leq x_j$ .

Hence, for  $x \in [x_{j-1}, x_{j+1}]$ , from the Cauchy-Schwartz inequality:

$$|D^k(f - \tilde{f})(x)| \leq \int_{x_{j-1}}^{x_j} |D^{k+1}(f - \tilde{f})| dt \leq \sqrt{h} \|D^{k+1}(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}.$$

Therefore

$$\|D^k(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}^2 \leq \int_{x_{j-1}}^{x_j} (D^k(f - \tilde{f}))^2 dt \leq h^2 \|D^{k+1}(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}^2,$$

Q.E.D.  $\square$

**Lemma 3.4.** *Let  $f \in H^m$ ,  $m = 2, 3$ , or  $4$ . Then*

$$\|D^2(f - \tilde{f})\| \leq h^{m-2} \|D^m f\|. \quad (3.65)$$

**Proof:** For  $1 \leq j \leq N+1$  we have the ‘orthogonality’ relation  $\int_{x_{j-1}}^{x_j} D^2(f - \tilde{f}) D^2 \tilde{f} = 0$ . (This follows from the observation that  $\int_{x_{j-1}}^{x_j} D^2(f - \tilde{f}) D^2 \tilde{f} =$  (integrating by parts and using  $D(f - \tilde{f})(x_j) = 0$  all  $j$ )  $= - \int_{x_{j-1}}^{x_j} D(f - \tilde{f}) D^3 \tilde{f} = - \left[ (f - \tilde{f})(x) D^3 \tilde{f}(x) \right]_{x=x_{j-1}}^{x_j} + \int_{x_{j-1}}^{x_j} (f - \tilde{f}) D^4 \tilde{f} = 0$ , since  $(f - \tilde{f})(x_j) = 0$ , all  $j$ , and  $D^4 \tilde{f}|_{[x_{j-1}, x_j]} = 0$  since  $\tilde{f} \in \mathbb{P}_3[x_{j-1}, x_j]$ .)

This ‘orthogonality’ relation implies, for  $f \in H^m$ ,  $m = 2, 3$ , or  $4$

$$\|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}^2 \leq \|D^{4-m}(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \|D^m f\|_{L^2(x_{j-1}, x_j)}, \quad 1 \leq j \leq N + 1 \quad (3.66)$$

To see (3.66), consider

$$\begin{aligned} \|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}^2 &= \int_{x_{j-1}}^{x_j} D^2(f - \tilde{f}) D^2(f - \tilde{f}) = (\text{‘orthogonality’}) = \\ &= \int_{x_{j-1}}^{x_j} D^2(f - \tilde{f}) D^2 f = (-1)^m \int_{x_{j-1}}^{x_j} D^{4-m}(f - \tilde{f}) D^m f, \end{aligned}$$

where the last equality is trivial for  $m = 2$ , requires one integration by parts and use of the fact that  $D(f - \tilde{f})(x_j) = 0$  all  $j$  if  $m = 3$ , and one more integration by parts and the fact that  $(f - \tilde{f})(x_j) = 0$  all  $j$ , if  $m = 4$ . The Cauchy-Schwarz inequality yields now (3.66).

If  $m = 2$ , (3.66) gives  $\|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \leq \|D^2 f\|_{L^2(x_{j-1}, x_j)}$ . Hence, squaring and summing over  $j$  we get

$$\|D^2(f - \tilde{f})\|^2 = \sum_{j=1}^{N+1} \|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}^2 \leq \sum_{j=1}^{N+1} \|D^2 f\|_{L^2(x_{j-1}, x_j)}^2 = \|D^2 f\|^2,$$

which is (3.65) for  $m = 2$ . If  $m = 3$ , (3.66) and (3.64) for  $k = 1$ , give

$$\begin{aligned} \|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}^2 &\leq \|D(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \|D^3 f\|_{L^2(x_{j-1}, x_j)} \\ &\leq h \|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \|D^3 f\|_{L^2(x_{j-1}, x_j)}. \end{aligned}$$

Hence,  $\|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \leq h \|D^3 f\|_{L^2(x_{j-1}, x_j)}$ , which gives (3.65) after squaring and summing over  $j$ . Finally, if  $m = 4$ , (3.66), and (3.64) for  $k = 0$  yield

$$\begin{aligned} \|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}^2 &\leq \|f - \tilde{f}\|_{L^2(x_{j-1}, x_j)} \|D^4 f\|_{L^2(x_{j-1}, x_j)} \\ &\leq h \|D(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \|D^4 f\|_{L^2(x_{j-1}, x_j)}. \end{aligned}$$

Again, by (3.64) for  $k = 1$ , we obtain

$$\|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}^2 \leq h^2 \|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \|D^4 f\|_{L^2(x_{j-1}, x_j)},$$

i.e.  $\|D^2(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)} \leq h^2 \|D^4 f\|_{L^2(x_{j-1}, x_j)}$ . Squaring and summing we get (3.65).

□

**Proof of Theorem 3.3:** It follows from Lemma 3.3 and 3.4 easily. For example, if  $m = 4$ , we have:

$$\begin{aligned} k = 0: \quad \|f - \tilde{f}\| &\leq (\text{Lemma 3.3}) \leq h \|D(f - \tilde{f})\| \leq (\text{Lemma 3.3}) \\ &\leq h^2 \|D^2(f - \tilde{f})\| \leq (\text{Lemma 3.4}) \leq h^4 \|D^4 f\|. \end{aligned}$$

$$\begin{aligned} k = 1: \quad \|(f - \tilde{f})'\| &\leq (\text{Lemma 3.3}) \leq h \|D^2(f - \tilde{f})\| \leq (\text{Lemma 3.4}) \\ &\leq h^3 \|D^4 f\|. \end{aligned}$$

$$\text{Hence } \|f - \tilde{f}\|_1^2 \leq ((h^4)^2 + (h^3)^2) \|D^4 f\|^2 \leq C h^6 \|D^4 f\|^2.$$

$$\begin{aligned} k = 2: \quad \|D^2(f - \tilde{f})\| &\leq (\text{Lemma 3.4}) \leq h^2 \|D^4 f\| \Rightarrow \\ \|f - \tilde{f}\|_2 &\leq C h^2 \|D^4 f\|. \end{aligned}$$

The cases  $m = 2, 0 \leq k \leq 2, m = 3, 0 \leq k \leq 2$ , also follow analogously.  $\square$

We conclude that given  $f \in H^m$  ( $m = 2, 3$ , or  $4$ ), there exists an element  $\chi \in \mathcal{H}$  (we took  $\chi = \tilde{f}$ ) such that

$$\sum_{k=0}^2 h^k \|f - \chi\|_k \leq C h^m \|D^m f\|. \quad (3.67)$$

We turn now to the Galerkin solution of the 2-pt. b.v.p. for the d.e. (3.1). Considering first Neumann boundary conditions, i.e. the b.c. (3.24), we see that  $u_h$  is the unique element of  $\mathcal{H}$  that satisfies

$$B(u_h, \phi) = (f, \phi) \quad \forall \phi \in \mathcal{H},$$

for which, as usual (taking as our Hilbert space  $H^1(I)$ ), there follows that

$$\|u - u_h\|_1 \leq C \inf_{\phi \in \mathcal{H}} \|u - \phi\|_1.$$

Hence, by (3.67) we conclude that if  $u \in H^m$ ,  $m = 2, 3$ , or  $4$ , then

$$\|u - u_h\|_1 \leq C h^{m-1} \|D^m u\|. \quad (3.68)$$

By a duality argument ('Nitsche trick') it follows, just as in the proof of Thm 3.1 (with  $H^1$  instead of  $\overset{0}{H}^1$ ), that in  $L^2$  we have

$$\|u - u_h\| \leq C h^m \|D^m u\|. \quad (3.69)$$

Hence, the (optimal) rate of accuracy one may achieve for  $u_h$  is 4 in  $L^2$  (3 in  $H^1$ ), provided  $u$ , the solution of (3.1), (3.24), is in  $H^4$ . (If for example  $p \in C^{m-1}$ ,  $q \in C^{m-2}$ ,  $f \in H^{m-2}$  for some  $m \leq 2$ , it may be shown that the elliptic regularity estimate (3.4) generalizes and gives that  $u \in H^m$  and

$$\|u\|_m \leq C_m \|f\|_{m-2}.)$$

In the case of homogeneous Dirichlet boundary conditions, i.e. the problem (3.1)-(3.2), we may take as our subspace of  $\overset{0}{H}^1(I)$  the vector space  $\overset{0}{\mathcal{H}} = \{\phi : \phi \in \mathcal{H}, \phi(a) = \phi(b) = 0\}$ . It is easily seen that  $\overset{0}{\mathcal{H}}$  is a  $(2N+2)$ -dimensional subspace of  $H^2 \cap \overset{0}{H}^1$ . Its basis consists of the elements  $V_j, 1 \leq j \leq N$ , and  $S_j, 0 \leq j \leq N+1$ . We may define the interpolant  $\tilde{f}$  of a function  $f \in H^2 \cap \overset{0}{H}^1$  in the natural way, and show as before that

$$\|f - \tilde{f}\|_k \leq C h^{m-k} \|D^m f\|, \quad k = 0, 1, 2,$$

provided  $f \in H^m \cap \overset{\circ}{H}^1$ , and  $m$  is 2, 3, or 4. The analogous estimates to (3.68) and (3.69) still hold, i.e. we have

$$\|u - u_h\|_j \leq C h^{m-j} \|D^m u\|, \quad j = 0, 1,$$

provided  $u$ , the solution of (3.1),(3.2), belongs to  $H^m \cap \overset{\circ}{H}^1$ .

The system that defines the Galerkin equations is now of size (say, for the Neumann problem)  $(2N + 4) \times (2N + 4)$ . Ordering the basis vectors  $\{\phi_1, \dots, \phi_{2N+4}\}$  as  $\{V_0, S_0, V_1, S_1, \dots, V_{N+1}, S_{N+1}\}$ , we see that e.g. the Gram matrix (with elements  $\int_a^b \phi_i \phi_j$ ) is *block-tridiagonal* with 2x2 blocks. Its general ‘line’ of blocks is:

$$\begin{array}{|c|c|} \hline \int V_j V_{j-1} & \int V_j S_{j-1} \\ \hline \int S_j V_{j-1} & \int S_j S_{j-1} \\ \hline \end{array} \quad \begin{array}{|c|c|} \hline \int (V_j)^2 & \int V_j S_j \\ \hline \int S_j V_j & \int (S_j)^2 \\ \hline \end{array} \quad \begin{array}{|c|c|} \hline \int V_j V_{j+1} & \int V_j S_{j+1} \\ \hline \int S_j V_{j+1} & \int S_j S_{j+1} \\ \hline \end{array}$$

The cubic Hermite elements will, then, give an accuracy of  $O(h^4)$  in  $L^2$  (if  $u \in H^4$ ) at the expense of solving a  $(2N + 4) \times (2N + 4)$  7-diagonal linear system.

We close this section with two remarks:

(i) We may define Hermite piecewise cubic functions on a general partition  $a = x_0 < x_1 < \dots < x_{N+1} = b$  of  $[a, b]$ . The associated basis  $\{V_j, S_j\}$ ,  $0 \leq j \leq N + 1$  may be defined in a straightforward way; its members have support in  $[x_{j-1}, x_{j+1}]$  for  $1 \leq j \leq N$ , etc. An interpolant may be defined in the natural way. The estimate (3.63) still holds with  $h := \max_j (x_{j+1} - x_j)$ .

(ii) The cubic Hermite functions are a special case of the space

$\mathcal{H}_m = \left\{ \phi \in C^{m-1}[a, b], \phi|_{[x_i, x_{i+1}]} \in \mathbb{P}_{2m-1} \right\}$ , on which the error of the Galerkin approximation is  $O(h^{2m})$  in  $L^2$ .

### 3.4.2 Cubic splines

The dimension of  $\mathcal{H}$ , the space of Hermite cubics is, as we saw,  $2N + 4$ . This, in particular, leads to rather large linear systems that have to be solved for the Galerkin approximation  $u_h$ . A natural idea is to lower the dimension of the space by requiring more continuity at the nodes  $x_j$ . In the cubic polynomial case this leads to the space

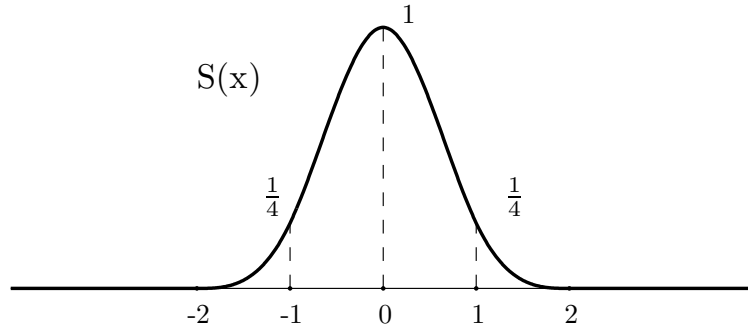
$$\mathcal{S} := \left\{ \phi : \phi \in C^2[a, b], \phi|_{[x_i, x_{i+1}]} \in \mathbb{P}_3, 0 \leq i \leq N \right\},$$

of the so-called *cubic splines*. (We suppose that the  $\{x_i\}$  define a uniform partition of  $[a, b]$  with  $x_0 = a$ ,  $x_{N+1} = b$ , of meshlength  $h = (b - a)/(N + 1)$ .)

A count of the free parameters and the constraints of functions in  $\mathcal{S}$  leads to the conjecture that  $\mathcal{S}$  is a  $(N+4)$ -dimensional subspace of  $H^3((a, b))$ . To prove this fact, we shall construct a basis of  $\mathcal{S}$ , in fact a basis with elements of *minimal support*.

It is not hard to see that there are no nontrivial elements of  $\mathcal{S}$  vanishing outside less than four adjacent mesh intervals. (For example, prove that the only element of  $\mathcal{S}$  which vanishes identically outside  $[x_{j-1}, x_{j+2}]$  is the zero element.) In addition, it is easy to see that there is a unique function  $S \in C^2[-2, 2]$ , such that  $\text{supp}(S) = [-2, 2]$ ,  $S|_{[k, k+1]} \in \mathbb{P}_3$  for  $k = -2, -1, 0, 1$ ,  $S(\pm 2) = S'(\pm 2) = S''(\pm 2) = 0$ ,  $S(0) = 1$ . This function is given by the formula

$$S(x) = \begin{cases} \frac{1}{4}(x+2)^3 & -2 \leq x \leq -1, \\ \frac{1}{4}[1 + 3(x+1) + 3(x+1)^2 - 3(x+1)^3] & -1 \leq x \leq 0, \\ \frac{1}{4}[1 + 3(1-x) + 3(1-x)^2 - 3(1-x)^3] & 0 \leq x \leq 1, \\ \frac{1}{4}(2-x)^3 & 1 \leq x \leq 2 \\ 0 & x \in \mathbb{R} - [-2, 2]. \end{cases} \quad (3.70)$$



Using  $S(x)$ , we define the functions  $\{\phi_j\}$ ,  $j = -1, 0, 1, \dots, N + 2$ , on  $[a, b]$  as follows: We introduce the extra nodes  $x_{-1} := a - h$ ,  $x_{N+2} := b + h$  and put

$$\phi_j(x) = S\left(\frac{x - x_j}{h}\right) \Big|_{[a, b]}, \quad -1 \leq j \leq N + 2,$$

(the *restriction* of  $S\left(\frac{x - x_j}{h}\right)$  to  $[a, b]$ ). It may be seen immediately that each  $\phi_j$ ,  $-1 \leq j \leq N + 2$ , is a cubic polynomial on each interval  $[x_k, x_{k+1}]$ ,  $0 \leq k \leq N$ , belongs to  $C^2[a, b]$ , and hence  $\phi_j \in \mathcal{S}$ ,  $1 \leq j \leq N + 2$ .



Moreover,  $\text{supp}(\phi_j) = [x_{j-2}, x_{j+2}]$ , for  $2 \leq j \leq N - 1$ ,

$\phi_{-1}$  vanishes for  $x \in [a, b] - [x_0, x_1]$ ,

$\phi_0$  vanishes for  $x \in [a, b] - [x_0, x_2]$ ,

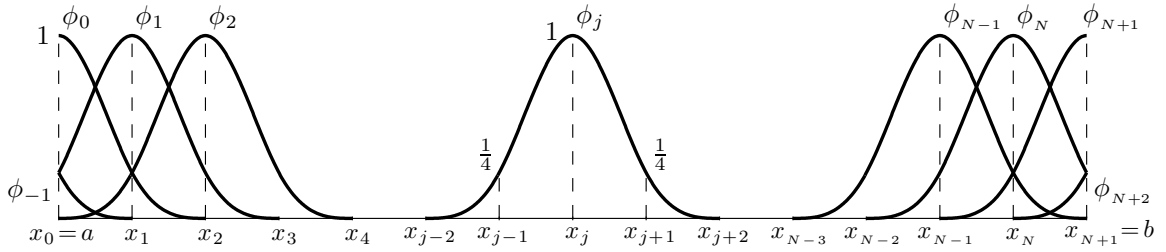
$\phi_1$  vanishes for  $x \in [a, b] - [x_0, x_3]$ ,

$\phi_N$  vanishes for  $x \in [a, b] - [x_{N-2}, x_{N+1}]$ ,

$\phi_{N+1}$  vanishes for  $x \in [a, b] - [x_{N-1}, x_{N+1}]$ ,

$\phi_{N+2}$  vanishes for  $x \in [a, b] - [x_N, x_{N+1}]$ ,

In addition,  $\phi_j(x_j) = 1$  for  $0 \leq j \leq N + 1$ .



The  $N + 4$  functions  $\phi_j$ ,  $-1 \leq j \leq N + 2$ , are known as *B-splines*, since they form a *basis* of  $\mathcal{S}$ . We shall show this in a series of lemmata:

**Lemma 3.5.** *The  $\{\phi_j\}_{j=-1}^{N+2}$  are linearly independent.*

**Proof:** Suppose that for real constants  $c_j$ ,  $-1 \leq j \leq N + 2$ , we have

$$\sum_{j=-1}^{N+2} c_j \phi_j(x) = 0, \quad x \in [a, b].$$

Then, in particular we have

$$\sum_{j=-1}^{N+2} c_j \phi_j(x_k) = 0, \quad k = 0, \dots, N + 1, \quad (3.71a)$$

$$\text{and} \quad \sum_{j=-1}^{N+2} c_j \phi_j'(x_k) = 0, \quad k = 0, \dots, N + 1. \quad (3.71b)$$

Using the facts that

$$\phi_j(x_k) = \begin{cases} 1 & \text{if } j = k, \\ \frac{1}{4} & \text{if } |j - k| = 1, \\ 0 & \text{if } |j - k| \geq 2, \end{cases}$$

we may write the relation (3.71a) as

$$\frac{1}{4}c_{k-1} + c_k + \frac{1}{4}c_{k+1} = 0, \quad k = 0, \dots, N+1. \quad (3.72)$$

Now, using the fact that  $S$  is even (about zero) and therefore that  $S'$  is an odd function, we conclude from the definition of  $\phi_j$  that

$$\phi'_j(x_j) = 0, \quad \phi'_j(x_{j-1}) = -\phi'_j(x_{j+1}), \quad \phi'_j(x_k) = 0, \quad |j - k| \geq 2.$$

Therefore, (3.71b) for  $k = 0$  gives

$$c_{-1}\phi'_{-1}(x_0) + c_1\phi'_1(x_0) = 0 \Rightarrow c_{-1} = c_1. \quad (3.73a)$$

In addition, (3.71b) for  $k = N+1$  gives

$$c_N\phi'_N(x_{N+1}) + c_{N+2}\phi'_{N+2}(x_{N+1}) = 0 \Rightarrow c_N = c_{N+2}. \quad (3.73b)$$

The relations (3.73a) and (3.73b) are used to eliminate the unknowns  $c_{-1}$  and  $c_{N+2}$  from the linear system (3.72), which takes the form

$$Ac = 0, \quad (3.74)$$

where  $c = [c_0, \dots, c_{N+1}]^T$  and  $A$  is the  $(N+2) \times (N+2)$  tridiagonal matrix

$$A = \begin{pmatrix} 4 & 2 & & & \mathbf{0} \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ \mathbf{0} & & & & 2 & 4 \end{pmatrix}.$$

The matrix  $A$  is strictly diagonally dominant and therefore is invertible by Gerschgorin's lemma. Therefore,  $c_j = 0$ ,  $0 \leq j \leq N+1$ , since  $c$  solves the homogeneous problem (3.74). By (3.73a) and (3.73b)  $c_{N+2} = 0$ ,  $c_{-1} = 0$ . We conclude that the  $\{\phi_j\}$ ,  $-1 \leq j \leq N+2$ , are linearly independent.  $\square$

We let now  $M := \langle \phi_{-1}, \dots, \phi_{N+2} \rangle$  be the subspace of  $\mathcal{S}$  which is *spanned* by the elements of the set  $\{\phi_j\}_{j=-1}^{N+2}$ . We shall show that in fact  $M = \mathcal{S}$ , thus completing the proof that  $\{\phi_j\}_{j=-1}^{N+2}$  is a basis for  $\mathcal{S}$ . This will be accomplished by two intermediate results of *interpolation*:

**Lemma 3.6.** *Let  $f \in C^1[a, b]$ . Then, there exists a unique element  $\tilde{f} \in M$  satisfying the interpolation conditions*

$$\left\{ \begin{array}{l} \tilde{f}(x_j) = f(x_j), \quad 0 \leq j \leq N + 1, \\ \tilde{f}'(a) = f'(a), \\ \tilde{f}'(b) = f'(b) \end{array} \right. \quad (3.75)$$

**Proof:**

The functions  $\{\phi_j\}_{j=-1}^{N+2}$  form a basis for  $M$ . We seek therefore a function  $\tilde{f} = \sum_{j=-1}^{N+2} c_j \phi_j \in M$  which satisfies (3.75), i.e. the linear system (for its coefficients  $c_j$ ):

$$\left. \begin{array}{l} \sum_{j=-1}^{N+2} c_j \phi_j(x_k) = f(x_k), \quad 0 \leq k \leq N + 1, \\ c_{-1} \phi'_{-1}(a) + c_1 \phi'_1(a) = f'(a), \\ c_N \phi'_N(b) + c_{N+2} \phi'_{N+2}(b) = f'(b). \end{array} \right\} \quad (3.76)$$

The linear system (3.76) has a unique solution, since the associated homogeneous system (obtained by putting  $f(x_k) = 0$ ,  $0 \leq k \leq N + 1$ ,  $f'(a) = f'(b) = 0$  in (3.76)), has only the trivial solution as we argued in the proof of Lemma 3.5. We conclude that there is a unique element  $\tilde{f} \in M$  satisfying the interpolation conditions (3.75).  $\square$

**Lemma 3.7.** *Let  $f \in C^1[a, b]$ . Then, there is a unique element  $\tilde{s} \in \mathcal{S}$  which satisfies the analogous to (3.75) interpolation conditions*

$$\left. \begin{array}{l} \tilde{s}(x_j) = f(x_j), \quad 0 \leq j \leq N + 1, \\ \tilde{s}'(a) = f'(a), \\ \tilde{s}'(b) = f'(b). \end{array} \right\} \quad (3.77)$$

**Proof:** See Akrivis-Dougalis “Introduction to Numerical Analysis”, Thm. 4.7. The function  $\tilde{s}(x)$  is explicitly constructed by the values of its second derivative at the points  $x_j$ ,  $0 \leq j \leq N + 1$ .  $\square$

We may now conclude that the  $\{\phi_i\}$  span  $\mathcal{S}$  i.e. that  $M = \mathcal{S}$ :

**Theorem 3.4.** *The  $\{\phi_j\}_{j=-1}^{N+2}$  are a basis of  $\mathcal{S}$ .*

**Proof:** In view of Lemma 3.5 we need only to show that  $\mathcal{S} \subset M$ . Let  $f \in \mathcal{S}$ . Then, by Lemma 3.6, there is a unique element  $\tilde{f} \in M$  that satisfies the interpolation conditions (3.75). But  $M \subset \mathcal{S}$ , hence  $\tilde{f} \in \mathcal{S}$ . However, by Lemma 3.7 there is a unique

element of  $\mathcal{S}$  that satisfies the interpolation conditions (3.77) which are the same as (3.75). This element coincides with  $f$  since  $f \in \mathcal{S}$ . We conclude that  $f = \tilde{f}$ . Hence  $f \in M$ , i.e.  $\mathcal{S} \subset M$ .  $\square$

Given  $f \in C^1[a, b]$ , we call  $\tilde{f}(= \tilde{s})$  its *cubic spline interpolant* (with “first derivative” end-point conditions  $\tilde{f}'(a) = f'(a)$ ,  $\tilde{f}'(b) = f'(b)$ .) It is not hard to see that if we are given  $f''(a)$ ,  $f''(b)$ , we may construct a cubic spline interpolant  $\tilde{\tilde{f}}$  satisfying, in addition to the conditions  $\tilde{\tilde{f}}(x_j) = f(x_j)$ ,  $0 \leq j \leq N + 1$ , the endpoint “second derivative” boundary conditions  $\tilde{\tilde{f}}''(a) = f''(a)$ ,  $\tilde{\tilde{f}}''(b) = f''(b)$ . Similarly, for *periodic*  $f \in C^1[a, b]$  we may construct the “periodic spline” interpolant etc. All these interpolant functions satisfy error estimates of  $O(h^4)$  accuracy in  $L^2$  provided  $f \in H^4(I)$ . For example, we shall prove the following result for the interpolant with first-derivative boundary conditions:

**Theorem 3.5.** *There exists a constant  $C > 0$  (independent of  $h$ ), such that for each  $f \in H^m$ ,  $m = 2, 3$ , or  $4$ , we have*

$$\|f - \tilde{f}\|_k \leq C h^{m-k} \|f^{(m)}\|_k, \quad k = 0, 1, 2. \quad \square \quad (3.78)$$

As before, we shall prove the theorem using two intermediate results. First, we prove the analog of Lemma 3.3; but now the norms are *global*, i.e. they are the  $L^2(a, b)$  norms.

**Lemma 3.8.** *Let  $f \in H^2$  and  $k = 0$  or  $1$ . Then, there exists a constant  $C$  independent of  $h$  and  $f$ , such that*

$$\|D^k(f - \tilde{f})\| \leq C h \|D^{k+1}(f - \tilde{f})\|. \quad (3.79)$$

*Proof:* For  $k = 0$  we may prove the local result as well. Since  $f$  coincides with  $\tilde{f}$  at the nodes  $x_j$ ,  $0 \leq j \leq N + 1$ , we have, for  $x \in [x_{j-1}, x_j]$ , for some  $1 \leq j \leq N + 1$ , that

$$f(x) - \tilde{f}(x) = \int_{x_{j-1}}^x D(f - \tilde{f})(y) dy.$$

Hence, for  $x_{j-1} \leq x \leq x_j$ , using the Cauchy-Schwarz inequality,

$$\begin{aligned} |(f - \tilde{f})(x)| &\leq \int_{x_{j-1}}^x |D(f - \tilde{f})(y)| dy \leq \sqrt{(x_j - x_{j-1})} \left( \int_{x_{j-1}}^{x_j} (D(f - \tilde{f})(y))^2 dy \right)^{\frac{1}{2}} \\ &\leq \sqrt{h} \|D(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}. \end{aligned}$$

Therefore,

$$\|f - \tilde{f}\|_{L^2(x_{j-1}, x_j)} \leq h \|D(f - \tilde{f})\|_{L^2(x_{j-1}, x_j)}, \quad (3.80)$$

from which, by squaring and summing, we get (3.79) for  $k = 0$  with  $C = 1$ .

To prove (3.79) for  $k = 1$ , we note that since

$$\phi(x_j) = 0, \quad 0 \leq j \leq N + 1,$$

where  $\phi := f - \tilde{f}$ , there follows, by Rolle's theorem, that there exist  $\xi_j \in (x_{j-1}, x_j)$ ,  $1 \leq j \leq N + 1$ , where  $\phi'(\xi_j) = 0$ .

$$\begin{array}{ccccccccccc} \xi_0 & \xi_1 & & & \xi_j & \xi_{j+1} & & & \xi_{N+1} & \xi_{N+2} & \\ \leftarrow & \leftarrow & & & \leftarrow & \leftarrow & & & \leftarrow & \leftarrow & \\ x_0 = a & x_1 & & & x_{j-1} & x_j & & & x_{N+1} & x_{N+2} = b & \end{array}$$

(Note that  $\phi \in H^2[a, b]$ ; hence,  $\phi \in C^1[a, b]$ ). Note also that, by definition of  $\tilde{f}$ ,  $\phi'(x_0) = \phi'(x_{N+1}) = 0$ . Introduce then  $\xi_0 := x_0$  and  $\xi_{N+2} := x_{N+1}$ , so that  $\phi'(\xi_j) = 0$ ,  $0 \leq j \leq N + 2$ . Therefore, for  $x \in [\xi_{j-1}, \xi_j]$ ,  $\phi'(x) = \int_{\xi_{j-1}}^x \phi''(y) dy$ . Hence  $|\phi'(x)| \leq \sqrt{(x - \xi_{j-1})} \|\phi''\|_{L^2(\xi_{j-1}, \xi_j)} \leq \sqrt{(\xi_j - \xi_{j-1})} \|\phi''\|_{L^2(\xi_{j-1}, \xi_j)} \leq \sqrt{2h} \|\phi''\|_{L^2(\xi_{j-1}, \xi_j)}$ . We conclude that

$$\|\phi'\|_{L^2(\xi_{j-1}, \xi_j)} \leq (\xi_j - \xi_{j-1})(2h) \|\phi''\|_{L^2(\xi_{j-1}, \xi_j)}^2 \leq 4h^2 \|\phi''\|_{L^2(\xi_{j-1}, \xi_j)}^2.$$

Hence,

$$\|\phi'\|^2 = \sum_{j=1}^{N+2} \|\phi'\|_{L^2(\xi_{j-1}, \xi_j)}^2 \leq 4h^2 \sum_{j=1}^{N+2} \|\phi''\|_{L^2(\xi_{j-1}, \xi_j)}^2 = 4h^2 \|\phi''\|^2.$$

We conclude that (3.79) is valid for  $k = 1$  as well, with  $C = 2$ .  $\square$

We proceed now to the analog of Lemma 3.4:

**Lemma 3.9.** *Let  $f \in H^m$ ,  $m = 2, 3$  or  $4$ . Then, for some  $C$  independent of  $h$  and  $f$ :*

$$\|D^2(f - \tilde{f})\| \leq C h^{m-2} \|D^m f\|. \quad (3.81)$$

*Proof:* First we prove the *orthogonality relation*

$$\int_a^b D^2(f - \tilde{f}) D^2 \tilde{f} = 0. \quad (3.82)$$

We have, using the endpoint condition  $f'(a) = \tilde{f}'(a)$ ,  $f'(b) = \tilde{f}'(b)$ ,

$$\begin{aligned} \int_a^b D^2(f - \tilde{f}) D^2 \tilde{f} &= \left[ (f - \tilde{f})' \tilde{f}'' \right]_a^b - \int_a^b (f - \tilde{f})' \tilde{f}''' \\ &= - \int_a^b (f - \tilde{f})' \tilde{f}'''. \quad (\text{Note that } \tilde{f} \in H^3(a, b)). \end{aligned}$$

Now  $\int_a^b (f - \tilde{f})' \tilde{f}''' = \sum_{j=0}^N \int_{x_j}^{x_{j+1}} (f - \tilde{f})' \tilde{f}''' = \sum_{j=0}^N \left\{ [(f - \tilde{f}) \tilde{f}''']_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} (f - \tilde{f}) D^4 \tilde{f} \right\} = 0$ , since  $(f - \tilde{f})(x_j) = 0$ ,  $0 \leq j \leq N + 1$ , and  $\tilde{f} \in \mathbb{P}_3$  in each  $[x_j, x_{j+1}]$ . Thus, (3.82) is proved. Using now (3.82) we have by Cauchy-Schwarz

$$\|D^2(f - \tilde{f})\|^2 = \int_a^b D^2(f - \tilde{f}) D^2(f - \tilde{f}) = \int_a^b D^2(f - \tilde{f}) D^2 f \leq \|D^2(f - \tilde{f})\| \|D^2 f\|.$$

Hence, (3.81) holds, if  $m = 2$ , with  $C = 1$ . If now  $m = 3$ , we have, integrating by parts once and using the endpoint derivative conditions:

$$\begin{aligned} \|D^2(f - \tilde{f})\|^2 &= \int_a^b D^2(f - \tilde{f}) D^2 f = \left[ D(f - \tilde{f}) D^2 f \right]_a^b - \int_a^b (f - \tilde{f})' D^3 f \\ &= - \int_a^b ((f - \tilde{f})' D^3 f \leq \|(f - \tilde{f})'\| \|D^3 f\|, \end{aligned}$$

from which, by (3.79)  $k = 1$ ,  $\|D^2(f - \tilde{f})\|^2 \leq C h \|D^2(f - \tilde{f})\| \|D^3 f\|$ , which shows that (3.81) holds if  $m = 3$ . Finally, if  $m = 4$ , we may integrate by parts once more; using the interpolation conditions  $f(a) = \tilde{f}(a)$ ,  $f(b) = \tilde{f}(b)$ , we have:

$$\begin{aligned} \|D^2(f - \tilde{f})\|^2 &= - \int_a^b (f - \tilde{f})' D^3 f = - \left[ (f - \tilde{f}) D^3 f \right]_a^b + \int_a^b (f - \tilde{f}) D^4 f \\ &= \int_a^b (f - \tilde{f}) D^4 f \leq \|f - \tilde{f}\| \|D^4 f\| \leq C h^2 \|D^2(f - \tilde{f})\| \|D^4 f\|. \end{aligned}$$

(In the last step, we used Lemma 3.8 twice). We conclude that (3.81) holds as well for  $m = 4$ .  $\square$

*Proof of Theorem 3.5:* It is straightforward to see that (3.78) follows from the results of Lemmata 3.8 and 3.9.  $\square$

We conclude that given  $f \in H^m$ , ( $m = 2, 3$ , or  $4$ ), there exists an element  $\chi \in \mathcal{S}$  (take  $\chi = \tilde{f}$ ), such that

$$\sum_{k=0}^2 h^k \|f - \chi\|_k \leq C h^m \|D^m f\|. \quad (3.83)$$

As before, it follows for the Galerkin approximation  $u_h$  to  $u$ , the solution of the b.v.p. (3.1),(3.24), that

$$\|u - u_h\|_k \leq C h^{m-k} \|D^m u\|, \quad (3.84)$$

for  $k = 0, 1$ , provided  $u \in H^m$ ,  $m = 2, 3$  or  $4$ . In the case of homogeneous Dirichlet b.c.'s, i.e. of the problem (3.1),(3.2), we take as subspace of  $\overset{0}{H}^1(I)$  the vector space

$\overset{0}{\mathcal{S}} = \{\phi : \phi \in \mathcal{S}, \phi(a) = \phi(b) = 0\}$ . It is easily seen that  $\overset{0}{\mathcal{S}}$  is a  $(N+2)$ -dimensional subspace of  $\overset{0}{H}^1 \cap H^3$ . A *basis* of this subspace consists of the previously defined B-splines  $\phi_j$ ,  $2 \leq j \leq N-1$ , plus four functions  $\tilde{\phi}_0, \tilde{\phi}_1, \tilde{\phi}_N, \tilde{\phi}_{N+1} \in \overset{0}{\mathcal{S}}$ , taken as (independent) linear combinations of the  $\phi_{-1}, \phi_0, \phi_1$  and  $\phi_N, \phi_{N+1}, \phi_{N+2}$  B-splines, which are such that  $\tilde{\phi}_0(a) = 0, \tilde{\phi}_1(a) = 0, \tilde{\phi}_N(b) = 0, \tilde{\phi}_{N+1}(b) = 0$ . For example, we take

$$\tilde{\phi}_0 := \phi_0 - 4\phi_{-1}, \quad \tilde{\phi}_1 := \phi_1 - \phi_{-1} \quad \text{etc.}$$

Given  $f \in \overset{0}{H}^1 \cap H^2$ , we construct again an interpolant  $\tilde{f} \in \overset{0}{\mathcal{S}}$  satisfying  $\tilde{f}(x_i) = f(x_i)$ ,  $0 \leq i \leq N+1$ , and  $\tilde{f}'(x_0) = f'(x_0), \tilde{f}'(x_{N+1}) = f'(x_{N+1})$ , for example. The error estimates are entirely analogous. (The linear system that defines the Galerkin equations has now a seven-diagonal, banded matrix  $B(\phi_i, \phi_j)$ .)

As before, we may prove the error estimates (3.83), (3.84) etc. for cubic splines defined on a general partition  $\{x_j\}$  of  $[a, b]$ . Also, one may define higher-order *smooth* spline spaces, as follows: For  $m \geq 2$  we let

$$\mathcal{S}_{(m)} = \left\{ \phi : \phi \in C^m[a, b], \phi|_{[x_i, x_{i+1}]} \in \mathbb{P}_{2m-1} \right\},$$

in which we may prove  $L^2$  error estimates for the Galerkin approximations of  $O(h^{2m})$ .

# Chapter 4

## Results from the Theory of Sobolev Spaces and the Variational Formulation of Elliptic Boundary–Value Problems in $\mathbb{R}^N$

In this chapter we let  $\Omega$  be a bounded domain in  $\mathbb{R}^N$  (i.e. an open, connected, bounded point set in  $\mathbb{R}^N$ ). Many of the results that we shall quote hold for arbitrary open sets in  $\Omega$  but we do not strive for generality here, having in mind applications in the case of boundary–value problems on bounded domains. Let  $x = (x_1, \dots, x_N)$  denote a generic point of  $\Omega$  or  $\mathbb{R}^N$ . All functions are real–valued.

### 4.1 The Sobolev space $H^1(\Omega)$ .

**Definition.** *The Sobolev space  $H^1 = H^1(\Omega)$  is defined by*

$$H^1(\Omega) = \{u \in L^2(\Omega) : \exists g_1, g_2, \dots, g_N \in L^2(\Omega), \text{ such that} \\ \int_{\Omega} u \frac{\partial \phi}{\partial x_i} = - \int_{\Omega} g_i \phi, \quad \forall \phi \in C_c^\infty(\Omega), \quad \forall i : 1 \leq i \leq N\}.$$

For  $u \in H^1$  we denote  $g_i = \frac{\partial u}{\partial x_i}$  and call  $g_i$  the weak (generalized) partial derivative of  $u$  with respect to  $x_i$ .

(Note e.g. that this definition does not need that  $\Omega$  be bounded).



**Remarks.**

- (i) Using Lemma 2.4 we conclude that each generalized partial derivative  $g_i$  in the above definition is unique, as in the case of one dimension.
- (ii) Again,  $C_c^1(\Omega)$  may be used in place of  $C_c^\infty(\Omega)$  for the *test functions*  $\phi$ .
- (iii) It is clear that if  $u \in C^1(\Omega) \cap L^2(\Omega)$  and if the classical partial derivatives  $\frac{\partial u}{\partial x_i}$  belong to  $L^2(\Omega)$  for  $i = 1, 2, \dots, N$ , then the weak derivatives exist and coincide with the classical; in particular  $u \in H^1(\Omega)$ . Of course  $C^1(\overline{\Omega}) \subset H^1(\Omega)$ . It can be shown with some care that, inversely, if  $u \in H^1(\Omega) \cap C(\Omega)$  and if the generalized derivatives  $\frac{\partial u}{\partial x_i}$  belong to  $C(\Omega)$  for  $1 \leq i \leq N$ , then  $u \in C^1(\Omega)$ .
- (iv) Since  $u \in L^2(\Omega) \Rightarrow u \in L_{\text{loc}}^1(\Omega)$ , we can define the *distributional* derivatives of  $u$ ,  $\frac{\partial u}{\partial x_i}$ , in the sense of the theory of distributions. We can say that  $H^1$  is the set of elements of  $L^2(\Omega)$  whose distributional derivatives  $\frac{\partial u}{\partial x_i}$ ,  $1 \leq i \leq N$ , are represented by functions in  $L^2$ .

It is clear that  $H^1$  is a subspace of  $L^2$ . Denoting by  $(\cdot, \cdot)$ ,  $\|\cdot\|$  the inner product, respectively the norm, on  $L^2(\Omega)$ , we introduce the quantities:

$$(u, v)_1 = (u, v) + \sum_{i=1}^N \left( \frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i} \right), \quad u, v \in H^1,$$
$$\|u\|_1 = \left( \|u\|^2 + \sum_{i=1}^N \left\| \frac{\partial u}{\partial x_i} \right\|^2 \right)^{\frac{1}{2}}, \quad u \in H^1,$$

which clearly define an *inner product*, resp. (the induced) *norm*, on  $H^1$ . Hence  $H^1$  is a normed linear space.

**Theorem 4.1.** *The space  $(H^1(\Omega), \|\cdot\|_1)$  is a Hilbert space.*

**Proof.** Adapt the 1-dim. proof.  $\square$

**Remark.**  $(H^1(\Omega), \|\cdot\|_1)$  is separable.

In the case of one dimension we had shown (Theorem 2.4) that the restrictions on  $\Omega = I$  of functions in  $C_c^\infty(\mathbb{R})$  form a dense set in  $H^1$ . In more than one dimensions this is not true for an arbitrary  $\Omega$ . We list below several “density” results:

1. (**Friedrichs**) Let  $u \in H^1(\Omega)$ . Then, there exists a sequence  $\{u_n\} \in C_c^\infty(\mathbb{R}^N)$  such that:

$$(\alpha) \quad u_n|_\Omega \rightarrow u \text{ in } L^2(\Omega)$$

$$(\beta) \quad \forall i : \frac{\partial u_n}{\partial x_i}|_\omega \rightarrow \frac{\partial u}{\partial x_i}|_\omega \text{ in } L^2(\omega) \text{ for every precompact } \omega \subset \Omega.$$

2. (**Meyers – Serrin**) If  $u \in H^1(\Omega)$ , then  $\exists$  sequence  $\{u_n\} \in C^\infty(\Omega) \cap H^1(\Omega)$  such that  $u_n \rightarrow u$  in  $H^1(\Omega)$ .
3. If  $\Omega$  is an arbitrary open (or even open and bounded) set and if  $u \in H^1(\Omega)$ , in general there does not exist a sequence  $u_n \in C_c^1(\mathbb{R}^N)$  such that  $u_n|_\Omega \rightarrow u$  in  $H^1(\Omega)$ . (The problem is at the boundary; however, compare with Theorem 4.2 below).

With the aid e.g. of Friedrichs' result (1), above, we can show the following analog of Proposition 2.1:

**Proposition 4.1.** *If  $u, v \in H^1(\Omega) \cap L^\infty(\Omega)$ , then  $uv \in H^1 \cap L^\infty$  and*

$$\frac{\partial}{\partial x_i}(uv) = \frac{\partial u}{\partial x_i}v + u \frac{\partial v}{\partial x_i}, \quad 1 \leq i \leq N.$$

As in the case of one dimension, many properties of  $H^1$  are established easier if  $\Omega = \mathbb{R}^N$ . Then an extension result is needed to establish the same property for an  $\Omega \subset \mathbb{R}^N$ . Such extensions (of functions i.e. of  $H^1(\Omega)$  to functions of  $H^1(\mathbb{R}^N)$ ) are not always possible to construct, unless the set  $\Omega$  has a “regular” boundary  $\partial\Omega$  in a certain sense.

**Definition.** *Let  $\Omega$  be a bounded domain in  $\mathbb{R}^N$ . We say that  $\Omega$  is of class  $C^1$  if there exist finitely many open balls  $B_i \subset \mathbb{R}^N$ ,  $i = 1, 2, \dots, M$  such that*

$$(i) \quad \partial\Omega \subset \cup_{i=1}^M B_i, \quad B_i \cap \partial\Omega \neq \emptyset.$$

(ii) *There is for each  $i$ ,  $1 \leq i \leq M$ , a function  $y = f^{(i)}(x)$  in  $C^1(\overline{B_i})$  which maps the ball  $B_i$  in an one-to-one and onto way onto a domain in  $\mathbb{R}^N$  so that  $\partial\Omega \cap \overline{B_i}$  gets mapped onto a subset of the hyperplane  $y_N = 0$  and  $\Omega \cap B_i$  into a simply connected domain in the half-space  $\{y : y_N > 0\}$ . Moreover the Jacobian determinant  $\det \left( \frac{\partial f_k^{(i)}}{\partial x_l} \right)$  does not vanish for  $x \in \overline{B_i}$ .*

The  $C^1$  property of  $\Omega$  permits for example the advertized “extension” result:

**Proposition 4.2.** *Suppose that  $\Omega$  is of class  $C^1$  (or that  $\Omega = \mathbb{R}^N_+ = \{x \in \mathbb{R}^N : x_N > 0\}$ ). Then there exists a linear extension operator  $P : H^1(\Omega) \rightarrow H^1(\mathbb{R}^N)$  such that,*

$$(i) \quad Pu|_{\Omega} = u \quad \forall u \in H^1(\Omega).$$

$$(ii) \quad \exists C \text{ such that } \|Pu\|_{L^2(\mathbb{R}^N)} \leq C \|u\|_{L^2(\Omega)}, \quad \forall u \in H^1(\Omega).$$

$$(iii) \quad \exists C' \text{ such that } \|Pu\|_{H^1(\mathbb{R}^N)} \leq C' \|u\|_{H^1(\Omega)} \quad \forall u \in H^1(\Omega).$$

Using e.g. this extension one may show the following important density result:

**Theorem 4.2.** *If  $\Omega$  is of class  $C^1$  and  $u \in H^1(\Omega)$ , then, there exists a sequence  $u_n \in C_c^\infty(\mathbb{R}^N)$  such that  $u_n|_{\Omega} \rightarrow u$  in  $H^1$ . That is to say, the restriction to  $\Omega$  of functions in  $C_c^\infty(\mathbb{R}^N)$  are dense in  $H^1(\Omega)$ .*

For the purposes of studying boundary–value problems, it is important to study the behavior of functions in  $H^1(\Omega)$  at the boundary  $\partial\Omega$  of  $\Omega$ . We suppose to this effect that  $\Omega$  is of class  $C^1$ . Then we can “measure” (in the sense of the above definition of the  $C^1$  domain) the content of pieces on the “hypersurface”  $\partial\Omega$  (through measuring “plane” surface pieces on the hyperplane  $y_N = 0$ , i.e. measuring the “area” of the images of the pieces of  $\partial\Omega$  that are mapped on  $y_N = 0$  by the functions  $y = f^{(i)}(x)$ ). One may define open sets on  $\partial\Omega$  as intersections of  $\partial\Omega$  and open sets in  $\mathbb{R}^N$ . These open sets on  $\partial\Omega$  one can then complete into a  $\sigma$ –algebra and then extend the elementary “content measure” into the Lebesgue measure on  $\partial\Omega$ . With respect to this measure we may define the surface integral  $\int_{\partial\Omega} g(y)dy$ . Consequently, one may consider the Hilbert space  $L^2(\partial\Omega)$  of functions defined on  $\partial\Omega$  with norm

$$\|g\|_{L^2(\partial\Omega)} \equiv \left( \int_{\partial\Omega} g^2(y) dy \right)^{\frac{1}{2}} \quad \text{for } g \in L^2(\partial\Omega).$$

One may show first that the following result holds for *smooth* functions:

**Lemma 4.1.** *Let  $\Omega$  be of class  $C^1$ . Then there exists a constant  $C$  such that for all functions  $f \in C^\infty(\overline{\Omega})$  we have*

$$\|f\|_{L^2(\partial\Omega)} \leq C \|f\|_{H^1(\Omega)}.$$

This lemma, along with Theorem 4.2, permits us to define boundary values for functions in  $H^1(\Omega)$  for  $C^1$  domains  $\Omega$ . Let  $f \in H^1(\Omega)$ . Then  $\exists f_n \in C^\infty(\overline{\Omega})$  such that  $f_n \rightarrow f$  in  $H^1(\Omega)$  (Theorem 4.2). By Lemma 4.1  $\{f_n\}$  is Cauchy in  $L^2(\partial\Omega)$ . So  $f_n \rightarrow g$  in  $L^2(\partial\Omega)$ . (It is easily seen that this  $g$  is independent of the chosen sequence  $f_n$ ). This limiting function  $g$  we denote by  $f|_{\partial\Omega}$  and call it “boundary value” on  $\partial\Omega$  of  $f$ , or *trace* of  $f$  on  $\partial\Omega$  (sometimes we say that  $g = f|_{\partial\Omega}$  is the boundary value of  $f$  in the sense of trace). We summarize in the following theorem.

**Theorem 4.3** (Trace theorem). *Let  $\Omega$  be of class  $C^1$ . Then, every  $f \in H^1(\Omega)$  possesses boundary values in the above sense (also denoted by  $f$ ), which belong to the Hilbert space  $L^2(\partial\Omega)$ . Moreover there exists a constant  $C$  such that*

$$\|f\|_{L^2(\partial\Omega)} \leq C \|f\|_{H^1(\Omega)} \quad \forall f \in H^1(\Omega).$$

**Remarks.**

(i) It can be shown that under our hypotheses on  $\Omega$ , the boundary value  $f|_{\partial\Omega}$  of a function  $f \in H^1(\Omega)$  actually belongs to the so-called “fractional-order” space  $H^{1/2}(\partial\Omega)$ , an intermediate space defined by interpolation between the spaces  $H^0(\partial\Omega) \equiv L^2(\partial\Omega)$  and  $H^1(\partial\Omega)$ .

(ii) Let  $\Omega$  be a  $C^1$  domain. Then we may show that *Green’s formula* (*Gauss’s theorem*) holds:  $\forall i : 1 \leq i \leq N$

$$\int_{\Omega} \frac{\partial u}{\partial x_i} v = - \int_{\Omega} u \frac{\partial v}{\partial x_i} + \int_{\partial\Omega} u v \nu_i dy$$

for  $u, v \in H^1(\Omega)$ . Here  $dy$  is the surface Lebesgue measure constructed on  $\partial\Omega$  as above and  $\nu_i = \vec{n} \cdot \vec{e}_i$  is the  $i^{\text{th}}$  component of the unit outward normal  $\vec{n}(y)$  defined on the boundary of the  $C^1$  domain  $\Omega$ . (Note that since  $u, v \in H^1(\Omega) \Rightarrow \frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i} \in L^2(\Omega)$ ,  $u, v \in L^2(\partial\Omega)$  and all terms in the above equality make sense).

## 4.2 The Sobolev space $\overset{\circ}{H}^1(\Omega)$ .

**Definition.** *We define the space  $\overset{\circ}{H}^1(\Omega)$  as the completion of  $C_c^\infty(\Omega)$  with respect to the  $H^1(\Omega)$  norm.*

Hence  $\left(\overset{\circ}{H}^1(\Omega), \|\cdot\|_1\right)$  is a Hilbert space (a closed subspace of  $H^1$ ). (We may show that the completion of  $C_c^\infty(\mathbb{R}^N)$  under the  $\|\cdot\|_{H^1(\mathbb{R}^N)}$  norm is  $H^1(\mathbb{R}^N)$  itself, i.e. that  $\overset{\circ}{H}^1(\mathbb{R}^N) = H^1(\mathbb{R}^N)$ ). But for  $\Omega \subset \mathbb{R}^N$  we have in general that  $\overset{\circ}{H}^1(\Omega) \subset H^1(\Omega)$ . More precisely, we may show that for sufficiently smooth  $\partial\Omega$  (e.g.  $C^1$ ) then  $\overset{\circ}{H}^1(\Omega)$  consists precisely of those functions in  $H^1(\Omega)$  which vanish (in the sense of trace) on  $\partial\Omega$ .

**Theorem 4.4.** *Let  $\Omega$  be of class  $C^1$ . Then*

$$\overset{\circ}{H}^1(\Omega) = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\},$$

(where by  $v|_{\partial\Omega} = 0$  we mean that the trace of  $v$  on  $\partial\Omega$  (a function in  $L^2(\partial\Omega)$ ), is equal to the zero function in  $L^2(\partial\Omega)$ ).

On  $\overset{\circ}{H}^1(\Omega)$  we also have the analog of Proposition 2.3:

**Proposition 4.3** (Inequality of Poincaré–Friedrichs). *Let  $\Omega$  be a bounded domain. Then, there exists a constant  $C_* = C_*(\Omega)$  such that*

$$\|u\| \leq C_* \left( \sum_{i=1}^N \left\| \frac{\partial u}{\partial x_i} \right\|^2 \right)^{\frac{1}{2}} \quad \forall u \in \overset{\circ}{H}^1(\Omega).$$

In particular the expression  $\left(\sum_{i=1}^N \left\| \frac{\partial u}{\partial x_i} \right\|^2\right)^{1/2}$  is a norm on  $\overset{\circ}{H}^1(\Omega)$ , equivalent to the norm  $\|\cdot\|_1$  on  $\overset{\circ}{H}^1(\Omega)$ . The quantity  $\int_{\Omega} \left(\sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i}\right) dx$  is an inner product on  $\overset{\circ}{H}^1(\Omega)$ .

**Remark.** As in the 1–dim case, if  $\Omega$  is of class  $C^1$ , then  $u \in \overset{\circ}{H}^1(\Omega)$  if and only if the extension

$$\bar{u}(x) = \begin{cases} u(x) & \text{if } x \in \Omega \\ 0 & \text{if } x \in \mathbb{R}^N \setminus \Omega \end{cases}$$

belongs to  $H^1(\mathbb{R}^N)$ . (In such a case also  $\overline{\frac{\partial u}{\partial x_i}} = \frac{\partial \bar{u}}{\partial x_i}$ ).

### 4.3 The Sobolev spaces $H^m(\Omega)$ , $m = 2, 3, 4, \dots$

For  $m \geq 2$  an integer we can define the spaces  $H^m(\Omega)$  recursively by

$$H^m = H^m(\Omega) = \{u \in H^{m-1}(\Omega) : \frac{\partial u}{\partial x_i} \in H^{m-1}(\Omega), \forall i = 1, 2, \dots, N\}.$$

We introduce some notation. A *multiindex*  $\alpha = (\alpha_1, \dots, \alpha_N)$  is an  $N$ -vector of non-negative integers  $\alpha_i \geq 0$ ,  $1 \leq i \leq N$ . If  $\alpha$  is a multiindex we let  $|\alpha| = \sum_{i=1}^N \alpha_i$ . Then, the partial derivatives of a function of  $N$  variables may be denoted by

$$D^\alpha \varphi = \frac{\partial^{|\alpha|} \varphi}{\partial x_1^{\alpha_1} \dots \partial x_N^{\alpha_N}}.$$

It follows that  $H^m$  is the set:

$$H^m = H^m(\Omega) = \left\{ u \in L^2(\Omega) : \forall \alpha \text{ with } |\alpha| \leq m, \exists g_\alpha \in L^2(\Omega) \text{ such that} \right. \\ \left. \int_{\Omega} u D^\alpha \varphi = (-1)^{|\alpha|} \int_{\Omega} g_\alpha \varphi, \forall \varphi \in C_c^\infty(\Omega) \right\}.$$

We call  $g_\alpha$  the generalized partial derivative of order  $\alpha$  of  $u$  and denote  $g_\alpha = D^\alpha u$ .

The space  $H^m$  with the norm

$$\|u\|_m = \left( \sum_{0 \leq |\alpha| \leq m} \|D^\alpha u\|^2 \right)^{\frac{1}{2}},$$

induced by the inner product

$$(u, v)_m = \sum_{0 \leq |\alpha| \leq m} (D^\alpha u, D^\alpha v)$$

is a Hilbert space. One may show that if  $\partial\Omega$  is sufficiently regular ( $C^1$  will certainly suffice), then the norm  $\|\cdot\|_m$  on  $H^m(\Omega)$  is equivalent to the norm

$$\left( \|u\|^2 + \sum_{|\alpha|=m} \|D^\alpha u\|^2 \right)^{\frac{1}{2}}.$$

In effect one may show that  $\forall \alpha$ ,  $0 < |\alpha| \leq m$  and  $\epsilon > 0$ , there exists a constant  $C = C(\alpha, \epsilon, \Omega)$  such that the *interpolation inequality*

$$\|D^\alpha u\| \leq \epsilon \sum_{|\beta|=m} \|D^\beta u\| + C \|u\|, \quad \forall u \in H^m(\Omega)$$

holds.

Now, since  $u \in H^m \Rightarrow D^\alpha u \in H^1(\Omega)$  for each  $\alpha$ :  $0 \leq |\alpha| \leq m-1$ , we can define by the trace theorem, boundary values on  $\partial\Omega$  (for  $\Omega$ , say, of class  $C^1$ ) for all derivatives  $D^\alpha u$ ,  $0 \leq |\alpha| \leq m-1$  of  $u$ . In this sense we can define e.g. the *normal derivative* on  $\partial\Omega$  of a function  $u \in H^2(\Omega)$  as the linear combination

$$\frac{\partial u}{\partial n} = \sum_{i=1}^N \frac{\partial u}{\partial x_i} \Big|_{\partial\Omega} n_i,$$

where  $\vec{n}(y)$  is the unit outer normal on  $\partial\Omega$ . For  $u \in H^2(\Omega)$ ,  $\frac{\partial u}{\partial n} \in L^2(\partial\Omega)$ . One has another formula of Green's too:

$$-\int_{\Omega} \Delta u v = \int_{\Omega} \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} - \int_{\partial\Omega} \frac{\partial u}{\partial n} v dy, \quad \forall u, v \in H^2(\Omega).$$

One may define again  $H^m_0(\Omega)$  as the completion of  $C_c^\infty(\Omega)$  in the  $\|\cdot\|_m$  norm. For  $\partial\Omega$  sufficiently smooth (e.g. for  $\Omega$  a domain of class  $C^m$  – replace  $C^1$  by  $C^m$  in the definition of  $C^1$  domain) one may show that  $H^m_0(\Omega)$  is equal to the subspace of  $H^m(\Omega)$  for which  $u \in H^m_0(\Omega) \Leftrightarrow u \in H^m(\Omega)$  and  $D^\alpha u|_{\partial\Omega} = 0$  (in the sense of trace),  $0 \leq |\alpha| \leq m-1$ . Again we note the difference between the spaces

$$H^2 \cap \dot{H}^1 = \{u \in H^2 : u|_{\partial\Omega} = 0\}$$

and

$$\dot{H}^2 = \{u \in H^2 : u|_{\partial\Omega} = \frac{\partial u}{\partial x_i}|_{\partial\Omega} = 0\}.$$

## 4.4 Sobolev's inequalities.

In one dimension we had proved that  $H^1(I) \subset C(\bar{I})$  for a bounded interval  $I$  and that  $\|u\|_{L^\infty(I)} \leq C \|u\|_{H^1(I)}$ . In more than one dimensions this is no longer true. There is a wealth of imbedding theorems of which we quote two results:

**Theorem 4.5** (Sobolev). *Let  $\Omega$  be of class  $C^1$ . Then*

(a) *If  $N = 2$ ,  $H^1 \subset L^p \forall p, 1 \leq p < \infty$ .*

*If  $N > 2$ ,  $H^1 \subset L^p$  where  $1 \leq p \leq \frac{2N}{N-2}$ .*

(b) *If  $m > \frac{N}{2}$  we have  $H^m(\Omega) \subset C^k(\bar{\Omega})$  where  $0 \leq k < m - \frac{N}{2}$  and*

$$\sup_{x \in \bar{\Omega}, 0 \leq |\alpha| \leq k} |D^\alpha u(x)| \leq C \|u\|_m \quad \forall u \in H^m(\Omega).$$

This theorem tells us that if  $N > 1$  the functions in  $H^1(\Omega)$  are no longer continuous (in the sense of a.e. equality as usual). For example if  $N = 2$  we need to go to  $H^2(\Omega)$  to obtain continuous functions in  $\bar{\Omega}$ . As a counterexample in this direction we may verify that the function

$$u = \left( \log \frac{1}{|x|} \right)^\alpha \quad \text{with } 0 < \alpha < \frac{1}{2}, \quad \Omega = \{x \in \mathbb{R}^2 : |x| < \frac{1}{2}\}$$

belongs to  $H^1(\Omega)$  but it is not bounded because of the singularity at  $x = 0$ .

## 4.5 Variational formulation of some elliptic boundary–value problems.

### 4.5.1 (a) Homogeneous Dirichlet boundary conditions.

We consider the following problem. Let  $\Omega \subset \mathbb{R}^N$  be a  $C^1$  domain. We seek a function  $u : \bar{\Omega} \rightarrow \mathbb{R}$  satisfying

$$-\Delta u + u = f \quad \text{in } \Omega, \quad \Delta = \sum_{i=1}^N \frac{\partial^2}{\partial x_i^2} \quad (4.1)$$

$$u = 0 \quad \text{on } \partial\Omega, \quad (4.2)$$

where  $f$  is a given function on  $\Omega$ . The boundary condition  $u = 0$  on  $\partial\Omega$  is called *homogeneous (zero) Dirichlet b.c.*

**Definition** Let  $f \in C(\bar{\Omega})$ . Then, a classical solution of (4.1), (4.2) is a function  $u \in C^2(\bar{\Omega})$  satisfying the P.D.E. (4.1) and the b.c. (4.2) in the usual (pointwise) sense. Let now  $f \in L^2(\Omega)$ . Then, a weak solution of (4.1), (4.2) is a function  $u \in \overset{\circ}{H}^1$  which satisfies the weak form of (4.1), (4.2), i.e. the relation

$$\int_{\Omega} \left( \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} + u v \right) dx = \int_{\Omega} f v \quad \forall v \in \overset{\circ}{H}^1. \quad (4.3)$$

**(i) A classical solution of (4.1), (4.2) is a weak solution.**

Let  $f \in C(\bar{\Omega})$  and let  $u \in C^2(\bar{\Omega})$  be a classical solution of (4.1), (4.2). Then, since  $u \in H^1(\Omega)$  and  $u = 0$  on  $\partial\Omega$ , it follows that  $u \in \overset{\circ}{H}^1$ . Multiplying  $-\Delta u + u = f$  by a function  $\varphi \in C_c^\infty(\Omega)$  we have, using Green's theorem (cf. p. 88) that

$$\int_{\Omega} \left( \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} + u \varphi \right) dx = \int_{\Omega} f \varphi$$

holds. Since now  $C_c^\infty(\Omega)$  is dense in  $(\overset{\circ}{H}^1, \|\cdot\|_1)$  (4.3) follows by the above by approximating in  $\overset{\circ}{H}^1$ ,  $v \in \overset{\circ}{H}^1(\Omega)$  by a sequence  $\varphi_i \in C_c^\infty(\Omega)$ . Hence  $u$  is a weak solution.

**(ii) Existence and uniqueness of the weak solution.**

Let  $f \in L^2(\Omega)$  and consider the bilinear form  $a(v, w)$  defined on  $\overset{\circ}{H}^1 \times \overset{\circ}{H}^1$  by

$$a(v, w) = \int_{\Omega} \left( \sum_{i=1}^N \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_i} + v w \right) dx. \quad (4.4)$$



By our hypotheses,  $a(\cdot, \cdot)$  is a bilinear, symmetric form on  $\overset{\circ}{H}^1 \times \overset{\circ}{H}^1$ . Moreover, for  $v, w \in \overset{\circ}{H}^1$  we have

$$\begin{aligned}
a(v, w) &\leq \sum_{i=1}^N \int_{\Omega} \left| \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_i} \right| + \int_{\Omega} |v w| \\
&\leq \sum_{i=1}^N \left\| \frac{\partial v}{\partial x_i} \right\| \left\| \frac{\partial w}{\partial x_i} \right\| + \|v\| \|w\| \\
&\leq \left( \sum_{i=1}^N \left\| \frac{\partial v}{\partial x_i} \right\|^2 \right)^{\frac{1}{2}} \left( \sum_{i=1}^N \left\| \frac{\partial w}{\partial x_i} \right\|^2 \right)^{\frac{1}{2}} + \|v\|_1 \|w\|_1 \\
&\leq 2 \|v\|_1 \|w\|_1,
\end{aligned} \tag{4.5}$$

i.e. that  $a(v, w)$  is continuous on  $\overset{\circ}{H}^1 \times \overset{\circ}{H}^1$ . Also

$$\begin{aligned}
a(v, v) &= \sum_{i=1}^N \int_{\Omega} \left( \frac{\partial v}{\partial x_i} \right)^2 + \int_{\Omega} v^2 \geq \sum_{i=1}^N \int_{\Omega} \left( \frac{\partial v}{\partial x_i} \right)^2 \\
&\geq c \|v\|_1^2, \quad c > 0
\end{aligned} \tag{4.6}$$

using the Poincaré–Friedrichs inequality. Hence  $a(\cdot, \cdot)$  satisfies the hypotheses of the Lax–Milgram theorem on the Hilbert space  $\overset{\circ}{H}^1$ . Since  $f \mapsto \int_{\Omega} f v$  is a bounded linear functional on  $\overset{\circ}{H}^1$ , it follows that (4.3) has a unique solution  $u \in \overset{\circ}{H}^1$  that satisfies

$$\|u\|_1 \leq C \|f\|.$$

(Note that by the symmetry of  $a(\cdot, \cdot)$ , the weak solution can be also characterized as the (unique) element of  $\overset{\circ}{H}^1$  that solves the minimization problem

$$J(u) = \inf_{v \in \overset{\circ}{H}^1(\Omega)} J(v),$$

where

$$J(v) = \frac{1}{2} \int_{\Omega} \left( \sum_{i=1}^N \left( \frac{\partial v}{\partial x_i} \right)^2 + v^2 \right) - \int_{\Omega} f v.$$

This is “Dirichlet’s principle”).

(Note also that simply  $a(v, w) = (v, w)_1$  on  $\overset{\circ}{H}^1 \times \overset{\circ}{H}^1$ . Hence an appeal to the Riesz theorem would solve the problem. Of course,  $a(v, w) \leq \|v\|_1 \|w\|_1$  and  $a(v, v) = \|v\|_1^2$  too. But the proof of (4.5), (4.6) above indicates the general way of proving (4.5) and (4.6) in the case e.g. of a positive definite form with variable coefficients).

**(iii) Regularity of the weak solution.**

If  $f \in L^2(\Omega)$  and  $\Omega$  is a domain of class  $C^2$  and if  $u \in \overset{\circ}{H}^1$  is a weak solution of (4.1), (4.2), i.e. a solution of (4.3), then one may show that in fact  $u \in H^2(\Omega)$  and that  $\|u\|_2 \leq C \|f\|$  (“elliptic regularity” estimate). Here  $C$  is a constant depending only on  $\Omega$ . More generally, if  $\Omega$  is of class  $C^{m+2}$  and if  $f \in H^m(\Omega)$ , then  $u \in H^{m+2}(\Omega)$  and  $\|u\|_{m+2} \leq C_m \|f\|_m$ . (In particular, using Sobolev’s Theorem 4.5 we may conclude that if  $m > \frac{N}{2}$ , then  $u \in C^2(\overline{\Omega})$ ).

**(iv) If the weak solution is in  $C^2(\overline{\Omega})$ , then it is classical.**

Let  $u$ , the weak solution of (4.1), (4.2) be in  $C^2(\overline{\Omega})$  and let  $f \in C(\overline{\Omega})$ . Since  $u \in \overset{\circ}{H}^1 \cap C^2(\overline{\Omega})$  we conclude that  $u|_{\partial\Omega} = 0$  (in the classical sense). Applying Green’s formula we have now from (4.3)

$$\int_{\Omega} (-\Delta u + u) v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in C_c^\infty(\Omega)$$

from which  $-\Delta u + u - f = 0$  a.e. in  $\Omega$  since  $C_c^\infty(\Omega)$  is dense in  $L^2(\Omega)$ . We conclude, by our hypotheses that  $-\Delta u + u = f \, \forall x \in \Omega$ . Hence  $u$  is a classical solution.

### Remarks

- (i) The above discussion extends with no extra difficulties to the case of a linear, self-adjoint elliptic operator with variable coefficients. Consider the problem of finding  $u : \overline{\Omega} \rightarrow \mathbb{R}$  such that

$$-\sum_{i,j=1}^N \frac{\partial}{\partial x_i} (a_{ij} \frac{\partial u}{\partial x_j}) + a_0 u = f \text{ in } \Omega \quad (4.7)$$

$$u = 0 \text{ on } \partial\Omega, \quad (4.8)$$

where we suppose that  $a_{ij}(x) = a_{ji}(x)$  are functions in  $C^1(\overline{\Omega})$  such that the matrix  $a_{ij}$  is symmetric and uniformly positive definite, i.e. that the *ellipticity condition*

$$\sum_{i,j=1}^N a_{ij}(x) \xi_i \xi_j \geq \alpha \sum_{i=1}^N \xi_i^2$$

holds for some  $\alpha > 0 \, \forall x \in \overline{\Omega}, \xi \in \mathbb{R}^N$ . We also suppose that  $a_0 \in C(\overline{\Omega})$  and that  $a_0(x) \geq 0$  on  $\overline{\Omega}$ . We define a classical solution of (4.7), (4.8) to be a function  $u \in C^2(\overline{\Omega})$  satisfying (4.7), (4.8) in the usual sense, while a *weak solution* is an element of  $\overset{\circ}{H}^1$  satisfying

$$A(u, v) = \int_{\Omega} \left( \sum_{i,j} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + a_0 uv \right) = (f, v) \quad \forall v \in \overset{\circ}{H}^1. \quad (4.9)$$

For  $f \in L^2(\Omega)$  we may show that (4.9) has a unique solution  $u \in \overset{\circ}{H}^1$ : It is easy to see that  $A(u, v) \leq C_1 \|u\|_1 \|v\|_1 \ \forall u, v \in \overset{\circ}{H}^1$  and  $A(u, v) \geq C_2 \|v\|_1^2 \ \forall v \in \overset{\circ}{H}^1$  for  $C_1, C_2 > 0$ . The regularity in  $H^2$  also holds under our hypotheses on  $a_{ij}, a_0$ . In general, for  $m \geq 1$  if  $f \in H^m(\Omega)$ ,  $a_{ij} \in C^{m+1}(\overline{\Omega})$ ,  $a_0 \in C^m(\overline{\Omega})$  yields  $u \in H^{m+2}(\Omega)$  (elliptic regularity).

- (ii) The fact that the weak solution of (4.1), (4.2) is in  $C^2(\overline{\Omega})$  if  $f \in H^m$  with  $m > \frac{N}{2}$  follows from Sobolev's theorem. There is a sharper theory based on *Schauder's estimates* which states that if  $\Omega$  is of class  $C^{2,a}$  (Hölder spaces) with  $0 < a < 1$  and  $f \in C^{0,a}(\overline{\Omega})$ , then  $\exists u \in C^{2,a}(\overline{\Omega})$ , unique solution of (4.1), (4.2) in the classical sense. Moreover if  $\Omega$  is of class  $C^{m+2}(\overline{\Omega})$  ( $m \geq 1$  integer) and  $f \in C^{m,a}(\overline{\Omega})$ , then  $u \in C^{m+2,a}(\overline{\Omega})$  and an analogous elliptic regularity result holds. Here

$$\begin{aligned} C^{0,a}(\overline{\Omega}) &= \{u \in C(\overline{\Omega}), \sup_{x,y \in \Omega, x \neq y} \frac{|u(x) - u(y)|}{|x - y|^a} < \infty\} \\ C^{m,a}(\overline{\Omega}) &= \{u \in C^m(\overline{\Omega}), D^\beta u \in C^{0,a}(\overline{\Omega}) \ \forall \beta : |\beta| = m\}. \end{aligned}$$

#### 4.5.2 (b) Homogeneous Neumann boundary conditions.

We now consider the problem of finding  $u : \overline{\Omega} \rightarrow \mathbb{R}$  such that

$$-\Delta u + u = f \quad \text{in } \Omega \quad (4.10)$$

$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega \quad (4.11)$$

with  $f$  given on  $\Omega$ . As usual  $\frac{\partial u}{\partial n} = \vec{\nabla} u \cdot \vec{n}$  is the normal derivative at the boundary  $\partial\Omega$  (again  $\Omega$  is of class  $C^1$ ). A *classical solution* of (4.10), (4.11) (for  $f \in C(\overline{\Omega})$ ) is a  $C^2(\overline{\Omega})$  function  $u$  satisfying (4.10), (4.11) in the classical sense. A *weak solution* is an element  $u \in H^1(\Omega)$  satisfying (for  $f \in L^2(\Omega)$  say)

$$a(u, v) \equiv \int_{\Omega} \left( \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} + u v \right) dx = \int_{\Omega} f v \quad \forall v \in H^1(\Omega). \quad (4.12)$$

**(i) Every classical solution is weak.**

Let  $u \in C^2(\overline{\Omega})$  be a classical solution of (4.10), (4.11). Then Green's formula gives

$$\int_{\Omega} \Delta u v = - \int_{\Omega} \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} + \int_{\partial\Omega} \frac{\partial u}{\partial n} v dy \quad \forall v \in C^\infty(\overline{\Omega}).$$

Hence (4.10), (4.11) yields that

$$\int_{\Omega} \left( \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} + u v \right) = \int_{\Omega} f v \quad \forall v \in C^{\infty}(\bar{\Omega})$$

and density of  $C^{\infty}(\bar{\Omega})$  in  $H^1(\Omega)$  yields then that (4.12) is satisfied.

(ii) An immediate application of the Lax–Milgram theorem yields that *there exists a unique weak solution.*

(iii) More refined theory yields again the *regularity of the weak solution.* Exactly the same results as in the Dirichlet b.c. case hold.

**(iv) If the weak solution is in  $C^2(\bar{\Omega})$ , then it is classical.**

For if this case we have by (4.12) that ( $f \in C(\bar{\Omega})$  here)

$$\int_{\Omega} (-\Delta u + u) v + \int_{\partial\Omega} \frac{\partial u}{\partial n} v dy = \int_{\Omega} f v \quad \forall v \in C^{\infty}(\bar{\Omega}).$$

Choosing  $v \in C_c^{\infty}(\Omega)$  we obtain as in the Dirichlet b.c. case that  $-\Delta u + u = f$  in  $\Omega$ .

It follows that  $\int_{\partial\Omega} \frac{\partial u}{\partial n} v dy = 0 \quad \forall v \in C^{\infty}(\bar{\Omega}) \Rightarrow \frac{\partial u}{\partial n} = 0$  on  $\partial\Omega$ .

# Chapter 5

## The Galerkin Finite Element Method for Elliptic Boundary–Value Problems

### 5.1 Introduction

Let  $\Omega$  be a bounded domain in  $\mathbb{R}^N$ . Consider the boundary–value problem of finding  $u = u(x)$ ,  $x \in \bar{\Omega}$ , such that

$$\left. \begin{aligned} Lu &= f, & x \in \Omega \\ u &= 0, & x \in \partial\Omega \end{aligned} \right\} \quad (5.1)$$

where  $L$  is the elliptic operator given by

$$Lu = - \sum_{i,j=1}^N \frac{\partial}{\partial x_i} \left( a_{ij}(x) \frac{\partial u}{\partial x_j} \right) + a_0(x) u.$$

As in the previous chapter, we assume that for  $1 \leq i, j \leq N$   $a_{ij}(x) = a_{ji}(x)$ ,  $x \in \bar{\Omega}$ , that  $\exists c > 0$  such that

$$\sum_{i,j=1}^N a_{ij}(x) \xi_i \xi_j \geq c \sum_{i=1}^N \xi_i^2 \quad \forall \xi \in \mathbb{R}^N,$$

that  $a_{ij}$ ,  $a_0$  are sufficiently smooth functions of  $x$ , and that  $a_0 \geq 0$  in  $\bar{\Omega}$ . Under these hypotheses, and if  $f \in L^2(\Omega)$ , we have shown that there exists a weak solution  $u \in \overset{\circ}{H}^1(\Omega)$  of (5.1) satisfying

$$B(u, v) = (f, v) \quad \forall v \in \overset{\circ}{H}^1, \quad (5.2)$$

where

$$B(u, v) = \int_{\Omega} \left( \sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + a_0(x)uv \right) dx.$$

We shall assume that the data are such that the unique solution  $u$  of (5.2) belongs to  $H^2 \cap \overset{\circ}{H}^1$  and satisfies the elliptic regularity estimate that for some  $C > 0$ , independent of  $f$  and  $u$ , we have

$$\|u\|_2 \leq C \|f\|. \quad (5.3)$$

As we have seen in Chapters 1 and 3, the (*standard*) *Galerkin method* for approximating the solution  $u$  of (5.2) amounts to constructing a family of finite-dimensional subspaces  $S_h$  of  $\overset{\circ}{H}^1$ , say for  $0 < h < 1$ , and seeking  $u_h \in S_h$  satisfying the linear system of equations

$$B(u_h, v_h) = (f, v_h) \quad \forall v_h \in S_h. \quad (5.4)$$

Under our hypotheses, we have seen that a unique solution  $u_h$  of (5.4) exists and satisfies

$$\|u - u_h\|_1 \leq C \inf_{\chi \in S_h} \|u - \chi\|_1 \quad (5.5)$$

for some constant  $C$  independent of  $h$ . Assuming e.g. that

$$\inf_{\chi \in S_h} (\|v - \chi\| + h \|v - \chi\|_1) \leq Ch^2 \|v\|_2 \quad \forall v \in H^2 \cap \overset{\circ}{H}^1, \quad (5.6)$$

we obtain from (5.5) the *optimal-rate  $H^1$ -error estimate*

$$\|u - u_h\|_1 \leq Ch \|u\|_2. \quad (5.7)$$

The  $L^2$ -error estimate is obtained again by the “Nitsche trick”, by letting  $e = u - u_h$  and considering  $w \in H^2 \cap \overset{\circ}{H}^1$ , the solution of the problem

$$B(w, v) = (e, v) \quad \forall v \in \overset{\circ}{H}^1. \quad (5.8)$$

Then,  $\|e\|^2 = (e, e) = B(w, e) = B(e, w) = B(e, w - \chi)$  for any  $\chi \in S_h$  – we used (5.2) and (5.4) –. By the continuity of  $B$  in  $H^1 \times H^1$  we have then

$$\|e\|^2 \leq C \|e\|_1 \|w - \chi\|_1 \stackrel{\text{by(5.6)}}{\leq} C \|e\|_1 h \|w\|_2 \leq Ch \|e\|_1 \|e\|.$$

Hence  $\|e\| \leq Ch \|e\|_1 \leq Ch^2 \|u\|_2$  by (5.7)

In general, assuming that for some  $r \geq 2$  (integer) we have

$$\inf_{\chi \in \mathcal{S}_h} (\|v - \chi\| + h \|v - \chi\|_1) \leq C h^r \|v\|_r \quad \forall v \in H^r \cap \mathring{H}^1, \quad (5.9)$$

and that the weak solution  $u$  of (5.2) is in  $H^r \cap \mathring{H}^1$ , we see that (5.5) and the Nitsche argument give

$$\|u - u_h\| + h \|u - u_h\|_1 \leq C h^r \|u\|_r. \quad (5.10)$$

Hence, our task is to construct subspaces of  $\mathring{H}^1$  (endowed with bases of small support so that the linear system (5.4) is sparse) so that (5.6) or, in general, (5.9) holds. In what follows we shall consider the subspace of piecewise linear, continuous functions on a polygonal domain in  $\mathbb{R}^2$  (subdivided into triangular elements).

## 5.2 Piecewise linear, continuous functions on a triangulation of a plane polygonal domain

(This section is largely based on Ciarlet (1978), chapter 3).

We consider a *convex* polygonal domain  $\Omega$  in  $\mathbb{R}^2$  and the elliptic b.v.p. (5.1) associated with it. Although  $\Omega$  is not  $C^1$ , it is known that (5.3) still holds.

We subdivide  $\Omega$  into triangles  $\tau_i$ ,  $1 \leq i \leq M$ ,  $M = M(h)$ , forming a triangulation  $\mathcal{T}_h = \{\tau_i\}$  of  $\Omega$ . We assume that the  $\tau_i$  are open and disjoint, that  $\max_i(\text{diam } \tau_i) \leq h$ ,  $0 < h < 1$ , and that  $\Omega = \text{Int}(\cup_{i=1}^M \overline{\tau}_i)$ . The vertices of the triangles are called *nodes* of the triangulation. We shall assume that  $\mathcal{T}_h$  is such that there are no nodes in the interior of the interior sides of triangles:

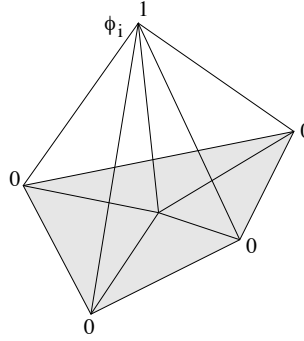


We let now  $S_h$  be the vector space of continuous functions on  $\overline{\Omega}$  that are linear on each  $\tau_i$  and vanish on  $\partial\Omega$ , i.e. let

$$S_h = \{\phi \in C(\overline{\Omega}), \phi|_{\tau_i} = \alpha_i + \beta_i x + \gamma_i y \quad (\text{i.e. } \phi \in \mathbb{P}_1(\tau_i)), \phi|_{\partial\Omega} = 0\}.$$

Let  $N = N(h)$  be the number of *interior* nodes  $P_i$  of the triangulation (i.e. those nodes are not on  $\partial\Omega$ ). Since these points which are not collinear define a single plane, it follows that an element  $\phi \in S_h$  is uniquely defined by its values  $\phi(P_j)$ ,  $1 \leq j \leq N$ . (This means that  $\dim S_h = N$ ). A suitable basis of  $S_h$  (for our purposes) consists of the functions  $\phi_i \in S_h$ ,  $1 \leq i \leq N$ , such that

$$\phi_i(P_j) = \delta_{ij}, \quad 1 \leq i, j \leq N.$$



It is clear that the support of  $\phi_i$  consists exactly of those triangles of  $\mathcal{T}_h$  that share  $P_i$  as a vertex. The  $\phi_i$  are linearly independent, since if  $\sum_{i=1}^N c_i \phi_i(x) = 0$ , then putting  $x = x_j$  yields  $c_j = 0$ ,  $1 \leq j \leq N$ . Moreover, given  $\psi \in S_h$ , we can write

$$\psi(x) = \sum_{j=1}^N \psi(P_j) \phi_j(x), \quad x \in \bar{\Omega}, \quad (5.11)$$

since both sides of (5.11) are elements of  $S_h$  that coincide at the interior nodes  $P_j$ ,  $1 \leq j \leq N$ . Hence,  $\{\phi_i\}_{i=1}^N$  form a basis for  $S_h$ .

$S_h$  is a subspace of  $\overset{0}{H}{}^1(\Omega)$ : Obviously,  $S_h \subset L^2(\Omega)$  and the elements of  $S_h$  vanish (pointwise) on  $\partial\Omega$ . Hence, to show that  $v \in S_h$  belongs to  $H^1(\Omega)$ , it suffices to prove that there exist  $g_i \in L^2(\Omega)$ ,  $i = 1, 2$ , such that

$$\int_{\Omega} v \frac{\partial \phi}{\partial x_i} = - \int_{\Omega} g_i \phi \quad \forall \phi \in C_c^{\infty}(\Omega), \quad i = 1, 2. \quad (5.12)$$

Let  $v^{(\tau)}$  be the restriction of  $v$  to  $\tau \in \mathcal{T}_h$ . For  $i = 1, 2$  let  $g_i \in L^2(\Omega)$  be defined as the piecewise constant function given by

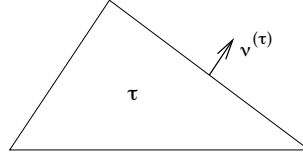
$$g_i = \frac{\partial}{\partial x_i} (v^{(\tau)}) \quad i = 1, 2, \quad \text{if } x \in \tau. \quad (5.13)$$



There follows that for  $\phi \in C_c^\infty(\Omega)$

$$\begin{aligned} \int_{\Omega} v \frac{\partial \phi}{\partial x_i} dx &= \sum_{\tau \in \mathcal{T}_h} \int_{\tau} v \frac{\partial \phi}{\partial x_i} dx = \sum_{\tau \in \mathcal{T}_h} \int_{\tau} v^{(\tau)} \frac{\partial \phi}{\partial x_i} dx \stackrel{\text{(Gauss theorem on } \tau)}{=} \\ &= - \sum_{\tau \in \mathcal{T}_h} \int_{\tau} g_i \phi dx + \sum_{\tau \in \mathcal{T}_h} \int_{\partial \tau} v^{(\tau)} \phi \nu_i^{(\tau)} dy, \end{aligned}$$

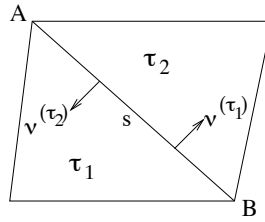
where  $\nu^{(\tau)} = (\nu_1^{(\tau)}, \nu_2^{(\tau)})^T$  is the unit outward normal on the boundary  $\partial \tau$  of  $\tau$ .



The first term of the right-hand side of the above equality is equal to  $-\int_{\Omega} g_i \phi dx$ , where  $g_i \in L^2(\Omega)$  was defined (piecewise) by (5.13). The second term vanishes. To see this, note that the second term is eventually the sum of terms

$$\int_s v^{(\tau)} \phi \nu_i^{(\tau)} dy$$

where  $s$  any side of any triangle  $\tau$ . If  $s \subset \partial \Omega$ , then the corresponding term is zero since  $\phi = 0$  on  $\partial \Omega$ . If  $s = AB$  is an interior side, suppose it is the common side of two adjacent triangles  $\tau_1$  and  $\tau_2$ .



In that case, there are precisely two terms in the sum  $\sum_{s: s \in \partial \tau} \int_s \dots$  involving  $AB$ , namely

$$\int_{AB} v^{(\tau_1)} \phi \nu_i^{(\tau_1)} dy \quad \text{and} \quad \int_{AB} v^{(\tau_2)} \phi \nu_i^{(\tau_2)} dy,$$

which cancel each other since  $v^{(\tau_1)}|_{AB} = v^{(\tau_2)}|_{AB}$  ( $v$  is continuous in  $\bar{\Omega}$  as an element of  $S_h$ ),  $\phi \in C_c^\infty(\Omega)$ , and  $\nu_i^{(\tau_1)} = -\nu_i^{(\tau_2)}$ ,  $i = 1, 2$ . We conclude then that (5.12) holds, i.e. that  $v \in H^1(\Omega)$ , q.e.d.

Given  $v \in C(\bar{\Omega})$ ,  $v|_{\partial\Omega} = 0$ , we define the *interpolant*  $I_h v$  of  $v$  in  $S_h$ , as the unique element  $I_h v$  of  $S_h$  that coincides with  $v$  at the interior nodes  $P_i$ ,  $1 \leq i \leq N$ , of the triangulation  $\mathcal{T}_h$ , i.e. as

$$(I_h v)(x) = \sum_{i=1}^N v(P_i) \phi_i(x), \quad x \in \bar{\Omega}.$$

It will be the objective of this section to show that for  $v \in H^2(\Omega) \cap \mathring{H}^1(\Omega)$  (note that  $v \in C(\bar{\Omega})$ ,  $v|_{\partial\Omega} = 0$ ), we have, for some constant  $C$  independent of  $h$  and  $v$ :

$$\|v - I_h v\| + h \|v - I_h v\|_1 \leq C h^2 |v|_{2,\Omega}. \quad (5.14)$$

(Given  $v \in H^m(\omega)$ , where  $\omega$  is a subdomain of  $\Omega$ , we define

$$\begin{aligned} |v|_{0,\omega} &= \|v\|_{L^2(\omega)} \\ |v|_{1,\omega} &= \left( \left\| \frac{\partial v}{\partial x_1} \right\|_{L^2(\omega)}^2 + \left\| \frac{\partial v}{\partial x_2} \right\|_{L^2(\omega)}^2 \right)^{\frac{1}{2}} \\ &\vdots \\ |v|_{m,\omega} &= \left( \sum_{|\alpha|=m} \|D^\alpha v\|_{L^2(\omega)}^2 \right)^{\frac{1}{2}}, \quad D^\alpha = \frac{\partial^{\alpha_1 + \alpha_2}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}}. \end{aligned}$$

Then  $|v|_{m,\omega}$  is in general a semi-norm on  $H^m(\Omega)$ ).

If (5.14) is established, then (5.6) holds and, as a consequence, we have our optimal-order  $L^2$  and  $H^1$  error estimates for  $u - u_h$ .

The estimate (5.14) will be proved as a consequence of two facts:

**(i) A local  $L^2$  and  $H^1$  estimate for the interpolant:**

Given a function  $v \in C(\bar{\tau})$ , where  $\tau$  is the triangle with vertices  $P_1, P_2, P_3$  define the (local) interpolant  $I_\tau v \in \mathbb{P}_1(\tau)$  as the unique linear polynomial in  $x_1, x_2$  on  $\tau$  such that

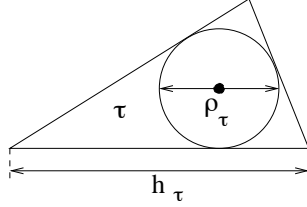
$$(I_\tau v)(P_i) = v(P_i), \quad i = 1, 2, 3.$$

Note that if  $v \in C(\bar{\Omega})$ , then  $I_h v|_\tau = I_\tau v$ .

Following Ciarlet, we shall prove that there exists a constant  $C$ , independent of  $\tau$  and  $\mathcal{T}_h$ , such that for each  $v \in H^2(\tau)$ ,  $\tau \in \mathcal{T}_h$

$$|v - I_\tau v|_{m,\tau} \leq C \frac{h_\tau^2}{\rho_\tau^m} |v|_{2,\tau}, \quad m = 0, 1, \quad (5.15)$$

where  $h_\tau = \text{diam}\tau \equiv$  the length of the largest side of the triangle  $\tau$ , and  $\rho_\tau$  is the diameter of the inscribed circle in the triangle  $\tau$ .



**(ii) The regularity of the triangulation:**

We shall assume that the triangulation  $\mathcal{T}_h$  is *regular*, in the sense that there exists a constant  $\sigma > 0$ , independent of  $\tau$  and  $\mathcal{T}_h$ , such that

$$\frac{h_\tau}{\rho_\tau} \leq \sigma \quad \forall \tau \in \mathcal{T}_h. \quad (5.16)$$

(The regularity condition (5.16) essentially states that  $h = \max_\tau h_\tau \rightarrow 0$ , i.e. the triangulation is refined, if and only if  $\max \rho_\tau \rightarrow 0$  i.e. all triangles tend to become ‘points’ and not ‘needles’, (for which  $\frac{h_\tau}{\rho_\tau}$  would become unbounded). It may be shown that (5.16) is equivalent to requiring that there exists a  $\theta_0 > 0$  independent of  $\mathcal{T}_h$  such that  $\theta_\tau \geq \theta_0 \quad \forall \tau \in \mathcal{T}_h$ , where  $\theta_\tau$  is the minimum (interior) angle of  $\tau$ . It is also equivalent to requiring that  $\exists c_0$ , independent of  $\mathcal{T}_h$ , such that  $\mu(\tau) \geq c_0 h_\tau^2 \quad \forall \tau \in \mathcal{T}_h$ , where  $\mu(\tau) = \text{area}(\tau)$ ).

Indeed, if (5.15) and (5.16) hold, we have

$$\frac{1}{\rho_\tau} \leq \frac{\sigma}{h_\tau} \Rightarrow \frac{1}{\rho_\tau^m} \leq \frac{\sigma^m}{h_\tau^m}, \quad m = 0, 1,$$

so that (5.15) gives

$$|v - I_\tau v|_{m,\tau} \leq C_m h_\tau^{2-m} |v|_{2,\tau}, \quad m = 0, 1, \quad \forall v \in H^2(\tau), \quad (5.17)$$

for constants  $C_0, C_1$  independent of  $\mathcal{T}_h$ .

We conclude then for  $v \in H^2(\Omega) \cap \overset{0}{H}{}^1(\Omega)$  ( $\Rightarrow v \in C(\bar{\Omega})$  by Sobolev’s theorem)

$$\begin{aligned} \|v - I_h v\|^2 &\equiv \|v - I_h v\|_{L^2(\Omega)}^2 = \sum_{\tau \in \mathcal{T}_h} |v - I_\tau v|_{0,\tau}^2 \leq C_0^2 \sum_{\tau \in \mathcal{T}_h} h_\tau^4 |v|_{2,\tau}^2 \\ &\leq C_0^2 h^4 \sum_{\tau \in \mathcal{T}_h} |v|_{2,\tau}^2 = C_0^2 h^4 |v|_{2,\Omega}^2. \end{aligned}$$

Hence

$$\| v - I_h v \| \leq C h^2 | v |_{2,\Omega} \quad (5.18)$$

for some constant  $C$  independent of  $\mathcal{T}_h$ .

In addition, analogously

$$\begin{aligned} | v - I_h v |_{1,\Omega}^2 &= \sum_{\tau \in \mathcal{T}_h} | v - I_\tau v |_{1,\tau}^2 \leq C_1^2 \sum_{\tau \in \mathcal{T}_h} h_\tau^2 | v |_{2,\tau}^2 \\ &\leq C_1^2 h^2 \sum_{\tau \in \mathcal{T}_h} | v |_{2,\tau}^2 = C_1^2 h^2 | v |_{2,\Omega}^2. \end{aligned}$$

Therefore

$$| v - I_h v |_{1,\Omega} \leq C h | v |_{2,\Omega}, \quad (5.19)$$

for some  $C$  independent of  $\mathcal{T}_h$ .

The estimates (5.18) and (5.19) yield then (5.14) as advertized. We turn then to proving (5.15). To accomplish this we shall need a series of results and definitions:

**Definition:** Let  $\Omega, \hat{\Omega}$  be two bounded domains in  $\mathbb{R}^N$ . We say that  $\Omega$  and  $\hat{\Omega}$  are affinely equivalent if there exists an invertible affine map  $F : \hat{\Omega} \rightarrow \Omega$  such that  $F(\hat{\Omega}) = \Omega$ .

In the definition of the invertible affine map  $B$  is an  $N \times N$  invertible matrix of constants and  $b \in \mathbb{R}^N$ . Hence, if  $x \in \mathbb{R}^N$ ,  $\hat{x} = F^{-1}(x) \equiv B^{-1}x - B^{-1}b$ .

Hence  $\hat{x} \in \hat{\Omega} \Leftrightarrow x = F(\hat{x}) \in \Omega$  and if  $\hat{v} : \hat{\Omega} \rightarrow \mathbb{R}$  is a real-valued function defined on  $\hat{\Omega}$ , then defining  $v = \hat{v} \circ F^{-1} : \Omega \rightarrow \mathbb{R}$ , we have if  $x \in \Omega$ ,  $x = F(\hat{x})$ , i.e. if  $\hat{x} = F^{-1}(x)$ , that

$$v(x) = (\hat{v} \circ F^{-1})(x) = \hat{v}(F^{-1}(x)) = \hat{v}(\hat{x}).$$

(Note that if  $v = \hat{v} \circ F^{-1} : \Omega \rightarrow \mathbb{R}$ , then  $\hat{v} = v \circ F : \hat{\Omega} \rightarrow \mathbb{R}$ ).

Using this notation we may prove:

**Lemma 5.1.** Let  $\Omega, \hat{\Omega}$  be two affinely equivalent bounded domains in  $\mathbb{R}^N$  and let  $F$  be the associated affine map such that  $F(\hat{\Omega}) = \Omega$ . Then  $v \in H^m(\Omega)$ ,  $m \geq 0$  integer, if and only if  $\hat{v} = v \circ F \in H^m(\hat{\Omega})$ . Moreover, there exist constants  $C$  and  $\hat{C}$  depending only on  $m$  and  $N$  such that

$$| \hat{v} |_{m,\hat{\Omega}} \leq C | B |^m | \det B |^{-\frac{1}{2}} | v |_{m,\Omega} \quad (5.20)$$

$$\text{and } | v |_{m,\Omega} \leq \hat{C} | B^{-1} |^m | \det B |^{\frac{1}{2}} | \hat{v} |_{m,\hat{\Omega}}. \quad (5.21)$$

Here  $|B|$  denotes the matrix norm induced by the Euclidean vector norm in  $\mathbb{R}^N$ , i.e.

$$|B| = \sup_{\mathbb{R}^N \ni x \neq 0} \frac{|Bx|}{|x|}, \text{ where } |x| = \left( \sum_{i=1}^N x_i^2 \right)^{\frac{1}{2}}.$$

**Proof:** We shall prove (5.20) for  $m = 0$  and  $m = 1$ . Recall that whenever  $x = F(\hat{x}) = B\hat{x} + b$ , i.e. whenever  $\hat{x} = B^{-1}x + c$ ,  $c = -B^{-1}b$ , then  $\hat{v}(\hat{x}) = v(x)$ . Hence

$$\int_{\hat{\Omega}} \hat{v}^2(\hat{x}) d\hat{x} = \int_{\Omega} v^2(x) |J| dx,$$

where  $J$  is the determinant of the Jacobian matrix of the transformation  $\hat{x} = B^{-1}x + c$ . Hence  $J = \det(B^{-1}) = (\det(B))^{-1}$  and we conclude that

$$\|\hat{v}\|_{0,\hat{\Omega}}^2 = \int_{\hat{\Omega}} \hat{v}^2(\hat{x}) d\hat{x} = |\det B|^{-1} \int_{\Omega} v^2(x) dx = |\det B|^{-1} \|v\|_{0,\Omega}^2,$$

which implies (5.20) with  $m = 0$ ,  $C = 1$ .

Let now  $m = 1$  and  $\hat{v} \in H^1(\hat{\Omega})$ . Then, if  $x = F(\hat{x})$  and  $1 \leq i \leq N$ ,

$$\frac{\partial \hat{v}}{\partial \hat{x}_i}(\hat{x}) = \frac{\partial}{\partial \hat{x}_i}(v(x)) = \sum_{j=1}^N \frac{\partial v}{\partial x_j}(x) \frac{\partial x_j}{\partial \hat{x}_i} = \sum_{j=1}^N \frac{\partial v}{\partial x_j}(x) \frac{\partial F_j(\hat{x})}{\partial \hat{x}_i},$$

where  $x_j = F_j(\hat{x}) \equiv \sum_{k=1}^N B_{jk}\hat{x}_k + b_j$ . Hence  $\frac{\partial F_j(\hat{x})}{\partial \hat{x}_i} = B_{ji}$  and we conclude that

$$\frac{\partial \hat{v}}{\partial \hat{x}_i}(\hat{x}) = \sum_{j=1}^N \frac{\partial v}{\partial x_j}(x) B_{ji}, \quad 1 \leq i \leq N,$$

which we may write as

$$(\hat{\nabla} \hat{v})(\hat{x}) = B^T (\nabla v)(x), \tag{5.22}$$

where

$$\hat{\nabla} \hat{v} = \left( \frac{\partial \hat{v}}{\partial \hat{x}_1}, \dots, \frac{\partial \hat{v}}{\partial \hat{x}_N} \right)^T \text{ and } \nabla v = \left( \frac{\partial v}{\partial x_1}, \dots, \frac{\partial v}{\partial x_N} \right)^T.$$

From (5.22), taking Euclidean norms we see that

$$|\hat{\nabla} \hat{v}| \leq |B^T| |\nabla v| = |B| |\nabla v| \tag{5.23}$$

since  $|B^T| = |B|$ . (To see this, recall that

$$|B| = \max_i \sqrt{\lambda_i(B^T B)} = \max_i \sqrt{\lambda_i(B B^T)} = |B^T|,$$

since  $B B^T$  and  $B^T B$  have the same eigenvalues).

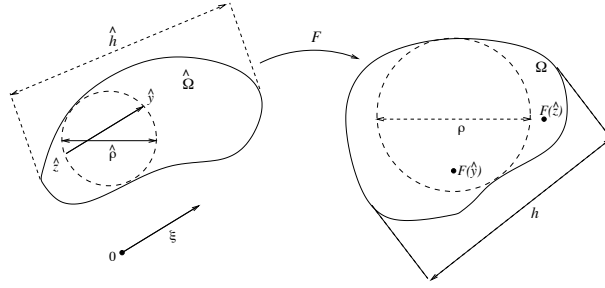
Hence, using (5.23) we have

$$\begin{aligned} |\hat{v}|_{1,\hat{\Omega}}^2 &= \sum_{i=1}^N \int_{\hat{\Omega}} \left( \frac{\partial \hat{v}}{\partial \hat{x}_i}(\hat{x}) \right)^2 d\hat{x} = \int_{\hat{\Omega}} |\widehat{\nabla} \hat{v}|^2 d\hat{x} \leq |B|^2 \int_{\Omega} |\nabla v|^2 |J| dx \\ &= |B|^2 |\det B|^{-1} |v|_{1,\Omega}^2, \end{aligned}$$

which is (5.20) for  $m = 1$ .

The rest of the proof is analogous.  $\square$

Given two bounded, affinely equivalent domains  $\Omega$  and  $\hat{\Omega}$  we let



$\hat{h} = \text{diam} \hat{\Omega} \equiv \sup_{\hat{x}, \hat{y} \in \hat{\Omega}} |\hat{x} - \hat{y}|$ ,  $\hat{\rho} = \text{diameter of the inscribed ball in } \hat{\Omega} = \sup\{\text{diam} \hat{S}, \hat{S} \text{ is a ball contained in } \hat{\Omega}\}$ , and  $h, \rho$  be the corresponding quantities for  $\Omega$ .

**Lemma 5.2.** *Let  $\Omega, \hat{\Omega}$  be two affinely equivalent bounded domains in  $\mathbb{R}^N$  such that  $\Omega = F(\hat{\Omega})$ ,  $F(\hat{x}) = B\hat{x} + b$ . Then*

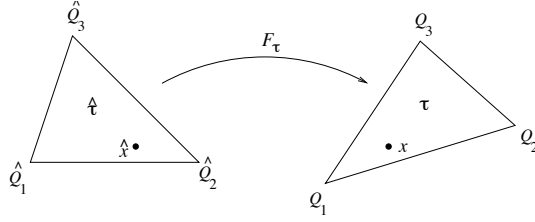
$$|B| \leq \frac{h}{\hat{\rho}}, \quad |B^{-1}| \leq \frac{\hat{h}}{\rho}. \quad (5.24)$$

**Proof:** An easy scaling argument yields that

$$|B| = \frac{1}{\hat{\rho}} \sup_{\xi \in \mathbb{R}^N, |\xi| = \hat{\rho}} |B\xi|.$$

Now given  $\xi \in \mathbb{R}^N$  such that  $|\xi| = \hat{\rho}$ , we may find two points  $\hat{y}, \hat{z} \in \hat{\Omega}$  such that  $\hat{y} - \hat{z} = \xi$  (see Fig. 5.6). For these points we have that  $F(\hat{y}) - F(\hat{z}) = B\hat{y} - B\hat{z} = B\xi$ , with  $F(\hat{y}), F(\hat{z}) \in \bar{\Omega}$ . Hence  $\sup_{\xi \in \mathbb{R}^N, |\xi| = \hat{\rho}} |B\xi| \leq h$ . We conclude that  $|B| \leq \frac{h}{\hat{\rho}}$ .  $\square$

Let us consider now our triangulation  $\mathcal{T}_h = \{\tau\}$  of the polygonal domain  $\Omega$  in  $\mathbb{R}^2$ .



It is clear that any triangle  $\tau$  of the triangulation is affinely equivalent to a fixed *reference triangle*  $\hat{\tau}$ . (E.g. we can take  $\hat{\tau}$  to be the triangle with vertices  $(0,0)$ ,  $(0,1)$ ,  $(1,0)$ ). Indeed we have for each  $\tau \in \mathcal{T}_h$

$$F_\tau(\hat{\tau}) = \tau$$

for some invertible affine map  $F_\tau(\hat{x}) = B_\tau \hat{x} + b_\tau$ ,  $\hat{x} \in \hat{\tau}$ , depending on  $\tau$ . (We construct the map  $F_\tau$  by requiring that  $Q_i = F_\tau(\hat{Q}_i)$ ,  $1 \leq i \leq 3$ , where  $\hat{Q}_i$ ,  $Q_i$ ,  $i = 1, 2, 3$ , are the vertices of  $\hat{\tau}$ ,  $\tau$ , respectively. These three 2–vector equations determine uniquely the six constants which are the entries of the matrix  $B_\tau$  and the vector  $b_\tau$ . Then we may easily check that  $Q_1 Q_2 = F_\tau(\hat{Q}_1 \hat{Q}_2)$  etc., and that each point  $\hat{x}$  in  $\hat{\tau}$  is mapped onto a uniquely defined point  $x$  in  $\tau$ , and that each point  $\hat{x} \in \partial \hat{\tau}$  is mapped onto a corresponding point  $x$  on  $\partial \tau$ . The idea is to work on  $\hat{\tau}$  and obtain corresponding estimates on  $\tau$  (such as the required (5.15)) by using the properties of the interpolant in spaces of piecewise linear continuous functions as well as the scaling and transformation inequalities of Lemmata 5.1 and 5.2.

To this effect, define the interpolant  $I_{\hat{\tau}}$  on the reference triangle  $\hat{\tau}$  as the map  $I_{\hat{\tau}} : C(\hat{\tau}) \rightarrow \mathbb{P}_1(\hat{\tau})$  such that

$$(I_{\hat{\tau}} \hat{v})(\hat{Q}_i) = \hat{v}(\hat{Q}_i), \quad 1 \leq i \leq 3. \quad (5.25)$$

for any continuous real–valued function  $\hat{v}$  defined on  $\hat{\tau}$ . Our basic step towards proving (5.15) is the following

**Lemma 5.3.** *Let  $\hat{v} \in H^2(\hat{\tau})$ . Then, there exists a constant  $C(\hat{\tau})$  such that for  $m = 0, 1, 2$*

$$| \hat{v} - I_{\hat{\tau}} \hat{v} |_{m, \hat{\tau}} \leq C(\hat{\tau}) | \hat{v} |_{2, \hat{\tau}}. \quad (5.26)$$

**Proof:** The estimates (5.26) will be proved as a consequence of three important facts:

(i) **The linear map  $I_{\hat{\tau}}$  preserves linear polynomials**, i.e.  $\forall \hat{p} \in \mathbb{P}_1(\hat{\tau}), I_{\hat{\tau}}\hat{p} = \hat{p}$ .

This is obvious and implies that for any  $\hat{v} \in H^2(\hat{\tau})$  and *any*  $\hat{p} \in \mathbb{P}_1(\hat{\tau})$

$$\hat{v} - I_{\hat{\tau}}\hat{v} = \hat{v} - \hat{p} - I_{\hat{\tau}}\hat{v} + I_{\hat{\tau}}\hat{p} = (\hat{v} - \hat{p}) - I_{\hat{\tau}}(\hat{v} - \hat{p}). \quad (5.27)$$

(ii)  $I_{\hat{\tau}}$  is “**stable**” on  $H^2(\hat{\tau})$ . By this we mean that there exists a constant  $\tilde{C} = \tilde{C}(\hat{\tau})$  such that

$$\| I_{\hat{\tau}}\hat{w} \|_{2,\hat{\tau}} \leq \tilde{C}(\hat{\tau}) \| \hat{w} \|_{2,\hat{\tau}} \quad \forall \hat{w} \in H^2(\hat{\tau}). \quad (5.28)$$

(To see this, let  $\hat{\phi}_i \in \mathbb{P}_1(\hat{\tau})$  be the “hat” basis functions associated with the vertices  $\hat{Q}_i$  of  $\hat{\tau}$ , i.e. let  $\hat{\phi}_i \in \mathbb{P}_1(\hat{\tau})$  be defined for  $i = 1, 2, 3$  by the relations

$$\hat{\phi}_i(\hat{Q}_j) = \delta_{ij}, \quad 1 \leq i, j \leq 3.$$

Then, for  $\hat{w} \in H^2(\hat{\tau})$

$$\begin{aligned} \| I_{\hat{\tau}}\hat{w} \|_{2,\hat{\tau}} &= \| \hat{w}(\hat{Q}_1)\hat{\phi}_1 + \hat{w}(\hat{Q}_2)\hat{\phi}_2 + \hat{w}(\hat{Q}_3)\hat{\phi}_3 \|_{2,\hat{\tau}} \leq \sum_{i=1}^3 | \hat{w}(\hat{Q}_i) | \| \hat{\phi}_i \|_{2,\hat{\tau}} \\ &\leq C'(\hat{\tau}) \max_{\hat{x} \in \hat{\tau}} | \hat{w}(\hat{x}) | \leq \tilde{C}(\hat{\tau}) \| \hat{w} \|_{2,\hat{\tau}}, \text{ by Sobolev's theorem in } \mathbb{R}^2. \end{aligned}$$

As a consequence of (5.27) and (5.28) note that for  $m = 0, 1, 2$  and  $\hat{v} \in H^2(\hat{\tau})$

$$\begin{aligned} | \hat{v} - I_{\hat{\tau}}\hat{v} |_{m,\hat{\tau}} &\leq | \hat{v} - \hat{p} |_{m,\hat{\tau}} + | I_{\hat{\tau}}(\hat{v} - \hat{p}) |_{m,\hat{\tau}} \leq \| \hat{v} - \hat{p} \|_{2,\hat{\tau}} + \| I_{\hat{\tau}}(\hat{v} - \hat{p}) \|_{2,\hat{\tau}} \\ &\leq \| \hat{v} - \hat{p} \|_{2,\hat{\tau}} + \tilde{C}(\hat{\tau}) \| \hat{v} - \hat{p} \|_{2,\hat{\tau}} \\ &\leq \tilde{C}(\hat{\tau}) \| \hat{v} - \hat{p} \|_{2,\hat{\tau}} \quad \forall \hat{p} \in \mathbb{P}_1(\hat{\tau}). \end{aligned} \quad (5.29)$$

We invoke now the

(iii) **Bramble–Hilbert Lemma**, which in our case asserts that there exists a constant  $C^*(\hat{\tau})$  such that:

$$\min_{\hat{p} \in \mathbb{P}_1(\hat{\tau})} \| \hat{v} - \hat{p} \|_{2,\hat{\tau}} \leq C^*(\hat{\tau}) | \hat{v} |_{2,\hat{\tau}} \quad \forall \hat{v} \in H^2(\hat{\tau}). \quad (5.30)$$

(We postpone for the moment the proof of this important result; we shall prove it later in more generality).

Putting together (5.29) and (5.30) yields now

$$| \hat{v} - I_{\hat{\tau}}\hat{v} |_{m,\hat{\tau}} \leq \tilde{C}(\hat{\tau}) \min_{\hat{p} \in \mathbb{P}_1(\hat{\tau})} \| \hat{v} - \hat{p} \|_{2,\hat{\tau}} \leq C(\hat{\tau}) | \hat{v} |_{2,\hat{\tau}},$$



which is the advertized result (5.26).  $\square$

Before turning to the proof of (5.30) we first complete the argument and show how (5.26) implies (5.15). To do this, recall that if  $\hat{w}$  is defined on  $\hat{\tau}$ , then for  $x = F_\tau(\hat{x})$  we have  $\hat{w}(\hat{x}) = w(x)$ , where  $w$  is the image function, i.e. where  $w = \hat{w} \circ F_\tau^{-1}$ , defined on  $\tau$ . Note that the  $\widehat{\phantom{x}}$  operation is linear, in the sense that  $\widehat{\mu u + \lambda v} = \mu \hat{u} + \lambda \hat{v}$ ,  $\mu, \lambda \in \mathbb{R}$ . (Indeed  $(\widehat{\mu u + \lambda v})(\hat{x}) = (\mu u + \lambda v)(x) = \mu u(x) + \lambda v(x) = \mu \hat{u}(\hat{x}) + \lambda \hat{v}(\hat{x})$ ).

Hence, if  $I_\tau v$  is the (local) linear interpolant on  $\mathbb{P}_1(\tau)$  of  $v \in H^2(\tau)$ , we have

$$(v - I_\tau v)^\wedge = \hat{v} - \widehat{I_\tau v}. \quad (5.31)$$

Now,  $\widehat{I_\tau v} = I_{\hat{\tau}} \hat{v}$ . To see this, note that for  $i = 1, 2, 3$

$$(\widehat{I_\tau v})(\hat{Q}_i) = (I_\tau v)(Q_i) = v(Q_i) = \hat{v}(\hat{Q}_i) = (I_{\hat{\tau}} \hat{v})(\hat{Q}_i)$$

i.e.  $\widehat{I_\tau v}$  and  $I_{\hat{\tau}} \hat{v}$  (which are real-valued functions defined on  $\bar{\hat{\tau}}$ ) coincide at the vertices  $\hat{Q}_i$  of  $\hat{\tau}$ . However  $I_{\hat{\tau}} \hat{v} \in \mathbb{P}_1(\hat{\tau})$  and  $(\widehat{I_\tau v})(\hat{x}) = (I_\tau v)(x) = (I_\tau v)(F_\tau(\hat{x})) \in \mathbb{P}_1(\hat{\tau})$ . Hence  $(\widehat{I_\tau v})(\hat{x}) = (I_{\hat{\tau}} \hat{v})(\hat{x}) \forall \hat{x} \in \hat{\tau}$ . Therefore, (5.31) gives that

$$(v - I_\tau v)^\wedge = \hat{v} - I_{\hat{\tau}} \hat{v}. \quad (5.32)$$

Now, using (5.21) for  $m = 0, 1, 2$ , we have ( $h_{\hat{\tau}} = \text{diam}(\hat{\tau})$ )

$$\begin{aligned} |v - I_\tau v|_{m,\tau} &\leq \tilde{C} |B_\tau^{-1}|^m |\det B_\tau|^{\frac{1}{2}} |v - I_\tau v|_{m,\hat{\tau}} \\ \text{(by (5.24), (5.26), (5.32))} &\leq \tilde{C} \frac{h_{\hat{\tau}}^m}{\rho_\tau^m} |\det B_\tau|^{\frac{1}{2}} C(\hat{\tau}) |\hat{v}|_{2,\hat{\tau}} \\ &\leq C'(\hat{\tau}) \frac{1}{\rho_\tau^m} |\det B_\tau|^{\frac{1}{2}} |\hat{v}|_{2,\hat{\tau}} \\ \text{(using again (5.20))} &\leq C''(\hat{\tau}) \frac{1}{\rho_\tau^m} |\det B_\tau|^{\frac{1}{2}} |B_\tau|^2 |\det B_\tau|^{-\frac{1}{2}} |v|_{2,\tau} \\ \text{(by (5.24))} &\leq C'''(\hat{\tau}) \frac{h_\tau^2}{\rho_\tau^m} \frac{1}{\rho_{\hat{\tau}}^2} |v|_{2,\tau} \\ &\leq C'''(\hat{\tau}) \frac{h_\tau^2}{\rho_\tau^m} |v|_{2,\tau}, \end{aligned}$$

which is (5.15), since  $C'''(\hat{\tau}) \equiv C$  is a constant depending only on the fixed reference triangle  $\hat{\tau}$  and is, hence, independent of  $\mathcal{T}_h$ .

We finally turn to proving (5.30). This will be done in some generality in the following Proposition in which  $\Omega$  is assumed to be a bounded, Lipschitz domain in  $\mathbb{R}^N$ .

**Proposition 5.1** (Bramble–Hilbert/Deny–Lions). *For  $k \geq 1$  there exists a constant  $C_k = C_k(\Omega)$  such that for every  $u \in H^k(\Omega)$*

$$\min_{p \in \mathbb{P}_{k-1}} \|u + p\|_k \leq C_k |u|_k. \quad (5.33)$$

**Remark:** (5.33) may be viewed as a type of “Taylor’s theorem”. Also, note that obviously  $|u|_k \leq \min_{p \in \mathbb{P}_{k-1}} \|u + p\|_k$ . Hence (5.33) implies that  $|\cdot|_k$  is a norm, equivalent to the quotient norm  $\min_{p \in \mathbb{P}_{k-1}} \|\cdot + p\|_k$  on the space  $H^k/\mathbb{P}_{k-1}$ . (From now on  $H^k = H^k(\Omega)$ ).

**Proof.** The estimate (5.33) is a direct consequence of two facts:

(i) For each  $u \in H^k$  there exists a unique  $q \in \mathbb{P}_{k-1}$  such that

$$\forall \alpha : |\alpha| \leq k-1, \quad \int_{\Omega} D^{\alpha} q \, dx = \int_{\Omega} D^{\alpha} u \, dx. \quad (5.34)$$

(ii) There exists  $C_k = C_k(\Omega)$  such that  $\forall u \in H^k$ :

$$\|u\|_k \leq C_k \left\{ |u|_k^2 + \sum_{|\alpha| < k} \left( \int_{\Omega} D^{\alpha} u \, dx \right)^2 \right\}^{\frac{1}{2}}. \quad (5.35)$$

Indeed, if (i) and (ii) hold, then for  $u \in H^k$ , with  $q$  as in (i),

$$\begin{aligned} \min_{p \in \mathbb{P}_{k-1}} \|u + p\|_k &\leq \|u - q\|_k \stackrel{(ii)}{\leq} C_k \left\{ |u - q|_k^2 + \sum_{|\alpha| < k} \left( \int_{\Omega} (D^{\alpha} u - D^{\alpha} q) \, dx \right)^2 \right\}^{\frac{1}{2}} \\ &\stackrel{(i)}{=} C_k |u - q|_k = C_k |u|_k \end{aligned}$$

since  $q \in \mathbb{P}_{k-1} \Rightarrow D^{\alpha} q = 0$  for  $|\alpha| = k$ .

Therefore, (5.33) is a consequence of (i) and (ii).

We now prove (i) and (ii).

(i). Let  $u \in H^k$  be given. We shall construct a polynomial  $q \in \mathbb{P}_{k-1}$  such that the relations (5.34) hold. Let  $q$  be of the form

$$q(x) = \sum_{m=0}^{k-1} \sum_{|\alpha|=m} c_{\alpha} x^{\alpha} \equiv \sum_{m=0}^{k-1} \sum_{|\alpha|=m} c_{\alpha_1 \dots \alpha_N} x_1^{\alpha_1} x_2^{\alpha_2} \dots x_N^{\alpha_N}.$$

We shall determine the unknown coefficients  $c_{\alpha}$ ,  $|\alpha| \leq k-1$ , from the relations (5.34) which represent a linear system of equations for the  $c_{\alpha}$  of size  $M_k \times M_k$  where  $M_k$  is the

number of multiindices  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)$  with  $|\alpha| \leq k - 1$ . We first determine the coefficients  $c_\alpha$  multiplying the terms  $x^\alpha$  of highest degree, i.e. the  $c_\alpha$  with  $|\alpha| = k - 1$ . For a multiindex  $\alpha$  with  $|\alpha| = k - 1$  we have

$$D^\alpha q = D^\alpha \left( \sum_{\beta: |\beta|=k-1} c_\beta x^\beta \right) + D^\alpha \underbrace{\left( \sum_{\beta: |\beta|<k-1} c_\beta x^\beta \right)}_{\in \mathbb{P}_{k-2}} = c_\alpha \alpha!,$$

where  $\alpha! \equiv \alpha_1! \alpha_2! \dots \alpha_N!$  for  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)$ .

(To see the last equality, consider any one term  $D^\alpha(C_\beta x^\beta)$  of the sum

$$D^\alpha \left( \sum_{\beta: |\beta|=k-1} c_\beta x^\beta \right).$$

Hence, for such a term,  $|\alpha| = |\beta| = k - 1$ . If  $\alpha \neq \beta$  then  $D^\alpha(C_\beta x^\beta) = 0$ . For if  $\alpha \neq \beta$  there must exist an index  $j$ ,  $1 \leq j \leq N$ , such that  $\beta_j \neq \alpha_j$ . Then for the corresponding factor in  $D^\alpha x^\beta$  we will have  $\left(\frac{\partial}{\partial x_j}\right)^{\alpha_j} x_j^{\beta_j} = 0$ . Now, if  $\alpha = \beta$  we have  $D^\alpha(x^\beta) = D^\alpha(x^\alpha) = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} x_1^{\beta_1} \dots \left(\frac{\partial}{\partial x_N}\right)^{\alpha_N} x_N^{\beta_N} = \alpha_1! \alpha_2! \dots \alpha_N! \equiv \alpha!$ .

Hence, the relation (5.34) for  $\alpha$  such that  $|\alpha| = k - 1$  yield

$$\begin{aligned} \int_{\Omega} D^\alpha u \, dx &= \int_{\Omega} D^\alpha q \, dx = \int_{\Omega} c_\alpha \alpha! \, dx = c_\alpha \alpha! \mu(\Omega) \Rightarrow \\ c_\alpha &= \frac{\int_{\Omega} D^\alpha u \, dx}{\alpha! \mu(\Omega)}, \quad |\alpha| = k - 1, \end{aligned} \quad (5.36)$$

i.e. the coefficients of  $q$  with  $|\alpha| = k - 1$  have been determined.

Write now  $q_{k-1}(x) = \sum_{|\alpha|=k-1} c_\alpha x^\alpha$  ( $q_{k-1}$  is now known). Hence

$$q(x) = \sum_{|\alpha|=k-2} c_\alpha x^\alpha + q_{k-1} + s(x),$$

where  $s \in \mathbb{P}_{k-3}$ , from which, for  $|\alpha| = k - 2$ , as before

$$D^\alpha q = c_\alpha \alpha! + D^\alpha q_{k-1}.$$

Therefore, (5.34) for  $|\alpha| = k - 2$  yield

$$\int_{\Omega} D^\alpha u \, dx = c_\alpha \alpha! \mu(\Omega) + \int_{\Omega} D^\alpha q_{k-1} \, dx$$

from which the  $c_\alpha$ ,  $|\alpha| = k - 2$  are determined. We continue in the same fashion to determine all  $c_\alpha$ .

Note that the polynomial  $q$  satisfying (5.34) is necessarily unique: if two such  $q_1, q_2 \in \mathbb{P}_{k-1}$  exist, then  $q = q_1 - q_2$  will satisfy  $\int_{\Omega} D^{\alpha} q dx = 0, |\alpha| \leq k-1$ , which is a homogeneous linear system for the associated  $c_{\alpha}$ . The formulas (5.36) would yield now  $c_{\alpha} = 0, |\alpha| = k-1$ . Therefore  $q_{k-1} = 0$ , i.e.  $c_{\alpha} = 0, |\alpha| = k-2$ , and so on, implying finally that  $q = 0$ .

(ii). We argue by contradiction: Suppose (5.36) does not hold. This means that for any constant  $C > 0$  there exists a  $u \in H^k$  such that

$$\|u\|_k > C \left\{ |u|_k^2 + \sum_{|\alpha| < k} \left( \int_{\Omega} D^{\alpha} u dx \right)^2 \right\}^{\frac{1}{2}},$$

i.e. such that

$$C \left\{ \frac{|u|_k^2}{\|u\|_k^2} + \frac{\sum_{|\alpha| < k} \left( \int_{\Omega} D^{\alpha} u dx \right)^2}{\|u\|_k^2} \right\}^{\frac{1}{2}} < 1.$$

This implies that for any constant  $C > 0$  there exists a  $v \in H^k$  with  $\|v\|_k = 1$  (take  $v = u/\|u\|_k$ ) such that

$$C \left\{ |v|_k^2 + \sum_{|\alpha| < k} \left( \int_{\Omega} D^{\alpha} v dx \right)^2 \right\}^{\frac{1}{2}} < 1.$$

Take  $C = n, n = 1, 2, \dots$ . Hence, there exists a sequence  $\{u_n\}$  of functions in  $H^k$  with  $\|u_n\|_k = 1$ , such that

$$|u_n|_k^2 + \sum_{|\alpha| < k} \left( \int_{\Omega} D^{\alpha} u_n dx \right)^2 < \frac{1}{n^2}. \quad (5.37)$$

We now use the fact (“Rellich’s theorem” cf. Adams) that for a domain such as  $\Omega$  (in fact for any bounded, Lipschitz domain) and for  $k \geq 1$ ,  $H^k$  may be compactly imbedded in  $H^{k-1}$ , in the sense that every bounded subset of  $H^k$  is relatively compact when viewed as a subset of  $H^{k-1}$ . This means that every bounded sequence in  $H^k$  has a subsequence which converges in the  $H^{k-1}$  norm. Therefore the bounded sequence  $u_n \in H^k$  ( $\|u_n\|_k = 1$ ) has a subsequence, which we denote again by  $u_n$  without loss of generality, and which converges in  $H^{k-1}$ . But (5.37) yields that  $|u_n|_k \rightarrow 0, n \rightarrow \infty$ , i.e. that  $D^{\alpha} u_n \rightarrow 0$  in  $L^2$  for  $|\alpha| = k$ . Since  $u_n$  converges in  $H^{k-1}$  already, we conclude that  $u_n$  converges in  $H^k$ . Let the limit of  $\{u_n\}$  in  $H^k$  be denoted by  $w$ . Since  $\|u_n\|_k = 1 \Rightarrow \|w\|_k = 1$ . But since  $D^{\alpha} u_n \rightarrow 0$  in  $L^2, |\alpha| = k$ , we conclude that  $D^{\alpha} w = 0, |\alpha| = k$ , i.e.  $w \in \mathbb{P}_{k-1}$ .

Now, (5.37) also yields  $\sum_{|\alpha|<k} \left( \int_{\Omega} D^{\alpha} u_n dx \right)^2 \rightarrow 0, n \rightarrow \infty$ , i.e. that

$$\int_{\Omega} D^{\alpha} u_n dx \rightarrow 0, \quad \text{for } |\alpha| \leq k - 1.$$

We conclude, since  $u_n \rightarrow w$  in  $H^k$ , that  $\int_{\Omega} D^{\alpha} w = 0$  for  $|\alpha| \leq k - 1$ . Since the polynomial  $w \in \mathbb{P}_{k-1}$  satisfies  $\int_{\Omega} D^{\alpha} w = 0$  for  $|\alpha| \leq k - 1$ , the construction in (i) yields that  $w = 0$ , a contradiction, since  $\|w\|_k = 1$ ; q.e.d.  $\square$

### 5.3 Implementation of the finite element method with $P_1$ triangles

In this section we shall study the details of the implementation of the standard Galerkin / finite element method for a simple elliptic boundary-value problem on a polygonal plane domain, based on the ideas of MODULEF, cf. Bernadou et al., 1985. We seek  $u(x) = u(x_1, x_2)$  defined on  $\bar{\Omega}$ , where  $\Omega$  is a convex, polygonal domain in  $\mathbb{R}^2$ , and satisfying

$$\left. \begin{aligned} -\Delta u + a(x)u &= f(x), & x \in \Omega, \\ u &= 0, & x \in \partial\Omega. \end{aligned} \right\} \quad (5.38)$$

Here  $f, a$  are given, say continuous, functions on  $\bar{\Omega}$  with  $a \geq 0$ . The weak formulation of the problem is, as usual, to seek  $u \in \overset{\circ}{H}^1 = \overset{\circ}{H}^1(\Omega)$ , such that

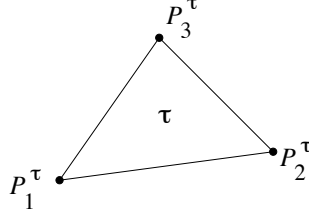
$$\int_{\Omega} (\nabla u \cdot \nabla v + a(x)uv) dx = \int_{\Omega} f v dx, \quad \forall v \in \overset{\circ}{H}^1. \quad (5.39)$$

Let  $S_h$  be a finite-dimensional subspace of  $\overset{\circ}{H}^1$ . The *standard Galerkin method* for the approximation of the solution of (5.39) consists in seeking  $u_h \in S_h$  such that

$$\int_{\Omega} (\nabla u_h \cdot \nabla v_h + a(x)u_h v_h) dx = \int_{\Omega} f v_h dx, \quad \forall v_h \in S_h. \quad (5.40)$$

We take  $S_h$  to be the space of continuous functions on  $\bar{\Omega}$  that vanish on  $\partial\Omega$  and are polynomials of degree at most 1 on each triangle  $\tau$  of a triangulation  $\mathcal{T}_h$  of  $\Omega$ . (For notation cf. Section 5.2). Accordingly, we refer to (5.40) as the “standard Galerkin / finite element method with  $P_1$  triangles”.

(i) **Local degrees of freedom.**



On each triangle  $\tau \in \mathcal{T}_h$ ,  $u_h$  is represented in terms of the *local basis functions*  $\varphi_j^\tau(x)$ ,  $j = 1, 2, 3$ , (which are  $\mathbb{P}_1$  polynomials on  $\tau$  such that  $\varphi_j^\tau(P_i^\tau) = \delta_{ij}$ ,  $1 \leq i, j \leq 3$ ) as

$$u_h(x) = \sum_{j=1}^3 \varphi_j^\tau(x) u_h(P_j^\tau), \quad x \in \tau. \quad (5.41)$$

The values  $\{u_h(P_j^\tau)\}$ ,  $1 \leq j \leq 3$ , coefficients of  $\varphi_j^\tau(x)$  in the linear combination of the  $\{\varphi_j^\tau(x)\}$  in the r.h.s. of (5.41), are, in our case, the “local degrees of freedom” that determine  $u_h(x)$  uniquely on  $\tau$ . (In general, there are  $N_\tau$  degrees of freedom – not all function values of  $u_h$  necessarily – on each triangle  $\tau$ . In our case  $N_\tau = 3 \forall \tau$ ). Introducing the  $1 \times 3$  *matrix of local basis functions*  $\Phi^\tau = \Phi^\tau(x)$  by

$$\Phi^\tau = [\varphi_1^\tau(x), \varphi_2^\tau(x), \varphi_3^\tau(x)] \quad (5.42)$$

and the  $3 \times 1$  vector

$$U^\tau = [u_h(P_1^\tau), u_h(P_2^\tau), u_h(P_3^\tau)]^T \quad (5.43)$$

of the *local degrees of freedom*, we may rewrite (5.41) as

$$u_h(x) = \Phi^\tau(x) U^\tau. \quad (5.44)$$

Let  $\nabla u_h = [\frac{\partial u_h}{\partial x_1}, \frac{\partial u_h}{\partial x_2}]^T$  denote the gradient of  $u_h$ . Then, (5.41) gives

$$\frac{\partial u_h}{\partial x_i} = \sum_{j=1}^3 \frac{\partial \varphi_j^\tau}{\partial x_i} u_h(P_j^\tau), \quad i = 1, 2,$$

i.e. that

$$\nabla u_h(x) = D\Phi^\tau(x) \cdot U^\tau, \quad x \in \tau, \quad (5.45)$$

where  $D\Phi^\tau = D\Phi^\tau(x)$ ,  $x \in \tau$ , denotes the  $2 \times 3$  matrix

$$D\Phi^\tau = \begin{pmatrix} \frac{\partial \varphi_1^\tau}{\partial x_1} & \frac{\partial \varphi_2^\tau}{\partial x_1} & \frac{\partial \varphi_3^\tau}{\partial x_1} \\ \frac{\partial \varphi_1^\tau}{\partial x_2} & \frac{\partial \varphi_2^\tau}{\partial x_2} & \frac{\partial \varphi_3^\tau}{\partial x_2} \end{pmatrix}. \quad (5.46)$$

The Galerkin equations (5.40) become, since  $\bar{\Omega} = \cup_{\tau \in \mathcal{T}_h} \bar{\tau}$ ,

$$\sum_{\tau \in \mathcal{T}_h} \int_{\tau} (\nabla u_h \cdot \nabla v_h + a(x) u_h v_h) dx = \sum_{\tau \in \mathcal{T}_h} \int_{\tau} f(x) v_h dx, \quad \forall v_h \in S_h. \quad (5.47)$$

We shall write (5.47) in matrix–vector form in terms of the local basis functions and the local degrees of freedom  $\{U^\tau\}$  and  $\{V^\tau\}$  of  $u_h$  and  $v_h$ , respectively. Writing

$$u_h = \Phi^\tau U^\tau, \quad v_h = \Phi^\tau V^\tau, \quad x \in \tau,$$

we have

$$\begin{aligned} u_h v_h &= (\Phi^\tau U^\tau)(\Phi^\tau V^\tau) = (V^\tau)^T (\Phi^\tau)^T \Phi^\tau U^\tau, \quad x \in \tau. \\ \nabla u_h \cdot \nabla v_h &= (D\Phi^\tau U^\tau) \cdot (D\Phi^\tau V^\tau) = (D\Phi^\tau V^\tau)^T (D\Phi^\tau U^\tau) \\ &= (V^\tau)^T (D\Phi^\tau)^T D\Phi^\tau U^\tau, \quad x \in \tau. \\ f v_h &= f \Phi^\tau V^\tau = (\Phi^\tau V^\tau)^T = (V^\tau)^T (\Phi^\tau)^T f, \quad x \in \tau. \end{aligned}$$

Using these expressions in (5.47) we have

$$\begin{aligned} \sum_{\tau \in \mathcal{T}_h} \int_{\tau} \{ (V^\tau)^T (D\Phi^\tau)^T D\Phi^\tau U^\tau + a(x) (V^\tau)^T (\Phi^\tau)^T \Phi^\tau U^\tau \} dx &= \\ = \sum_{\tau \in \mathcal{T}_h} \int_{\tau} (V^\tau)^T (\Phi^\tau)^T f(x) dx, \quad \forall v_h \in S_h. \end{aligned} \quad (5.48)$$

Let  $K^\tau$ ,  $M^\tau$  denote, respectively, the  $3 \times 3$  local *stiffness* and *mass* matrix. These are given by the formulas

$$K^\tau := \int_{\tau} (D\Phi^\tau)^T D\Phi^\tau dx, \quad (5.49)$$

$$M^\tau := \int_{\tau} a(x) (\Phi^\tau)^T \Phi^\tau dx. \quad (5.50)$$

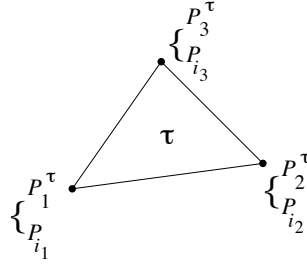
Let also  $A^\tau := K^\tau + M^\tau$ , and  $b^\tau$  be the  $3 \times 1$  vector

$$b^\tau := \int_{\tau} f(x) (\Phi^\tau)^T dx. \quad (5.51)$$

Using these local quantities in (5.48) yields the desired matrix–vector form of (5.47):

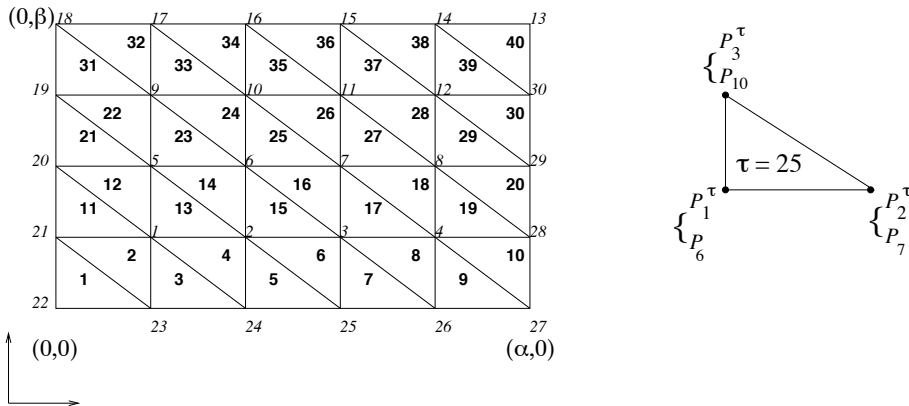
$$\sum_{\tau \in \mathcal{T}_h} (V^\tau)^T A^\tau U^\tau = \sum_{\tau \in \mathcal{T}_h} (V^\tau)^T b^\tau, \quad \forall v_h \in S_h. \quad (5.52)$$

(ii) The global - to - local degrees of freedom map.



Suppose that the triangulation  $\mathcal{T}_h$  consists of  $N$  nodes (vertices) (including the nodes on the boundary  $\partial\Omega$ ), that are denoted by  $P_i, i = 1, 2, \dots, N$ , in some *global indexing scheme*. Then the vertices  $P_1^\tau, P_2^\tau, P_3^\tau$  of of a given triangle  $\tau \in \mathcal{T}_h$  correspond to the points  $P_{i_1}, P_{i_2}, P_{i_3}$ , respectively, in the global enumeration. We would like to find an efficient way of expressing the correspondence  $P_{i_k} \rightarrow P_k^\tau$ . Specifically, we are really interested in expressing the local degrees of freedom, i.e. in our case the values  $u_h(P_k^\tau)$ ,  $k = 1, 2, 3$ , in terms of the global degrees of freedom, i.e. the values  $u_h(P_i)$ ,  $i = 1, 2, \dots, N$ .

Let us consider an example: let  $\Omega$  be the rectangle  $(0, \alpha) \times (0, \beta)$ . Subdivide it in 20 rectangles of size  $\Delta x_1 \times \Delta x_2$ , where  $\Delta x_1 = \frac{\alpha}{5}$ ,  $\Delta x_2 = \frac{\beta}{4}$  and then in 40 equal triangles as shown, by bisecting the rectangles.



We number the triangles  $\tau$  from 1 to 40 as shown and introduce the following global indexing scheme for the nodes: The interior nodes are the points  $P_i$  shown, with  $i = 1, 2, \dots, 12$ , and the boundary nodes are the points  $P_i, i = 13, \dots, 30$ . Here  $N = 30$ ,



therefore. For example, the triangle with  $\tau = 25$  is defined by the nodes  $P_6, P_7$  and  $P_{10}$  in the global numbering. These coincide with the local nodes  $P_1^\tau, P_2^\tau, P_3^\tau$ , ( $\tau = 25$ ), respectively. For  $\tau = 25$ , let the local degrees of freedom be represented by the  $3 \times 1$  vector  $U^\tau = [U_1^\tau, U_2^\tau, U_3^\tau]^T$  ( $U_j^\tau = u_h(P_j^\tau)$ ,  $j = 1, 2, 3$ ). The global degrees of freedom  $u_h(P_i)$ ,  $i = 1, 2, \dots, N = 30$ , are arranged in the  $30 \times 1$  vector  $U = [U_1, \dots, U_{30}]^T$ . Clearly, we have  $U^\tau = \mathcal{G}^\tau U$ , where  $\mathcal{G}^\tau$  is a  $3 \times 30$  matrix whose elements are 0 or 1 (Boolean matrix). In our case, i.e. for  $\tau = 25$ , we have

$$\begin{pmatrix} U_1^\tau \\ U_2^\tau \\ U_3^\tau \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \overset{6\text{th}}{1} & \overset{7\text{th}}{0} & 0 & 0 & \overset{10\text{th}}{0} & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_{30} \end{pmatrix}.$$

In general, let  $\mathcal{T}_h$  be the triangulation consisting of triangles  $\{\tau\}$  that we label, abusing notation a bit, as  $\tau = 1, 2, \dots, J$ . For each  $\tau$ , let

$$i = g(\tau, j)$$

be the map that associates the local index  $j$ ,  $1 \leq j \leq 3$ , of the local vertices  $P_j^\tau$  to the global index  $i$ ,  $1 \leq i \leq 30$ , of the corresponding points  $P_i$ . Then  $g$  is a function defined on the set  $\{1, 2, \dots, J\} \times \{1, 2, 3\}$  with values onto the set  $\{1, 2, \dots, N\}$  that can be easily stored. In our example, the values of  $i$  for  $\tau = 25$  and  $j = 1, 2, 3$  are

$$\begin{aligned} g(25, 1) &= 6, \\ g(25, 2) &= 7, \\ g(25, 3) &= 10. \end{aligned}$$

Let  $\mathcal{G}^\tau$ , for each  $\tau$ , denote the  $3 \times N$  matrix whose elements are given by

$$\mathcal{G}_{kl}^\tau = \delta_{g(\tau, k), l} \quad 1 \leq k \leq 3, \quad 1 \leq l \leq N, \quad (5.53)$$

where  $\delta_{i,j}$  is the Kronecker delta, i.e.  $\delta_{i,j} = 1$  if  $i = j$ ,  $\delta_{i,j} = 0$  if  $i \neq j$ . Then, the relation between the global degrees of freedom vector  $U$ ,  $U_i = u_h(P_i)$ ,  $1 \leq i \leq N$ , and the local degrees of freedom vector  $U^\tau$  on  $\tau$ ,  $U_j^\tau = u_h(P_j^\tau)$ ,  $j = 1, 2, 3$ , is expressed as

$$U^\tau = \mathcal{G}^\tau U. \quad (5.54)$$

Substituting (5.54), and the corresponding expression  $V^\tau = \mathcal{G}^\tau V$ , into (5.52), we have

$$\sum_{\tau \in \mathcal{T}_h} V^T (\mathcal{G}^\tau)^T A^\tau \mathcal{G}^\tau U = \sum_{\tau \in \mathcal{T}_h} V^T (\mathcal{G}^\tau)^T b^\tau, \quad \forall v_h \in S_h$$

or, since  $V, U$  do not depend on  $\tau$ ,

$$V^T \left\{ \sum_{\tau \in \mathcal{T}_h} (\mathcal{G}^\tau)^T A^\tau \mathcal{G}^\tau \right\} U = V^T \sum_{\tau \in \mathcal{T}_h} (\mathcal{G}^\tau)^T b^\tau, \quad \forall v_h \in S_h.$$

Hence, (5.52) in terms of global degrees of freedom may be written as

$$V^T A U = V^T b, \quad \forall V \in \mathbb{R}^N \text{ such that } v_h \in S_h, \quad (5.55)$$

where  $A$  is the  $N \times N$  matrix defined by

$$A := \sum_{\tau \in \mathcal{T}_h} (\mathcal{G}^\tau)^T A^\tau \mathcal{G}^\tau, \quad (5.56)$$

and  $b$  the  $N \times 1$  vector given by

$$b := \sum_{\tau \in \mathcal{T}_h} (\mathcal{G}^\tau)^T b^\tau. \quad (5.57)$$

**(iii) Assembly of  $A$  and  $b$ .**

The matrix  $A$  and the vector  $b$  defined by (5.56) and (5.57) should be *assembled* from their local contributions  $A^\tau$  and  $b^\tau$ . In doing this we should *not* form  $\mathcal{G}^\tau$  and perform the indicated matrix–matrix and matrix–vector operations, since this would be very costly in terms of storage and number of operations. Instead, recalling the definition (5.53) of  $\mathcal{G}^\tau$ , we have, for the vector  $b = \{b_i\}$ ,  $1 \leq i \leq N$ :

$$b_i = \sum_{\tau \in \mathcal{T}_h} ((\mathcal{G}^\tau)^T b^\tau)_i = \sum_{\tau \in \mathcal{T}_h} \left( \sum_{j=1}^3 \mathcal{G}_{ji}^\tau b_j^\tau \right) = \sum_{\tau \in \mathcal{T}_h} \left( \sum_{j=1}^3 \delta_{g(\tau,j),i} b_j^\tau \right).$$

We conclude that the following algorithm computes the  $b_i$ :

$$\left[ \begin{array}{l} \text{For } i = 1, 2, \dots, N \text{ do:} \\ \quad b_i = 0 \\ \text{For } i = 1, 2, \dots, N \text{ do:} \\ \quad \left[ \begin{array}{l} \text{For } \tau \in \mathcal{T}_h \text{ do:} \\ \quad \left[ \begin{array}{l} \text{For } j = 1, 2, 3 \text{ do:} \\ \quad b_i = b_i + \delta_{g(\tau,j),i} b_j^\tau. \end{array} \right. \end{array} \right. \end{array} \right.$$

Since  $\delta_{g(\tau,j),i} = 0$  unless  $i = g(\tau, j)$ , the last term contributes only to the  $b_i$  with  $i = g(\tau, j)$ . Hence, the above algorithm can be written, more compactly, in the form:

$$\left[ \begin{array}{l} \text{For } i = 1, 2, \dots, N \text{ do:} \\ b_i = 0 \\ \text{For } \tau \in \mathcal{T}_h \text{ do:} \\ \left[ \begin{array}{l} \text{For } j = 1, 2, 3 \text{ do:} \\ b_{g(\tau,j)} = b_{g(\tau,j)} + b_j^\tau, \end{array} \right. \end{array} \right. \quad (5.58)$$

i.e. the  $b_j^\tau$  is added to the previous value of that  $b_i$  that has  $i = g(\tau, j)$ . Similarly by (5.56) we have

$$A_{ij} = \sum_{\tau \in \mathcal{T}_h} \left( \sum_{k,l=1}^3 \mathcal{G}_{ki}^\tau A_{kl}^\tau \mathcal{G}_{lj}^\tau \right) = \sum_{\tau \in \mathcal{T}_h} \left( \sum_{k,l=1}^3 \delta_{g(\tau,k),i} A_{kl}^\tau \delta_{g(\tau,l),j} \right),$$

i.e. the term  $A_{kl}^\tau$  contributes to the element  $A_{ij}$  if  $i = g(\tau, k)$  and  $j = g(\tau, l)$ . Consequently, the following algorithm may be used to assemble the matrix  $A$ :

$$\left[ \begin{array}{l} \text{For } i = 1, 2, \dots, N \text{ do:} \\ \left[ \begin{array}{l} \text{For } j = 1, 2, \dots, N \text{ do:} \\ A_{ij} = 0 \end{array} \right. \\ \text{For } \tau \in \mathcal{T}_h \text{ do:} \\ \left[ \begin{array}{l} \text{For } k = 1, 2, 3 \text{ do:} \\ \left[ \begin{array}{l} \text{For } l = 1, 2, 3 \text{ do:} \\ A_{g(\tau,k),g(\tau,l)} = A_{g(\tau,k),g(\tau,l)} + A_{kl}^\tau. \end{array} \right. \end{array} \right. \end{array} \right. \quad (5.59)$$

The algorithms (5.58) and (5.59) implement the *assembly* of the (global) matrix  $A$  and vector  $b$  in the equations (5.55) from their local parts  $A^\tau$  and  $b^\tau$ .

**(iv) Reduction of (5.55) to a linear system of equations.**

The components of the global vector of degrees of freedom  $U \in \mathbb{R}^N$ , may be ordered (although this is not necessary always) so that  $U$  be of the form

$$U = [U_1, U_2, \dots, U_{N-N_0}, U_{N-N_0+1}, \dots, U_N]^T,$$

so that the degrees of freedom  $U_1, U_2, \dots, U_{N-N_0}$  are the values of  $u_h$  at the interior nodes  $P_1, P_2, \dots, P_{N-N_0}$  and  $U_{N-N_0+1}, \dots, U_N$  are the values of  $u_h$  at the  $N_0$  boundary nodes  $P_{N-N_0+1}, \dots, P_N$ . (In the example on p. 114,  $N_0 = 18$ ,  $N = 30$ ).

Then, writing

$$U = \begin{bmatrix} U_I \\ U_{II} \end{bmatrix}, \quad \text{with } U_I \in \mathbb{R}^{N-N_0}, \quad U_{II} \in \mathbb{R}^{N_0},$$

and partitioning conformably  $V$ ,  $A$  and  $b$  in (5.55), we can write it in the form

$$[V_I^T \quad V_{II}^T] \begin{bmatrix} A_{I,I} & A_{I,II} \\ A_{II,I} & A_{II,II} \end{bmatrix} \begin{bmatrix} U_I \\ U_{II} \end{bmatrix} = [V_I^T \quad V_{II}^T] \begin{bmatrix} b_I \\ b_{II} \end{bmatrix}, \quad \forall v_h \in S_h. \quad (5.60)$$

Since  $S_h$  is such that  $v_h \in S_h \Leftrightarrow V_I \in \mathbb{R}^{N-N_0}$ ,  $V_{II} = 0$  in  $\mathbb{R}^{N_0}$ , and since  $u_h \in S_h$  as well, we rewrite the above as

$$V_I^T A_{I,I} U_I = V_I^T b_I \quad \forall V_I \in \mathbb{R}^{N-N_0},$$

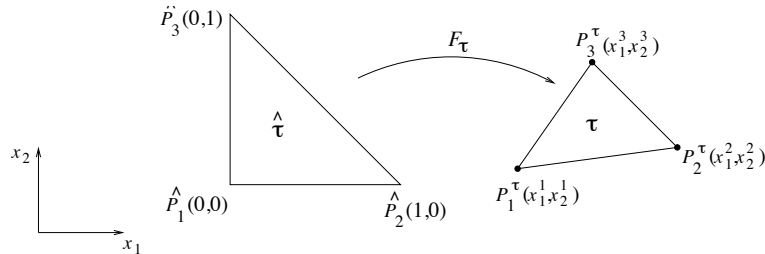
which is of course equivalent to the  $(N - N_0) \times (N - N_0)$  system

$$A_{I,I} U_I = b_I. \quad (5.61)$$

The reduction of the  $N \times N$  system (5.60) to the system (5.61) is usually referred to as “taking into account the boundary conditions of the problem”.

**(v) Computing  $K^\tau$ ,  $M^\tau$  and  $b^\tau$ .**

There remains one important issue of implementation, namely the computation of the local stiffness and mass matrices  $K^\tau$  and  $M^\tau$ , as well as the computation of the (local)  $b^\tau$ , cf. (5.49)–(5.51). This can be accomplished efficiently by letting  $\tau$  be the affine map of a fixed *reference triangle*  $\hat{\tau}$  and transforming the integrals in the formulas (5.49)–(5.51) to  $\hat{\tau}$ . To this end we will use the notation introduced in Section 5.2.



Let  $\hat{P}_1 \hat{P}_2 \hat{P}_3$  be the unit right triangle with vertices  $(0,0)$ ,  $(1,0)$ ,  $(0,1)$ . Let  $\tau \in \mathcal{T}_h$  be an arbitrary triangle in the triangulation with vertices  $P_j^\tau$ ,  $1 \leq j \leq 3$ , where

$P_j^\tau = (x_1^j, x_2^j)$ ,  $j = 1, 2, 3$ . Consider the affine map  $F_\tau$  that maps  $\hat{\tau}$  onto  $\tau$  and is defined by the requirements that  $P_j^\tau = F_\tau(\hat{P}_j)$ ,  $j = 1, 2, 3$ . Then  $F_\tau$  is of the form

$$x = F_\tau(\hat{x}) = B_\tau \hat{x} + c_\tau, \quad (5.62)$$

where  $B_\tau$  is the  $2 \times 2$  invertible matrix

$$B_\tau = \begin{pmatrix} x_1^2 - x_1^1 & x_1^3 - x_1^1 \\ x_2^2 - x_2^1 & x_2^3 - x_2^1 \end{pmatrix}$$

and  $c_\tau = (x_1^1, x_2^1)^T$ . Notice that  $|\det B_\tau| = 2 |\tau| := 2 \text{area}(\tau)$ .

Recall that functions on  $\hat{\tau}$  are transformed to the corresponding functions on  $\tau$  by  $\hat{v}(\hat{x}) = v(x)$ , whenever  $x = F_\tau(\hat{x})$ . I.e.  $\hat{v}(\hat{x}) = v(F_\tau(\hat{x}))$  and  $v(x) = \hat{v}(F_\tau^{-1}(x))$ . (Under our assumption on  $B_\tau$ , the map  $F_\tau$  is invertible, with  $\hat{x} = F_\tau^{-1}(x) = B_\tau^{-1}x - B_\tau^{-1}c_\tau$ ). (In our example (p. 114) the triangle  $\tau = 25$  has

$$B_\tau = \begin{pmatrix} \Delta x_1 & 0 \\ 0 & \Delta x_2 \end{pmatrix}, \quad c_\tau = \begin{pmatrix} 2 \Delta x_1 \\ 2 \Delta x_2 \end{pmatrix}, \quad \text{where } \Delta x_1 = \frac{\alpha}{5}, \Delta x_2 = \frac{\beta}{4}.$$

Let  $\hat{\varphi}_i(\hat{x})$ ,  $1 \leq i \leq 3$  be the local basis functions on the triangle  $\hat{\tau}$ , i.e. let  $\hat{\varphi}_i \in P_1$  be defined by the relations  $\hat{\varphi}_i(\hat{P}_j) = \delta_{ij}$ . Then, we easily see that

$$\hat{\varphi}_1(\hat{x}_1, \hat{x}_2) = 1 - \hat{x}_1 - \hat{x}_2,$$

$$\hat{\varphi}_2(\hat{x}_1, \hat{x}_2) = \hat{x}_1,$$

$$\hat{\varphi}_3(\hat{x}_1, \hat{x}_2) = \hat{x}_2.$$

It is easy to see that the corresponding local basis functions on  $\tau$ , i.e. the elements of  $P_1$  that satisfy  $\varphi_i^\tau(P_j^\tau) = \delta_{ij}$ ,  $1 \leq i, j \leq 3$ , are given by the relations

$$\varphi_i^\tau(x) = \hat{\varphi}_i(\hat{x}), \quad \text{whenever } x = F_\tau(\hat{x}).$$

Defining the  $1 \times 3$  matrix of reference basis functions  $\hat{\Phi}$  by

$$\hat{\Phi} = \hat{\Phi}(\hat{x}) = [\hat{\varphi}_1(\hat{x}), \hat{\varphi}_2(\hat{x}), \hat{\varphi}_3(\hat{x})]^T, \quad (5.63)$$

we see that  $\Phi^\tau(x) = \hat{\Phi}(\hat{x})$ , whenever  $x = F_\tau(\hat{x})$ , where  $\Phi^\tau(x)$  is the matrix of the local basis functions on  $\tau$  defined by (5.42).

We now compute the quantities  $b^\tau$ ,  $M^\tau$  and  $K^\tau$  in terms of integrals of functions defined on  $\hat{\tau}$ . We already seen that the area element transforms so that  $dx = |\det B_\tau| d\hat{x}$  i.e.  $dx = 2 |\tau| d\hat{x}$ . Then we have

$$b^\tau = \int_\tau f(x) (\Phi^\tau(x))^T dx = 2 |\tau| \int_{\hat{\tau}} \hat{f}(\hat{x}) (\hat{\Phi}(\hat{x}))^T d\hat{x}. \quad (5.64)$$

Hence, to compute  $b^\tau$  we must evaluate the integral on  $\hat{\tau}$  of the vector-valued function  $f(F_\tau(\hat{x})) (\hat{\Phi}(\hat{x}))^T$ , where  $\hat{\Phi}$  is defined in (5.63). Unless  $\hat{f}(\hat{x})$  is a very simple function, such integrals are evaluated numerically by an integration rule on  $\hat{\tau}$ . A simple but effective rule is the *barycenter rule*, which is exact for  $P_1$  polynomials and states that

$$\int_{\hat{\tau}} \hat{v}(\hat{x}) d\hat{x} \cong |\hat{\tau}| \hat{v}(\hat{M}),$$

where  $|\hat{\tau}| = \text{area}(\hat{\tau}) = 1/2$  and  $\hat{M} = (1/3, 1/3)$  is the barycenter of  $\hat{\tau}$ . Hence using

$$\int_{\hat{\tau}} \hat{v}(\hat{x}) d\hat{x} \cong \frac{1}{2} \hat{v}(1/3, 1/3)$$

in (5.64) we see that

$$b^\tau \cong |\tau| \hat{f}(1/3, 1/3) \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \text{ i.e. } b_i^\tau \cong \frac{|\tau|}{3} f(F_\tau(1/3, 1/3)), \quad 1 \leq i \leq 3.$$

Similarly, we may easily compute the elements of  $M^\tau$ . Since

$$M^\tau = \int_\tau a(x) (\Phi^\tau(x))^T \Phi^\tau(x) dx = 2 |\tau| \int_{\hat{\tau}} \hat{a}(\hat{x}) (\hat{\Phi}(\hat{x}))^T \hat{\Phi}(\hat{x}) d\hat{x},$$

we have, for  $1 \leq i, j \leq 3$

$$M_{ij}^\tau = 2 |\tau| \int_{\hat{\tau}} a(F_\tau(\hat{x})) \hat{\varphi}_i(\hat{x}) \hat{\varphi}_j(\hat{x}) d\hat{x}.$$

(If numerical integration with the barycenter rule is used, we have that

$$M_{ij}^\tau \cong \frac{|\tau|}{9} a(F_\tau(1/3, 1/3)), \quad 1 \leq i, j \leq 3).$$

The computation of

$$K^\tau = \int_\tau [D\Phi^\tau(x)]^T [D\Phi^\tau(x)] dx$$

requires transforming the matrix  $D\Phi^\tau$ . We have, for  $1 \leq i, j \leq 3$

$$(D\Phi^\tau(x))_{ij} = \frac{\partial \varphi_j^\tau(x)}{\partial x_i} = \frac{\partial}{\partial x_i}(\hat{\varphi}_j(\hat{x})) = \sum_{k=1}^3 \frac{\partial}{\partial \hat{x}_k}(\hat{\varphi}_j(\hat{x})) \frac{\partial \hat{x}_k}{\partial x_i}.$$

In analogy to (5.46) let  $\hat{D}\hat{\Phi}(\hat{x})$  be the  $2 \times 3$  matrix with elements

$$(\hat{D}\hat{\Phi}(\hat{x}))_{ij} = \frac{\partial \hat{\varphi}_j(\hat{x})}{\partial \hat{x}_i}. \quad (5.65)$$

In addition, note that from  $\hat{x} = B_\tau^{-1}x - B_\tau^{-1}c_\tau$ , we infer that

$$\frac{\partial \hat{x}_k}{\partial x_i} = (B_\tau^{-1})_{ki}.$$

Hence,

$$(D\Phi^\tau(x))_{ij} = \sum_{k=1}^3 (B_\tau^{-1})_{ki} (\hat{D}\hat{\Phi}(\hat{x}))_{kj}, \quad 1 \leq i, j \leq 3,$$

i.e.

$$D\Phi^\tau(x) = (B_\tau^{-1})^T \hat{D}\hat{\Phi}(\hat{x}). \quad (5.66)$$

We conclude that

$$K^\tau = 2 |\tau| \int_{\hat{\tau}} (\hat{D}\hat{\Phi}(\hat{x}))^T B_\tau^{-1} (B_\tau^{-1})^T \hat{D}\hat{\Phi}(\hat{x}) d\hat{x}.$$

The quantities inside the integral are independent of  $\hat{x}$ . Indeed,

$$\hat{D}\hat{\Phi}(\hat{x}) := J = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix},$$

and therefore

$$K^\tau = |\tau| J^T (B_\tau^T B_\tau)^{-1} J.$$

# Chapter 6

## The Galerkin Finite Element Method for the Heat Equation

### 6.1 Introduction. Elliptic projection

In this chapter we shall construct and analyze Galerkin finite element methods for the following model *parabolic* initial-boundary-value problem. Let  $\Omega$  be a bounded domain in  $\mathbb{R}^d$ ,  $d = 1, 2$ , or  $3$ . We seek a real function  $u = u(x, t)$ ,  $x \in \bar{\Omega}$ ,  $t \geq 0$ , such that

$$\begin{cases} u_t - \Delta u = f, & x \in \Omega, \quad t \geq 0, \\ u = 0, & x \in \partial\Omega, \quad t \geq 0, \\ u(x, 0) = u^0(x), & x \in \bar{\Omega}. \end{cases} \quad (6.1)$$

Here  $f = f(x, t)$  and  $u^0$  are given real functions on  $\Omega \times [0, \infty)$  and  $\bar{\Omega}$ , respectively. We shall assume that the initial-boundary-value problem (ibvp) (6.1) has a unique solution which is sufficiently smooth for the purposes of the analysis of its numerical approximation. For the theory of existence-uniqueness and regularity of problems like (6.1) see [2.2] - [2.5]. In this chapter we will just introduce some basic issues of approximating ibvp's like (6.1) with Galerkin methods. The reader is referred to [3.6] and [3.7] for many other related topics.

The spatial approximation of functions defined in  $\bar{\Omega}$  will be effected by a Galerkin finite element method. For this purpose we suppose that for  $h > 0$  we have a finite-dimensional subspace  $S_h$  of  $\overset{\circ}{H}^1 = \overset{\circ}{H}^1(\Omega)$  such that, for integer  $r \geq 2$  and  $h$  sufficiently



small, there holds

$$\inf_{\chi \in S_h} \{ \|v - \chi\| + h \|\nabla(v - \chi)\| \} \leq Ch^s \|v\|_s \quad \text{for } v \in H^s \cap \overset{\circ}{H}^1, \quad 2 \leq s \leq r, \quad (6.2)$$

where  $C$  is a positive constant independent of  $h$  and  $v$ . (In (6.2) we have used the notation  $\|\nabla v\| \equiv |v|_1 \equiv \left( \sum_{i=1}^d \left\| \frac{\partial v}{\partial x_i} \right\|^2 \right)^{1/2} = \left( \int_{\Omega} \nabla v \cdot \nabla v \, dx \right)^{1/2} \equiv (\nabla v, \nabla v)^{1/2}$ ). For example, (6.2) holds in  $\mathbb{R}^2$  when  $S_h = \left\{ \phi \in C(\overline{\Omega}), \phi|_{\tau} \in \mathbb{P}_1 \quad \forall \tau \in \mathcal{T}_h, \phi|_{\Omega - \Omega_h} = 0 \right\}$ , where  $\Omega_h$  is a polygonal domain included in  $\Omega$  and  $\mathcal{T}_h$  a regular triangulation of  $\Omega_h$  with triangles  $\tau$  with  $h = \max(\text{diam } \tau)$  whose vertices on  $\partial\Omega_h$  lie on  $\partial\Omega$ , cf. [3.5].

Given  $v \in H^1$ , we define  $R_h v$ , the *elliptic projection* of  $v$  in  $S_h$ , by the linear mapping  $R_h : H^1 \rightarrow S_h$ , such that

$$(\nabla R_h v, \nabla \chi) = (\nabla v, \nabla \chi), \quad \forall \chi \in S_h. \quad (6.3)$$

Given  $v \in H^1$  it is easy to see that  $R_h v$  exists uniquely in  $S_h$  and satisfies  $\|\nabla R_h v\| \leq \|\nabla v\|$ . The following error estimates follow from (6.2) and (6.3). (We have essentially seen their proof in Section 5.1 but we repeat it here for the convenience of the reader).

**Proposition 6.1.** *Suppose that  $v \in H^s \cap \overset{\circ}{H}^1$ , where  $2 \leq s \leq r$ . Then, there exists a constant  $C$  independent of  $v$  and  $h$  such that*

$$\|R_h v - v\| + h \|\nabla(R_h v - v)\| \leq Ch^s \|v\|_s, \quad 2 \leq s \leq r. \quad (6.4)$$

**Proof.** We have, by (6.3)

$$\begin{aligned} \|\nabla(R_h v - v)\|^2 &= (\nabla(R_h v - v), \nabla R_h v - \nabla v) \\ &= -(\nabla(R_h v - v), \nabla v) = (\nabla(R_h v - v), \nabla \chi - \nabla v) \end{aligned}$$

for any  $\chi \in S_h$ . Hence, by (6.2)

$$\|\nabla(R_h v - v)\|^2 \leq \|\nabla(R_h v - v)\| \|\nabla(v - \chi)\| \leq C \|\nabla(R_h v - v)\| h^{s-1} \|v\|_s,$$

from which the  $H^1$  estimate in (6.4) follows. For the  $L^2$  estimate we use *Nitsche's trick*. Given  $g \in L^2 = L^2(\Omega)$ , consider the bvp

$$\begin{cases} -\Delta \psi = g & \text{in } \Omega \\ \psi = 0 & \text{on } \partial\Omega. \end{cases}$$

This problem has a unique solution  $\psi \in H^2 \cap \overset{\circ}{H}^1$  such that  $\|\psi\|_2 \leq C \|\Delta\psi\| = C \|g\|$  (Elliptic regularity). For  $v \in H^s \cap \overset{\circ}{H}^1$ ,  $2 \leq s \leq r$ , we have by Gauss's theorem and (6.3)

$$(R_h v - v, g) = -(R_h v - v, \Delta\psi) = (\nabla(R_h v - v), \nabla\psi) = (\nabla(R_h v - v), \nabla(\psi - \chi))$$

for any  $\chi \in S_h$ . Hence, the  $H^1$  estimate in (6.4), (6.2), and elliptic regularity imply that

$$(R_h v - v, g) \leq Ch^{s-1} \|v\|_s h \|\psi\|_2 \leq Ch^s \|v\|_s \|g\|.$$

Taking  $g = R_h v - v$  gives that the  $L^2$  estimate in (6.4). □

## 6.2 Standard Galerkin semidiscretization

(In this and in the sections 6.2 and 6.3 we generally follow Thomée, [3.6, Ch.1])

Multiplying the pde in (6.1) by a function  $v \in \overset{\circ}{H}^1$ , and integrating over  $\Omega$  using Gauss's theorem, we see that

$$(u_t, v) + (\nabla u, \nabla v) = (f, v), \quad t \geq 0. \quad (6.5)$$

Motivated by (6.5), for each  $t \geq 0$  we approximate  $u(t) = u(\cdot, t)$  by a function  $u_h(t) = u_h(\cdot, t)$  in  $S_h$ , called the (*standard Galerkin*) *semidiscrete approximation* (or *spatial discretization*) of  $u$  in  $S_h$ , and defined by the equations

$$\begin{cases} (u_{ht}, \phi) + (\nabla u_h, \nabla \phi) = (f, \phi), & \forall \phi \in S_h, \quad t \geq 0, \\ u_h(0) = u_h^0, \end{cases} \quad (6.6)$$

where  $u_h^0$  is an approximation of  $u^0$  in  $S_h$  to be specified later.

The equations (6.6), that will be called the (*standard Galerkin*) *semidiscretization* of (6.1) in  $S_h$ , are equivalent to a linear system of ordinary differential equations (ode's). To see this, let  $\{\phi_j\}_{j=1}^{N_h}$  be a basis of  $S_h$ , where  $N_h = \dim S_h$ , and let

$$u_h(x, t) = \sum_{j=1}^{N_h} \alpha_j(t) \phi_j(x)$$

be the unknown semidiscrete approximation of  $u$ . Substituting this expression for  $u_h$  in (6.6) and taking  $\phi = \phi_k$ ,  $k = 1, \dots, N_h$ , we see that

$$\begin{cases} \sum_{j=1}^{N_h} \alpha_j'(t)(\phi_j, \phi_k) + \alpha_j(t)(\nabla\phi_j, \nabla\phi_k) = (f, \phi_k), & 1 \leq k \leq N_h, \quad t \geq 0, \\ \alpha_j(0) = \alpha_j^0, & 1 \leq j \leq N_h, \end{cases}$$

where  $u_h^0 = \sum_{j=1}^{N_h} \alpha_j^0 \phi_j$ . Hence the vector of unknowns  $\alpha = \alpha(t) = [\alpha_1, \dots, \alpha_{N_h}]^T$  satisfies the initial-value problem

$$\begin{cases} G\dot{\alpha} + S\alpha = F(t), & t \geq 0, \\ \alpha(0) = \alpha^0, \end{cases} \quad (6.7)$$

where  $G = (G_{ij})$  is the  $N_h \times N_h$  *mass (Gram)* matrix defined by  $G_{ij} = (\phi_j, \phi_i)$ ,  $S = (S_{ij})$  the  $N_h \times N_h$  *stiffness* matrix given by  $S_{ij} = (\nabla\phi_j, \nabla\phi_i)$ , and  $F_i = (f, \phi_i)$ . As we know,  $G$  and  $S$  are real, symmetric, positive definite matrices. In particular  $G$  is invertible and the ivp (6.7) has a unique solution  $\alpha(t)$  for all  $t \geq 0$ . We conclude that the Galerkin semidiscrete approximation  $u_h$  exists uniquely for all  $t \geq 0$ .

For each  $t > 0$  choose  $\phi = u_h$  in (6.6). Then

$$(u_{ht}, u_h) + \|\nabla u_h\|^2 = (f, u_h).$$

Since  $(u_{ht}, u_h) = \frac{1}{2} \int_{\Omega} \partial_t (u_h^2(\cdot, t)) \, dx = \frac{1}{2} \frac{d}{dt} \|u_h(t)\|^2$ , we have

$$\frac{1}{2} \frac{d}{dt} \|u_h\|^2 + \|\nabla u_h\|^2 = (f, u_h) \leq \|f\| \|u_h\|, \quad t \geq 0.$$

Recall the Poincarè-Friedrichs inequality, i.e. that

$$\|v\| \leq C_p \|\nabla v\|, \quad v \in \mathring{H}^1(\Omega), \quad (6.8)$$

valid for some  $C_p = C_p(\Omega)$ . Using (6.8) in the above gives

$$\frac{1}{2} \frac{d}{dt} \|u_h\|^2 + \|\nabla u_h\|^2 \leq C_p \|f\| \|\nabla u_h\| \leq \frac{C_p^2}{2} \|f\|^2 + \frac{1}{2} \|\nabla u_h\|^2,$$

from which

$$\frac{d}{dt} \|u_h\|^2 + \|\nabla u_h\|^2 \leq C_p^2 \|f\|^2, \quad t \geq 0.$$

We conclude that for any  $t > 0$  there holds that

$$\|u_h(t)\|^2 + \int_0^t \|\nabla u_h(s)\|^2 \, ds \leq \|u_h^0\|^2 + C_p^2 \int_0^t \|f(s)\|^2 \, ds. \quad (6.9)$$

In particular, for  $f = 0$ , we get  $\|u_h(t)\| \leq \|u_h^0\|$  for  $t \geq 0$ , i.e. that  $u_h$  is stable in  $L^2$ .

We now prove the main error estimate of this section.

**Theorem 6.1.** *Let  $u_h, u$  be the solutions of (6.6), (6.1), respectively. Then, there exists a constant  $C > 0$ , independent of  $h$  such that*

$$\|u_h(t) - u(t)\| \leq \|u_h^0 - u^0\| + Ch^r \left( \|u^0\|_r + \int_0^t \|u_t\|_r ds \right), \quad t \geq 0. \quad (6.10)$$

**Proof.** Following Wheeler (SIAM J. Numer. Anal., 10 (1973), 723-759), we write

$$u_h - u = \theta + \rho, \quad (6.11)$$

with  $\theta = u_h - R_h u$ , where  $R_h$  is the elliptic projection operator onto  $S_h$  defined by (6.3), and  $\rho = R_h u - u$ . Note that  $\theta \in S_h$  for  $t \geq 0$ , and in order to estimate  $\|u_h - u\|$  we should estimate  $\|\theta\|$  and  $\|\rho\|$ . For the latter, we have

$$\|\rho(t)\| = \|R_h u(t) - u(t)\| \leq Ch^r \|u(t)\|_r, \quad t \geq 0,$$

by (6.4). Since  $u(x, t) = u^0(x) + \int_0^t u_t(x, s) ds$ , assuming  $u^0 \in H^r \cap \overset{\circ}{H}^1$  and  $u_t \in H^r \cap \overset{\circ}{H}^1$  for  $t \geq 0$  with  $\int_0^t \|u_t\|_r ds < \infty$ , we see that  $u \in H^r \cap \overset{\circ}{H}^1$  for  $t \geq 0$  and  $\|u(t)\|_r \leq \|u^0\|_r + \int_0^t \|u_t\|_r ds$ .

Therefore

$$\|\rho(t)\| \leq Ch^r \left( \|u^0\|_r + \int_0^t \|u_t\|_r ds \right) \quad \text{for } t \geq 0. \quad (6.12)$$

In order to get an equation for  $\theta$  note that for  $t \geq 0$  and any  $\chi \in S_h$  we have

$$(\theta_t, \chi) + (\nabla \theta, \nabla \chi) = (u_{ht}, \chi) + (\nabla u_h, \nabla \chi) - ((R_h u)_t, \chi) - (\nabla R_h u, \nabla \chi).$$

Hence, using (6.6), (6.3), and the fact that  $((R_h u)_t, \chi) = (R_h u_t, \chi)$  for  $\chi \in S_h$  (this follows from (6.3) by differentiating both sides with respect to  $t$ ), we have

$$(\theta_t, \chi) + (\nabla \theta, \nabla \chi) = (f, \chi) - (R_h u_t, \chi) - (\nabla u, \nabla \chi).$$

Therefore, by (6.5) and the definition of  $\rho$ , we see that

$$(\theta_t, \chi) + (\nabla \theta, \nabla \chi) = -(\rho_t, \chi), \quad \forall \chi \in S_h, \quad t \geq 0. \quad (6.13)$$

Given  $t > 0$ , take  $\chi = \theta$  in the above to obtain

$$\frac{1}{2} \frac{d}{dt} \|\theta\|^2 + \|\nabla \theta\|^2 = -(\rho_t, \theta). \quad (6.14)$$

We would like to argue now that  $\frac{1}{2} \frac{d}{dt} \|\theta\|^2 = \|\theta\| \frac{d}{dt} \|\theta\|$  and conclude from (6.14) that  $\|\theta(t)\| \leq \|\theta(0)\| + \int_0^t \|\rho_t\| ds$ , but we don't know whether  $\frac{d}{dt} \|\theta(t)\|$  exists if  $\theta = 0$  for some

$t$ . Therefore we argue as follows: For all  $\varepsilon > 0$  we have from (6.14)  $\frac{1}{2} \frac{d}{dt} \|\theta\|^2 = \frac{1}{2} \frac{d}{dt} (\|\theta\|^2 + \varepsilon^2) \leq \|\rho_t\| \|\theta\|$ . Hence  $\sqrt{\|\theta\|^2 + \varepsilon^2} \frac{d}{dt} \sqrt{\|\theta\|^2 + \varepsilon^2} \leq \|\rho_t\| \|\theta\| \leq \|\rho_t\| \sqrt{\|\theta\|^2 + \varepsilon^2}$ . Therefore,  $\frac{d}{dt} \sqrt{\|\theta\|^2 + \varepsilon^2} \leq \|\rho_t\|$  for all  $t > 0$ , from which  $\sqrt{\|\theta(t)\|^2 + \varepsilon^2} \leq \int_0^t \|\rho_t\| ds + \sqrt{\|\theta(0)\|^2 + \varepsilon^2}$ . Letting  $\varepsilon \rightarrow 0$  we obtain the desired inequality

$$\|\theta(t)\| \leq \int_0^t \|\rho_t\| ds + \|\theta(0)\|, \quad t \geq 0. \quad (6.15)$$

Since  $\rho_t = R_h u_t - u_t$ , we have  $\int_0^t \|\rho_t\| ds \leq Ch^r \int_0^t \|u_t\|_r ds$  from (6.4). On the other hand,

$$\|\theta(0)\| \leq \|u_h^0 - u^0\| + \|R_h u^0 - u^0\| \leq \|u_h^0 - u^0\| + Ch^r \|u^0\|_r.$$

Therefore, (6.15), (6.11) and (6.12) yield the desired estimate (6.10).  $\square$

### Remarks

**a.** Theorem 6.1, and subsequent error estimates, depend on assumptions of sufficient *regularity* of the solution  $u$  of the continuous problem. Such assumptions will not normally be explicitly made in the statements of theorems but will appear in the conclusions or in the course of proofs of the error estimates. For example, in the case of the estimate at hand the proof requires that  $u^0 \in H^r \cap \mathring{H}^1$  and  $u_t(\cdot, s) \in H^r \cap \mathring{H}^1$  for  $0 \leq s \leq t$ . These assumptions guarantee in particular that the bound of  $\|u_h(t) - u(t)\|$  in (6.10) is of the form  $\|u_h^0 - u^0\| + O(h^r)$  where the  $O(h^r)$  term is of *optimal order of convergence* in  $L^2$  for  $S_h$ , as evidenced by the approximation property (6.2).

**b.** The initial value  $u_h^0$  may be chosen in various ways so that  $\|u_h^0 - u^0\| = O(h^r)$ . For example, we could choose it as  $u_h^0 = R_h u^0$  or  $u_h^0 = Pu^0$  (the  $L^2$  projection of  $u^0$  onto  $S_h$ ) or equal to an interpolant of  $u^0$  in  $S_h$ . For example, if one of the first two choices is made, (6.2) and (6.4) yield that  $\|u_h^0 - u^0\| \leq Ch^r \|u^0\|_r$ , provided  $u^0 \in H^r \cap \mathring{H}^1$ , and the overall optimal-order accuracy  $O(h^r)$  is preserved in the right-hand side of (6.10).

**Exercise 1.** In the proof of Theorem 6.1 take in (6.11)  $u_h - u = \theta + \rho$ , where, for example,  $\theta = u_h - Pu$ , where  $P$  is the  $L^2$ -projection operator onto  $S_h$ , or  $\theta = u_h - I_h u$ , where  $I_h u$  is an interpolant of  $u$  in  $S_h$ , satisfying  $\|v - I_h v\| + h \|\nabla(v - I_h v)\| \leq Ch^r \|v\|_r$  for  $v \in H^r \cap \mathring{H}^1$ . For these choices show that the best  $L^2$ -error estimate that one could obtain is of  $O(h^{r-1})$ , i.e. of suboptimal order. Try to understand from these

considerations why Wheeler's choice  $\theta = u_h - R_h u$  is crucial.

We now prove an optimal-order  $H^1$  estimate for  $u_h$ .

**Proposition 6.2.** *Under the hypotheses of Theorem 6.1, there holds*

$$\|\nabla(u_h(t) - u(t))\| \leq \|\nabla(u_h^0 - u^0)\| + Ch^{r-1} \left[ \|u^0\|_r + \|u(t)\|_r + \left( \int_0^t \|u_t\|_{r-1}^2 ds \right)^{1/2} \right], t \geq 0. \quad (6.16)$$

**Proof:** As before, we write  $\nabla(u_h - u) = \nabla\theta + \nabla\rho$ , where  $\theta = u_h - R_h u$ ,  $\rho = R_h u - u$ . Note that  $\|\nabla\rho\| \leq Ch^{r-1}\|u\|_r$ . From (6.13) with  $\chi = \theta_t$  it follows that

$$\|\theta_t\|^2 + \frac{1}{2} \frac{d}{dt} \|\nabla\theta\|^2 = -(\rho_t, \theta_t) \leq \frac{1}{2} \|\rho_t\|^2 + \frac{1}{2} \|\theta_t\|^2.$$

Therefore

$$\frac{d}{dt} \|\nabla\theta\|^2 \leq \|\rho_t\|^2, \quad t \geq 0$$

form which

$$\|\nabla\theta(t)\| \leq \|\nabla\theta(0)\| + \left( \int_0^t \|\rho_t\|^2 ds \right)^{1/2}, \quad t \geq 0.$$

We conclude that

$$\begin{aligned} \|\nabla(u_h - u)\| &\leq \|\nabla\theta(0)\| + \left( \int_0^t \|\rho_t\|^2 ds \right)^{1/2} + \|\nabla\rho\| \\ &\leq \|\nabla(u_h^0 - u^0)\| + \|\nabla(R_h u^0 - u^0)\| + \|\nabla\rho\| + Ch^{r-1} \left( \int_0^t \|u_t\|_{r-1}^2 ds \right)^{1/2}, \end{aligned}$$

from which (6.16) follows.  $\square$

**Exercise 2.** Consider instead of (6.1) the initial-boundary value problem with *Neumann* boundary conditions:

$$\begin{cases} u_t - \Delta u = f, & x \in \Omega, \quad t \geq 0, \\ \frac{\partial u}{\partial n} = 0, & x \in \partial\Omega, \quad t \geq 0, \\ u(x, 0) = u^0(x), & x \in \bar{\Omega}. \end{cases}$$

Construct the standard Galerkin semidiscretization for this problem in a finite-dimensional subspace  $S_h$  of  $H^1$  and prove an analog of Theorem 6.1. (Define now an elliptic projection  $R_h : H^1 \rightarrow S_h$  by the equations  $a(R_h v, \chi) = a(v, \chi)$ ,  $\forall \chi \in S_h$  for  $v \in H^1$ , where

$$a(v, w) = (\nabla v, \nabla w) + (v, w) \text{ for } v, w \in H^1. )$$

**Exercise 3.** Generalize the results of this section to the case of the initial-boundary value problem with variable coefficients

$$\begin{cases} u_t - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left( a_{ij}(x) \frac{\partial u}{\partial x_j} \right) + a_0(x)u = f(x, t), & x \in \Omega, \quad t \geq 0, \\ u = 0 \text{ on } \Omega, & t \geq 0, \\ u(x, 0) = u_0(x), & x \in \bar{\Omega}, \end{cases}$$

where the functions  $a_{ij}$  satisfy  $a_{ij}(x) = a_{ji}(x)$ ,  $x \in \bar{\Omega}$  and  $\sum_{i,j=1}^d a_{ij}\xi_i\xi_j \geq c_0 \sum_{i=1}^d \xi_i^2$ ,  $\forall x \in \bar{\Omega}$ ,  $\forall \xi = (\xi_1, \dots, \xi_d) \in \mathbb{R}^d$ , for some positive constant  $c_0$  independent of  $x$  and  $\xi$ , i.e. when the matrix-valued function  $a_{ij}$  is symmetric and uniformly positive definite for  $x \in \bar{\Omega}$ , and where  $a_0(x) \geq 0$ ,  $x \in \bar{\Omega}$ . Assume that the coefficients  $a_{ij}$ ,  $a_0$  are smooth enough on  $\bar{\Omega}$ . The weak formulation of this ibvp is to find  $u \in \mathring{H}^1$  for  $t \geq 0$  such that

$$\begin{cases} (u_t, v) + a(u, v) = (f(t), v), & \forall v \in \mathring{H}^1, \quad t \geq 0, \\ u(0) = u_0, \end{cases}$$

where  $a(u, v) := \sum_{i,j=1}^d \int_{\Omega} a_{ij}uv dx + \int_{\Omega} a_0uv dx$ . (Establish first that there exist positive constants  $C_1, C_2$  such that  $|a(v, w)| \leq C_1\|v\|_1\|w\|_1 \quad \forall v, w \in \mathring{H}^1$ , and  $a(v, v) \geq C_2\|v\|_1^2 \quad \forall v \in \mathring{H}^1$ , and introduce now the elliptic projection of  $v \in \mathring{H}^1$  onto  $S_h$  by  $a(R_h v, \chi) = a(v, \chi) \quad \forall \chi \in S_h$ . Use the Lax-Milgram theorem and Nitsche's trick to prove analogous properties of  $R_h$  to those of Section 6.1, assuming elliptic regularity, i.e. that  $\|u\|_2 \leq C\|f\|$  holds for the associated elliptic bvp

$$\begin{cases} - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left( a_{ij}(x) \frac{\partial u}{\partial x_j} \right) + a_0(x) u = f(x), & x \in \tilde{\Omega}, \\ u = 0 \text{ on } \partial\Omega. \end{cases}$$

)

### 6.3 Full discretization with the implicit Euler and the Crank-Nicolson method

To solve the o.d.e. system (6.6) (or (6.7)) we need to discretize it in  $t$  and obtain a *fully discrete method*. This may be done by various time-stepping techniques. In this section we shall examine two of them, the implicit Euler and the Crank-Nicolson methods, that are 'unconditionally stable' in a sense that we shall make precise.

Let  $\Delta t = k$  be the length of the (uniform) timestep and  $t^n = nk$ ,  $n = 0, 1, 2, \dots$ . The *implicit Euler* full discretization of (6.16) is defined as follows. We seek for  $n = 0, 1, 2, \dots$  approximations  $U^n \in S_h$  of  $u_h(t^n)$  that satisfy

$$\begin{cases} \left( \frac{U^n - U^{n-1}}{k}, \chi \right) + (\nabla U^n, \nabla \chi) = (f^n, \chi), & \forall \chi \in S_h, \quad n \geq 1, \\ U^0 = u_h^0, \end{cases} \quad (6.17)$$

where  $f^n = f(\cdot, t^n)$ .

Finding  $U^n$  for  $n \geq 1$ , given  $U^{n-1}$ , requires solving a linear system for the coefficients of  $U^n$  with respect to the basis  $\{\phi_j\}_{j=1}^{N_h}$  of  $S_h$ . Let  $U^n = \sum_{i=1}^{N_h} \alpha_i^n \phi_i$ , where  $\alpha^n = (\alpha_1^n, \dots, \alpha_{N_h}^n) \in \mathbb{R}^{N_h}$ . Then, putting  $\chi = \phi_j$ ,  $1 \leq j \leq N_h$ , in (6.17) gives

$$(G + kS)\alpha^n = G\alpha^{n-1} + kF^n, \quad n \geq 1, \quad (6.18)$$

where  $F_i^n = (f^n, \phi_i)$ ,  $1 \leq i \leq N_h$ . Thus, computing  $\alpha^n$  requires forming the right-hand side of (6.18) and solving a  $N_h \times N_h$  linear system with the matrix  $G + kS$ , which is symmetric positive definite, and has the sparsity structure of  $G$  and  $S$ . If a direct method, like Cholesky's method, is used to solve this linear system, the  $LL^T$  analysis of  $G + kS$  may be done only once and  $\alpha^n$  computed for each  $n$  using two backsolves with  $L$  and  $L^T$ . In more than one spatial dimensions such linear systems are usually solved by a preconditioned conjugate-gradient type method.

Putting  $\chi = U^n$  in (6.17) gives

$$\|U^n\|^2 + k \|\nabla U^n\|^2 = (U^{n-1}, U^n) + k(f^n, U^n) \leq (\|U^{n-1}\| + k\|f^n\|)\|U^n\|.$$

Hence,  $\|U^n\| \leq \|U^{n-1}\| + k\|f^n\|$ , for  $n = 1, 2, \dots$ . This implies that

$$\|U^n\| \leq \|U^0\| + k \sum_{j=1}^n \|f^j\|, \quad n = 1, 2, \dots$$



If  $f = 0$  we see that  $\|U^n\| \leq \|U^0\|$ , i.e. that the implicit Euler scheme is  $L^2$ -stable. In fact for each  $n$  we have  $\|U^n\| \leq \|U^{n-1}\|$ , i.e. the  $L^2$  norm of  $U^n$  is non-increasing. In particular, if  $f^n = 0$ ,  $U^{n-1} = 0$ , we see that  $U^n = 0$ , i.e. that the homogenous linear system of equations of the form (6.18) has only the trivial solution, implying that (6.18) has a unique solution; this is an alternative to the matrix argument used previously to show the same result. Note that the  $L^2$  stability of the scheme was proved without any assumption on the timestep  $k$ , i.e. that the scheme (6.17) is *unconditionally stable* in  $L^2$ .

As expected, the implicit Euler method is first-order accurate in the time variable as the following estimate suggests.

**Theorem 6.2.** *Let  $U^n$ ,  $u(t) = u(\cdot, t)$ , be the solutions of (6.17), (6.1), respectively. Then, there exists a positive constant  $C$ , independent of  $h$ ,  $k$ , and  $n$ , such that for  $n \geq 0$*

$$\|U^n - u(t^n)\| \leq \|u_h^0 - u^0\| + Ch^r \left[ \|u^0\|_r + \int_0^{t^n} \|u_t\|_r ds \right] + k \int_0^{t^n} \|u_{tt}\| ds. \quad (6.19)$$

**Proof:** As in the proof of Theorem 6.1 we write  $U^n - u(t^n) = \theta^n + \rho^n$ , where  $\theta^n = U^n - R_h u(t^n)$ ,  $\rho^n = R_h u(t^n) - u(t^n)$ .

Using (6.4) we see that

$$\|\rho^n\| \leq Ch^r \|u(t^n)\|_r \leq Ch^r \left[ \|u^0\|_r + \int_0^{t^n} \|u_t\|_r ds \right], \quad (6.20)$$

and it remains to estimate  $\|\theta^n\|$ . Let  $\bar{\partial}U^n := \frac{1}{k}(U^n - U^{n-1})$ . By (6.17), (6.5) and for  $\chi \in S_h$  we have

$$\begin{aligned} (\bar{\partial}\theta^n, \chi) + (\nabla\theta^n, \nabla\chi) &= (\bar{\partial}U^n, \chi) + (\nabla U^n, \nabla\chi) - (\bar{\partial}R_h u(t^n), \chi) - (\nabla R_h u(t^n), \nabla\chi) \\ &= (f(t^n), \chi) - (\bar{\partial}R_h u(t^n), \chi) - (\nabla u(t^n), \nabla\chi) \\ &= (u_t(t^n) - R_h \bar{\partial}u(t^n), \chi), \end{aligned}$$

i.e.

$$(\bar{\partial}\theta^n, \chi) + (\nabla\theta^n, \nabla\chi) = -(\omega^n, \chi), \quad \forall \chi \in S_h, \quad n \geq 1, \quad (6.21)$$

where

$$\begin{aligned} \omega^n &= R_h \bar{\partial}u(t^n) - u_t(t^n) \\ &= (R_h - I)\bar{\partial}u(t^n) + (\bar{\partial}u(t^n) - u_t(t^n)) =: \omega_1^n + \omega_2^n. \end{aligned} \quad (6.22)$$

Putting  $\chi = \theta^n$  in (6.21) gives

$$(\bar{\partial}\theta^n, \theta^n) + \|\nabla\theta^n\|^2 \leq \|\omega^n\| \|\theta^n\|, \text{ i.e.}$$

$$\|\theta^n\|^2 - (\theta^{n-1}, \theta^n) \leq h \|\omega^n\| \|\theta^n\|, \text{ from which}$$

$$\|\theta^n\| \leq \|\theta^{n-1}\| + k \|\omega^n\|, \quad n \geq 1.$$

Therefore, in view of (6.22), summation with respect to  $n$  gives

$$\|\theta^n\| \leq \|\theta^0\| + k \sum_{j=1}^n \|\omega_1^j\| + k \sum_{j=1}^n \|\omega_2^j\|, \quad n \geq 1. \quad (6.23)$$

Now

$$\begin{aligned} \omega_1^j &= (R_h - I)\bar{\partial}u(t^j) = (R_h - I)\frac{1}{k}(u(t^j) - u(t^{j-1})) \\ &= (R_h - I)\frac{1}{k} \int_{t^{j-1}}^{t^j} u_t(s) ds = \frac{1}{2} \int_{t^{j-1}}^{t^j} (R_h - I)u_t(x) ds. \end{aligned}$$

Therefore by (6.4)

$$\begin{aligned} \|\omega_j^1\| &\leq C \frac{h^r}{k} \int_{t^{j-1}}^{t^j} \|u_t\|_r ds, \text{ giving} \\ k \sum_{j=1}^n \|\omega_1^j\| &\leq C h^r \int_0^{t^n} \|u_t\|_r ds. \end{aligned} \quad (6.24)$$

For the last term in (6.23) we note

$$\omega_2^j = \bar{\partial}u(t^j) - u_t(t^j) = \frac{1}{k}(u(t^j) - u(t^{j-1})) - u_t(t^j).$$

For a real function  $v = v(t)$  recall Taylor's theorem with integral remainder:

$$v(t) = v(a) + (t-a)v'(a) + \dots + \frac{(t-a)^p}{p!}v^{(p)}(a) + \frac{1}{p!} \int_a^t (t-s)^p v^{(p+1)}(s) ds.$$

Hence

$$u(t^{j-1}) = u(t^j) - k u_t(t^j) + \int_{t^j}^{t^{j-1}} (t^{j-1} - s) u_{tt}(s) ds.$$

We conclude

$$\omega_2^j = -\frac{1}{k} \int_{t^j}^{t^{j-1}} (t^{j-1} - s) u_{tt}(s) ds,$$

from which

$$\|\omega_2^j\| \leq \frac{1}{k} \int_{t^{j-1}}^{t^j} (s - t^{j-1}) \|u_{tt}(s)\| ds \leq \int_{t^{j-1}}^{t^j} \|u_{tt}(s)\| ds,$$

yielding

$$k \sum_{j=1}^n \|\omega_2^j\| \leq k \int_0^{t^n} \|u_{tt}(s)\| ds. \quad (6.25)$$

Since  $\|\theta^0\| = \|U^0 - R_h u^0\| \leq \|u_h^0 - u^0\| + C h^r \|u^0\|_r$ , we conclude from (6.23), (6.24), (6.25) that

$$\|\theta^n\| \leq \|u_h^0 - u^0\| + C h^r \left[ \|u^0\|_r + \int_0^{t^n} \|u_t\|_r ds \right] + k \int_0^{t^n} \|u_{tt}(s)\| ds, \quad n \geq 0,$$

which, in view of the inequality  $\|U^n - u(t^n)\| \leq \|\theta^n\| + \|\rho^n\|$  and (6.20) gives (6.19).

**Remark.** The inequality (6.23) is essentially a *stability* inequality for the ‘error’ equation (6.21), whereas the estimates (6.24) and (6.25) are bounds on the spatial and temporal ‘truncation’ errors in the right-hand side of (6.23) and express the *consistency* of the fully discrete scheme in that they tend to zero as  $h \rightarrow 0$ ,  $k \rightarrow 0$  under the implied regularity assumptions on  $u$ . In fact, they imply that the spatial accuracy of the scheme in  $L^2$  is of  $O(h^r)$  and the temporal accuracy is of  $O(k)$ . Thus the proof of Theorem 6.2 is an illustration of the general principle that ‘stability + consistency  $\Rightarrow$  convergence’. Such a statement has to be verified in any given particular case and depends on the choice of norms and the regularity of solutions.

We turn now to the *Crank - Nicolson* scheme for discretizing (6.6) in  $t$  with second-order accuracy and retaining unconditional stability. We seek for  $n \geq 0$  approximations  $U^n \in S_h$  of  $u_h(t^n)$  satisfying

$$\begin{cases} \left( \frac{U^n - U^{n-1}}{k}, \chi \right) + \frac{1}{2} (\nabla(U^n + U^{n-1}), \nabla \chi) = (f^{n-1/2}, \chi), \quad \forall \chi \in S_h, \quad n \leq 1, \\ U^0 = u_h^0, \end{cases} \quad (6.26)$$

where  $f^{n-1/2} = f(\cdot, t^n - \frac{k}{2})$ . Using our previous notation, we see that for  $n \geq 1$  the matrix-vector representation of (6.26) is

$$\left( G + \frac{k}{2} S \right) \alpha^n = \left( G - \frac{k}{2} S \right) \alpha^{n-1} + k F^{n-1/2}, \quad n \geq 1, \quad (6.27)$$

where again  $\alpha^n$  is the vector of coefficients of  $U^n$  with respect to the basis of  $S_h$ . The matrix  $G + \frac{k}{2} S$  is again sparse, symmetric and positive definite, and similar remarks hold for computing  $\alpha^n$  as in the case of the implicit Euler scheme. Putting  $\chi = U^n + U^{n-1}$

in (6.26) gives for  $n \geq 1$

$$\begin{aligned} \|U^n\|^2 - \|U^{n-1}\|^2 + \frac{k}{2} \|\nabla(U^n + U^{n-1})\|^2 &= k(f^{n-1/2}, U^{n+1} + U^n) \\ &\leq k \|f^{n-1/2}\| (\|U^n\| + \|U^{n-1}\|). \end{aligned}$$

Hence  $\|U^n\| \leq \|U^{n-1}\| + k\|f^{n-1/2}\|$ , for  $n \geq 1$ , implying that

$$\|U^n\| \leq \|U^0\| + k \sum_{j=1}^n \|f^{j-1/2}\|, \quad n = 1, 2, \dots$$

From these relations, if  $f = 0$ , we see that the Crank-Nicolson method is  $L^2$ -stable and that  $\|U^n\|$  is non-increasing, unconditionally. In particular, if  $f^{n-1/2} = 0$ ,  $U^{n-1} = 0$ , it follows that  $U^n = 0$ , which means that the homogenous linear system of the form (6.27) has only the trivial solution, implying that (6.27) has a unique solution given  $\alpha^{n-1}$  and  $F^{n-1/2}$ . We proceed with an error estimate for the scheme.

**Theorem 6.3.** *Let  $U^n$ ,  $u(t) = u(\cdot, t)$ , be the solutions of (6.26), (6.1), respectively. Then, there exists a positive constant  $C$ , independent of  $h$ ,  $k$  and  $n$ , such that for  $n \geq 0$*

$$\|U^n - u(t^n)\| \leq \|u_h^0 - u^0\| + C h^r \left[ \|u^0\|_r + \int_0^{t^n} \|u_t\|_r ds \right] + C k^2 \int_0^{t^n} (\|u_{ttt}\| + \|\Delta u_{tt}\|) ds. \quad (6.28)$$

**Proof:** We write again  $U^n - u(t^n) = \theta^n + \rho^n$ , where  $\theta^n = U^n - R_h u(t^n)$ ,  $\rho^n = R_h u(t^n) - u(t^n)$  and note that

$$\|\rho^n\| \leq C h^r \left[ \|u^0\|_r + \int_0^{t^n} \|u_t\|_r ds \right]. \quad (6.29)$$

From (6.26), (6.5) and for  $\chi \in S_h$  we have

$$\begin{aligned} (\bar{\partial}\theta^n, \chi) + \frac{1}{2} (\nabla(\theta^n + \theta^{n-1}), \nabla\chi) \\ &= (f(t^{n-1/2}), \chi) - (R_h \bar{\partial}u(t^n), \chi) - \frac{1}{2} (\nabla(u(t^n) + u(t^{n-1})), \nabla\chi) \\ &= (f(t^{n-1/2}), \chi) - (u_t(t^{n-1/2}), \chi) + (\Delta u(t^{n-1/2}), \chi) \\ &\quad + (u_t(t^{n-1/2}) - R_h \bar{\partial}u(t^n), \chi) - (\Delta u(t^{n-1/2}) - \frac{1}{2} \Delta(u(t^n) + u(t^{n-1})), \chi), \end{aligned}$$

where we used Gauss's theorem (integration by parts) in the last term. Hence

$$(\bar{\partial}\theta^n, \chi) + \frac{1}{2} (\nabla(\theta^n + \theta^{n-1}), \nabla\chi) = -(\omega^n, \chi), \quad \forall \chi \in S_h, \quad n \geq 1, \quad (6.30)$$

where

$$\begin{aligned}\omega^n &= \omega_1^n + \omega_2^n + \omega_3^n := (R_h - I)\bar{\partial}u(t^n) + (\bar{\partial}u(t^n) - u_t(t^{n-1/2})) \\ &\quad + \Delta(u(t^{n-1/2}) - \frac{1}{2}(u(t^n) + u(t^{n-1}))).\end{aligned}\quad (6.31)$$

If we put  $\chi = \theta^n + \theta^{n-1}$  in (6.30), we obtain, as in the stability proof of the scheme,

$$\|\theta^n\| \leq \|\theta^0\| + k \sum_{j=1}^n \|\omega_1^j\| + k \sum_{j=1}^n \|\omega_2^j\| + k \sum_{j=1}^n \|\omega_3^j\|, \quad n \geq 1. \quad (6.32)$$

As in the case of the implicit Euler scheme (cf. (6.24)), we have

$$k \sum_{j=1}^n \|\omega_1^j\| \leq C h^r \int_0^{t^n} \|u_t\|_r ds. \quad (6.33)$$

To estimate  $\omega_2^j$  we note that for  $j \geq 1$

$$\omega_2^j = \bar{\partial}u(t^j) - u_t(t^{j-1/2}) = \frac{1}{k}(u(t^j) - u(t^{j-1})) - u_t(t^{j-1/2}).$$

Using Taylor's theorem gives

$$\begin{aligned}u(t^j) &= u(t^{j-1/2}) + \frac{k}{2}u_t(t^{j-1/2}) + \frac{k^2}{2!4}u_{tt}(t^{j-1/2}) + \frac{1}{2!} \int_{t^{j-1/2}}^{t^j} (t^j - s)^2 u_{ttt}(s) ds, \\ u(t^{j-1}) &= u(t^{j-1/2}) - \frac{k}{2}u_t(t^{j-1/2}) + \frac{k^2}{2!4}u_{tt}(t^{j-1/2}) + \frac{1}{2!} \int_{t^{j-1/2}}^{t^{j-1}} (t^{j-1} - s)^2 u_{ttt}(s) ds,\end{aligned}$$

so that

$$\omega_2^j = \frac{1}{2k} \left[ \int_{t^{j-1/2}}^{t^j} (t^j - s)^2 u_{ttt}(s) ds + \int_{t^{j-1}}^{t^{j-1/2}} (s - t^{j-1})^2 u_{ttt}(s) ds \right].$$

Therefore, for  $1 \leq j$

$$\|\omega_2^j\| \leq \frac{1}{2k} \left( \frac{k^2}{4} \int_{t^{j-1/2}}^{t^j} \|u_{ttt}\| ds + \frac{k^2}{4} \int_{t^{j-1}}^{t^{j-1/2}} \|u_{ttt}\| ds \right) = \frac{k}{8} \int_{t^{j-1}}^{t^j} \|u_{ttt}\| ds,$$

i.e.

$$k \sum_{j=1}^n \|\omega_2^j\| \leq \frac{k^2}{8} \int_0^{t^n} \|u_{ttt}\| ds. \quad (6.34)$$

For  $\omega_3^j$ ,  $j \leq 1$ , we have

$$\omega_3^j = \Delta \left( u(t^{j-1/2}) - \frac{1}{2}(u(t^j) + u(t^{j-1})) \right).$$

By Taylor's theorem

$$u(t^j) = u(t^{j-1/2}) + \frac{k}{2}u_t(t^{j-1/2}) + \int_{t^{j-1/2}}^{t^j} (t^j - s)u_{tt}(s)ds,$$

$$u(t^{j-1}) = u(t^{j-1/2}) - \frac{k}{2}u_t(t^{j-1/2}) + \int_{t^{j-1/2}}^{t^{j-1}} (t^{j-1} - s)u_{tt}(s)ds,$$

so that

$$\omega_3^j = -\frac{1}{2} \int_{t^{j-1/2}}^{t^j} (t^j - s)\Delta u_{tt}ds - \frac{1}{2} \int_{t^{j-1}}^{t^{j-1/2}} (s - t^{j-1})\Delta u_{tt}ds.$$

Therefore

$$k \sum_{j=1}^n \|\omega_3^j\| \leq \frac{k^2}{4} \int_0^{t^n} \|\Delta u_{tt}\|ds. \quad (6.35)$$

The desired estimate (6.28) follows now from the inequalities (6.29), (6.32)-(6.35), and the fact that  $\|\theta^0\| = \|U^0 - R_h u_0\| \leq \|u_h^0 - u^0\| + \|u^0 - R_h u^0\| \leq \|u_h^0 - u^0\| + C h^r \|u^0\|_r$ .  $\square$

**Exercise 1.** Let  $\frac{1}{2} \leq \alpha \leq 1$  and consider the following family of fully discrete schemes

$$\left\{ \begin{array}{l} \left( \frac{U^n - U^{n-1}}{k}, \chi \right) + (\nabla(\alpha U^n + (1 - \alpha)U^{n-1}), \nabla \chi) \\ \qquad \qquad \qquad = (\alpha f(t^n) + (1 - \alpha)f(t^{n-1}), \chi) \quad \forall \chi \in S_h, n \geq 1, \\ U^0 = u_h^0. \end{array} \right.$$

Show that the schemes are  $L^2$ -stable and prove error estimates of the form

$$\|U^n - u(t^n)\| \leq \|u_h^0 - u^0\| + O(k^p + h^r), \text{ where } p = 1 \text{ if } 1/2 < \alpha \leq 1 \text{ and } p = 2 \text{ if } \alpha = 1/2.$$

**Exercise 2.** Consider the implicit Euler method with *variable* step  $k_n$ , where  $k_n = t^n - t^{n-1}$ ,  $n \geq 1$ :

$$\left\{ \begin{array}{l} \left( \frac{U^n - U^{n-1}}{k_n}, \chi \right) + (\nabla U^n, \nabla \chi) = (f(t^n), \chi), \quad \forall \chi \in S_h, n \geq 1, \\ U^0 = u_h^0. \end{array} \right.$$

Show that the scheme is  $L^2$ -stable and prove an error estimate of the form (6.19) with  $k = \max_n k_n$ .

## 6.4 The explicit Euler method. Inverse inequalities and stiffness

The *explicit Euler* full discretization of (6.6) is the following scheme: We seek for  $n = 0, 1, 2, \dots$  approximations  $U^n \in S_h$  of  $u_h(t^n)$  that satisfy

$$\begin{cases} \left( \frac{U^n - U^{n-1}}{k}, \chi \right) + (\nabla U^{n-1}, \nabla \chi) = (f^{n-1}, \chi), & \forall \chi \in S_h, \quad n \geq 1, \\ U^0 = u_h^0. \end{cases} \quad (6.36)$$

Finding  $U^n$  for  $n \geq 1$ , given  $U^{n-1}$ , requires solving the linear system

$$G \alpha^n = (G - k S) \alpha^{n-1} + k F^{n-1}, \quad n \geq 1, \quad (6.37)$$

where  $G$ ,  $S$ ,  $F^n$  have been previously defined and  $\alpha^n$  is as usual the vector of coefficients of  $U^n$  with respect to the basis of  $S_h$ . (Note that although the time-stepping method is explicit as a scheme for solving initial-value problems for ode's, we still have to solve linear systems with the mass matrix  $G$  at each time step.) Obviously the linear system (6.37) has a unique solution  $\alpha^n$  given  $\alpha^{n-1}$  and  $F^{n-1}$ .

In order to study the stability of the scheme put  $\chi = U^n$  in (6.36). Then, for  $n \geq 1$  we have

$$\|U^n\|^2 - (U^{n-1}, U^n) + k (\nabla U^{n-1}, \nabla U^n) = k (f^{n-1}, U^n).$$

Use now the identities

$$\begin{aligned} -(U^{n-1}, U^n) &= \frac{1}{2} (\|U^n - U^{n-1}\|^2 - \|U^n\|^2 - \|U^{n-1}\|^2), \\ (\nabla U^{n-1}, \nabla U^n) &= \frac{1}{4} \|\nabla(U^n + U^{n-1})\|^2 - \frac{1}{4} \|\nabla(U^n - U^{n-1})\|^2 \end{aligned}$$

to obtain

$$\begin{aligned} \|U^n - U^{n-1}\|^2 + \|U^n\|^2 - \|U^{n-1}\|^2 + \frac{k}{2} \|\nabla(U^n + U^{n-1})\|^2 - \frac{k}{2} \|\nabla(U^n - U^{n-1})\|^2 = \\ 2k (f^{n-1}, U^n), \quad n \geq 1. \end{aligned} \quad (6.38)$$

In the left-hand side of this identity the troublesome term is  $-\frac{k}{2} \|\nabla(U^n - U^{n-1})\|^2$  which is non-positive. To resolve this problem we write (6.38) in the form

$$\begin{aligned} \|U^n - U^{n-1}\|^2 + \|U^n\|^2 - \|U^{n-1}\|^2 + \frac{k}{2} \|\nabla(U^n + U^{n-1})\|^2 \\ = \frac{k}{2} \|\nabla(U^n - U^{n-1})\|^2 + 2k (f^{n-1}, U^n) \end{aligned}$$

and use the *inverse inequality* (valid for quasiuniform partitions of  $\Omega$ )

$$\|\nabla\chi\| \leq \frac{C_*}{h}\|\chi\|, \quad \chi \in S_h, \quad (6.39)$$

where  $C_*$  is a constant independent of  $h$  and  $\chi$ , in the first term of the right-hand side to get

$$\|U^n - U^{n-1}\|^2 + \|U^n\|^2 - \|U^{n-1}\|^2 \leq \frac{k}{2} \frac{C_*^2}{h^2} \|U^n - U^{n-1}\|^2 + 2k \|f^{n-1}\| \|U^n\|,$$

i.e.

$$\left(1 - \frac{k}{h^2} \frac{C_*^2}{2}\right) \|U^n - U^{n-1}\|^2 + \|U^n\|^2 - \|U^{n-1}\|^2 \leq 2k \|f^{n-1}\| \|U^n\|.$$

Therefore, if

$$\frac{k}{h^2} \leq \frac{2}{C_*^2}, \quad (6.40)$$

the above inequality gives

$$\|U^n\|^2 - \|U^{n-1}\|^2 \leq 2k \|f^{n-1}\| \|U^n\| \leq 2k \|f^{n-1}\| (\|U^n\| + \|U^{n-1}\|),$$

from which

$$\|U^n\| \leq \|U^{n-1}\| + 2k \|f^{n-1}\|, \quad n \geq 1,$$

and finally

$$\|U^n\| \leq \|U^0\| + 2k \sum_{j=0}^{n-1} \|f^j\|,$$

which is the required  $L^2$ -stability inequality, analogous to those that were derived for the implicit Euler and the Crank-Nicolson schemes. However, whereas such inequalities in the case of the previous schemes were valid unconditionally, the explicit Euler scheme needs a *stability condition* of the form (6.40). This condition is very restrictive in that it requires taking  $k = O(h^2)$ , i.e. very small time steps. It can be shown that such a condition is also necessary for stability. (A simple numerical experiment with piecewise linear functions in 1D gives a clear indication!).

We postpone for the time being the proof of the inverse inequality (6.39) in order to prove an  $L^2$  error estimate for the fully discrete scheme (6.36):

**Theorem 6.4.** *Suppose that (6.39) and (6.40) hold and let  $U^n, u$  be the solutions of (6.36), (6.1) respectively. Then, there exists a positive constant  $C$ , independent of  $h, k$  and  $u$ , such that for  $n \geq 0$*

$$\|U^n - u(t^n)\| \leq \|u_h^0 - u^0\| + Ch^r \left[ \|u^0\|_r + \int_0^{t^n} \|u_t\|_r ds \right] + 2k \int_0^{t^n} \|u_{tt}\| ds. \quad (6.41)$$



**Proof:** Writing again  $U^n - u(t^n) = \theta^n + \rho^n$ ,  $\theta^n = U^n - R_h u(t^n)$ ,  $\rho^n = R_h u(t^n) - u(t^n)$ , and noting that (6.29) holds for  $\rho^n$ , we obtain for  $n \geq 1$  and any  $\chi \in S_h$

$$(\bar{\partial}\theta^n, \chi) + (\nabla\theta^{n-1}, \nabla\chi) = -(\omega^n, \chi),$$

where

$$\omega^n = \omega_1^n + \omega_2^n := (R_h - I)\bar{\partial}u(t^n) + (\bar{\partial}u(t^n) - u_t(t^{n-1})).$$

Putting  $\chi = \theta^n$  we see, as in the derivation of (6.38), that

$$\begin{aligned} \|\theta^n\|^2 - \|\theta^{n-1}\|^2 + \|\theta^n - \theta^{n-1}\|^2 + \frac{k}{2}\|\nabla(\theta^n + \theta^{n-1})\|^2 &= \frac{k}{2}\|\nabla(\theta^n - \theta^{n-1})\|^2 - 2k(\omega^n, \theta^n) \\ &\stackrel{(6.39)}{\leq} \frac{k}{2}\frac{C_*^2}{h^2}\|\theta^n - \theta^{n-1}\|^2 - 2k(\omega^n, \theta^n). \end{aligned}$$

Therefore, using (6.40), as in the stability proof

$$\|\theta^n\| \leq \|\theta^0\| + 2k \sum_{j=1}^n \|\omega^j\| \leq \|\theta^0\| + 2k \sum_{j=1}^n \|\omega_1^j\| + 2k \sum_{j=1}^n \|\omega_2^j\|. \quad (6.42)$$

For the  $\omega_1^j$  term we have as in (6.24)

$$k \sum_{j=1}^n \|\omega_1^j\| \leq C h^r \int_0^{t^n} \|u_t\|_r ds.$$

Since by Taylor's theorem

$$k \omega_2^j = u(t^j) - u(t^{j-1}) - k u_t(t^{j-1}) = \int_{t^{j-1}}^{t^j} (t^j - s) u_{tt}(s) ds, \quad j \geq 1,$$

we see that

$$k \sum_{j=1}^n \|\omega_2^j\| \leq k \int_0^{t^n} \|u_{tt}\| ds,$$

and (6.41) follows from (6.42).  $\square$

We proceed now with verifying the inverse inequality (6.39). This is straightforward to do in 1D. Consider an interval  $(\lambda, \mu)$  with  $\mu - \lambda < 1$  and let  $\mathbb{P}_k$  be the polynomials of degree  $\leq k$ . Then for some constant  $C = C(k)$  independent of  $\lambda$  and  $\mu$  it holds that

$$\|\phi\|_{H^1(\lambda, \mu)} \leq \frac{C(k)}{\mu - \lambda} \|\phi\|_{L^2(\lambda, \mu)}, \quad \forall \phi \in \mathbb{P}_k. \quad (6.43)$$

To see this, observe that there exists a constant  $C = C(k)$  such that

$$\|\phi\|_{H^1(0,1)} \leq C \|\phi\|_{L^2(0,1)}, \quad \forall \phi \in \mathbb{P}_k, \quad (6.44)$$

as a consequence of the fact that  $\mathbb{P}_k$  is a finite-dimensional vector space and the norms  $\|\cdot\|_{H^1(0,1)}$  and  $\|\cdot\|_{L^2(0,1)}$  are equivalent on  $\mathbb{P}_k(0,1)$ . To derive (6.43) from (6.44) requires just a change of scale. Write the inequality (6.44) for  $\phi \in \mathbb{P}_k$  as

$$\int_0^1 (\phi^2(x) + (\phi'(x))^2) dx \leq C^2 \int_0^1 \phi^2(x) dx,$$

and make in the integrals the change of variable  $x \mapsto y$ ,  $y = (\mu - \lambda)x + \lambda$ , that maps  $[0, 1]$  onto  $[\lambda, \mu]$ . Then the above inequality becomes

$$\frac{1}{\mu - \lambda} \int_\lambda^\mu \left[ \phi^2(y) + (\mu - \lambda)^2 \left( \frac{d\phi}{dy}(y) \right)^2 \right] dy \leq \frac{C^2}{\mu - \lambda} \int_\lambda^\mu \phi^2(y) dy,$$

giving

$$(\mu - \lambda)^2 \int_\lambda^\mu \left[ \phi^2(y) + \left( \frac{d\phi}{dy}(y) \right)^2 \right] dy \leq C^2 \int_\lambda^\mu \phi^2(y) dy,$$

since we assumed  $\mu - \lambda < 1$ . Hence (6.43) follows.

Let now  $(a, b)$  be any fixed finite interval and let  $a = x_0 < x_1 < \dots < x_{J+1} = b$  be an arbitrary partition of  $[a, b]$ . Letting  $h_i = x_{i+1} - x_i$ ,  $0 \leq i \leq J$ , we obtain from (6.43) that

$$\|\phi\|_{H^1(x_i, x_{i+1})} \leq \frac{C(k)}{h_i} \|\phi\|_{L^2(x_i, x_{i+1})}, \quad \forall \phi \in \mathbb{P}_k(x_i, x_{i+1}). \quad (6.45)$$

Let now  $S_h = \left\{ \phi \in C[a, b] : \phi|_{[x_i, x_{i+1}]} \in \mathbb{P}_k \right\}$ . Since  $S_h \subset H^1$ , using (6.45) we get for any  $\phi \in S_h$

$$\|\phi\|_{H^1(a,b)}^2 = \sum_{i=0}^J \|\phi\|_{H^1(x_i, x_{i+1})}^2 \leq C^2(k) \sum_{i=0}^J \frac{1}{h_i^2} \|\phi\|_{L^2(x_i, x_{i+1})}^2. \quad (6.46)$$

We now assume that the partition  $\{x_i\}$  of  $[a, b]$  is *quasiuniform*, i.e. that there is a constant  $\nu$  independent of the partition (in the sense that as the partition is refined  $\nu$  does not change) such that

$$\frac{h}{h_i} \leq \nu, \quad \forall i, \quad (6.47)$$

where  $h = \max_i h_i$ . In view of (6.47), (6.46) gives

$$\|\phi\|_{H^1(a,b)} \leq \frac{C_*}{h} \|\phi\|_{L^2(a,b)}, \quad \forall \phi \in S_h, \quad (6.48)$$

with  $C_* = C(k)\nu$ , from which (6.39) follows in 1D.

In order to prove (6.39) in 2D, we assume that  $\Omega$  is a polygonal domain and let  $\mathcal{T}_h = \{\tau\}$  be a regular (cf. (5.16)) triangulation of  $\Omega$ . We recall from section 5.2 that

any triangle  $\tau$  of the triangulation is affinely equivalent to a fixed reference triangle  $\widehat{\tau}$ . As in the case of 1D, there exists a constant  $C_1 = C_1(\widehat{\tau}, k)$  such that

$$\|\widehat{\phi}\|_{1,\widehat{\tau}} \leq C_1 \|\widehat{\phi}\|_{0,\widehat{\tau}}, \quad \forall \widehat{\phi} \in \mathbb{P}_k(\widehat{\tau}), \quad (6.49)$$

as a consequence again of the fact that  $\mathbb{P}_k(\widehat{\tau})$  is a finite-dimensional space and that  $\|\cdot\|_{1,\widehat{\tau}} = \|\cdot\|_{H^1(\widehat{\tau})}$  and  $\|\cdot\|_{0,\widehat{\tau}} = \|\cdot\|_{L^2(\widehat{\tau})}$  are norms on  $\mathbb{P}_k(\widehat{\tau})$ . Suppose now that  $S_h = \{\phi \in C(\overline{\Omega}) : \phi|_{\tau} \in \mathbb{P}_k(\tau)\}$  and let  $F_{\tau}$  be the affine map that maps  $\widehat{\tau}$  one-one onto  $\tau$ . Following the notation of section 5.2 we write  $F_{\tau}(\widehat{x}) = B_{\tau}\widehat{x} + b_{\tau}$ , for  $\widehat{x} = (\widehat{x}_1, \widehat{x}_2) \in \widehat{\tau}$ , where  $B_{\tau}$  is a  $2 \times 2$  invertible matrix and  $b_{\tau}$  a 2-vector. If  $\phi \in \mathbb{P}_k(\tau)$  we define  $\widehat{\phi} \in \mathbb{P}_k(\widehat{\tau})$  by  $\phi(x) = \widehat{\phi}(\widehat{x})$ , where  $x = (x_1, x_2)$  and  $x = F_{\tau}(\widehat{x})$  as usual. Then for  $\phi \in \mathbb{P}_k(\tau)$ , using the transformation norm inequalities (5.20), (5.21) and (6.49) we get

$$\begin{aligned} |\phi|_{1,\tau} &\leq \widehat{C}(k) |B_{\tau}^{-1}| |\det B_{\tau}|^{1/2} |\widehat{\phi}|_{1,\widehat{\tau}} \\ &\leq C_1 \widehat{C}(k) |B_{\tau}^{-1}| |\det B_{\tau}|^{1/2} \|\widehat{\phi}\|_{0,\widehat{\tau}} \\ &\leq C(k, \widehat{\tau}) |B_{\tau}^{-1}| |\det B_{\tau}|^{1/2} \cdot |\det B_{\tau}|^{-1/2} \|\phi\|_{0,\tau}. \end{aligned}$$

Therefore, using the regularity of the triangulation and (5.24) we see that there exists a constant  $C$  independent of  $\tau$  such that

$$|\phi|_{1,\tau} \leq \frac{C}{h_{\tau}} \|\phi\|_{0,\tau}, \quad \forall \phi \in \mathbb{P}_k(\tau),$$

where  $h_{\tau} = \text{diam}(\tau)$ . Since  $S_h \subset H^1(\Omega)$  we see that

$$|\phi|_{1,\Omega}^2 = \sum_{\tau \in \mathcal{T}_h} |\phi|_{1,\tau}^2 \leq C^2 \sum_{\tau \in \mathcal{T}_h} \frac{1}{h_{\tau}^2} \|\phi\|_{0,\tau}^2, \quad \forall \phi \in S_h.$$

If the triangulation is *quasiuniform* in the sense that for some constant  $\nu$  independent of the partition

$$\frac{h}{h_{\tau}} \leq \nu, \quad \forall \tau \in \mathcal{T}_h, \quad (6.50)$$

where  $h = \max_{\tau} h_{\tau}$ , we see that

$$|\phi|_{1,\Omega} \leq \frac{C\nu}{h} \|\phi\|_{0,\Omega}, \quad \forall \phi \in S_h, \quad (6.51)$$

from which (6.39) (and also  $\|\phi\|_1 \leq Ch^{-1}\|\phi\|$ ) follows.

We mention that similar scaling arguments yield, for quasiuniform partitions, the more general inverse inequalities

$$\|\chi\|_{\alpha} \leq Ch^{\beta-\alpha} \|\chi\|_{\beta}, \quad \forall \chi \in S_h,$$

provided  $\alpha, \beta$  are nonnegative integers so that  $\beta < \alpha$  and  $S_h \subset H^\alpha(\Omega)$ , and

$$\|\chi\|_\infty \leq Ch^{-d/2}\|\chi\|, \quad \forall \chi \in S_h$$

for  $\Omega \in \mathbb{R}^d$ . Of interest is also the nonstandard inverse “almost Sobolev” inequality

$$\|\chi\|_\infty \leq C|\ln h|^{1/2}\|\chi\|_1, \quad \forall \chi \in S_h,$$

valid for  $\Omega \subset \mathbb{R}^2$ , cf. [3.6, p.68].

We close this section with a few remarks about stiff systems of ode’s and the interpretation of the stability condition (6.40) as a restriction on the time step related to the size of the eigenvalues of the matrix  $G^{-1}S$ . Let  $A$  be a symmetric and positive definite real  $m \times m$  matrix and consider the ode ivp for  $y = y(t) \in \mathbb{R}^m$

$$\begin{aligned} \dot{y} + Ay &= 0, \quad t \geq 0, \\ y(0) &= y_0. \end{aligned} \tag{6.52}$$

Let  $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$  be the eigenvalues of  $A$ . Then from the theory of numerical solution of ode’s, we know that if a method for the numerical solution of ivp’s has interval of absolute stability  $[-\alpha, 0]$ , where  $\alpha > 0$  ( $\alpha = +\infty$  if the method is  $A_0$ -stable), it will give stable approximations, when applied to (6.52), provided the time step  $\Delta t$  is chosen so that  $(-\lambda_i)\Delta t \in [-\alpha, 0]$  for all  $i$ , i.e. so that  $\Delta t \leq \frac{\alpha}{\lambda_m}$ . We recall that for the explicit Euler method  $\alpha = 2$ , while  $\alpha = +\infty$  for the implicit Euler and the trapezoidal method. Hence, the latter two methods are suitable for stiff systems, i.e. systems for which  $\lambda_1 = O(1)$  and  $\lambda_m \gg 1$ .

Consider now the ode system (6.7) corresponding to the Galerkin semidiscretization of the ibvp (6.1). For  $f = 0$  we write the system as

$$\begin{aligned} Gy + Sy &= 0, \quad t \geq 0, \\ y(0) &= y_0, \end{aligned} \tag{6.53}$$

where  $y \in \mathbb{R}^m$  with  $m = N_h = \dim S_h$ . This system is of the form (6.52) with  $A = G^{-1}S$  but  $G^{-1}S$  is not symmetric. To transform the system into the form (6.52) with a symmetric positive definite matrix, consider the matrix  $G^{1/2}$ . Since  $G$  is symmetric and positive definite,  $G^{1/2}$  is defined e.g. using the spectral representation of  $G$  and is

symmetric and positive definite. Multiplying by  $G^{-1/2}$  both sides of the ode system in (6.53) we have

$$G^{1/2}\dot{y} + G^{-1/2}S G^{-1/2} G^{1/2}y = 0,$$

or

$$\dot{z} + Az = 0, \quad t \geq 0, \quad (6.53')$$

where  $z(t) = G^{1/2}y(t)$ , and  $A = G^{-1/2}S G^{-1/2}$  is easily seen to be symmetric and positive definite. Let  $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$  be the eigenvalues of  $A$ . Then they satisfy

$$G^{-1/2}S G^{-1/2}x = \lambda_i x, \quad (6.54)$$

where  $x \in \mathbb{R}^m$  is a corresponding to  $\lambda_i$  eigenvector. It follows that  $S G^{-1/2}x = \lambda_i G^{1/2}x$ . Therefore, with  $G^{-1/2}x = w$ , i.e.  $Gw = G^{1/2}x$ , we see that the eigenvalue problem (6.54) is equivalent to the generalized eigenvalue problem

$$Sw = \lambda_i Gw \quad (6.55)$$

and so the  $\lambda_i$  are eigenvalues of  $G^{-1}S$ . It follows from (6.55) that

$$\lambda_i = \frac{w^T S w}{w^T G w}.$$

Let now  $\phi = \sum_{i=1}^m w_i \phi_i \in S_h$ , where  $\{\phi_i\}$  is the chosen basis of  $S_h$ . Then  $\|\phi\|^2 = w^T G w$ ,  $\|\nabla \phi\|^2 = w^T S w$ , and  $\lambda_i = \frac{\|\nabla \phi\|^2}{\|\phi\|^2}$ . Therefore, if the inverse inequality (6.39) holds, we have that

$$\lambda_m = \lambda_{\max}(G^{-1}S) \leq \frac{C_*^2}{h^2}. \quad (6.56)$$

Hence, a sufficient condition for the stability of the explicit Euler scheme for the ivp (6.53') or, equivalently, for the ivp (6.53), is, since  $\left(-\frac{C_*^2}{h^2}\right)k \leq (-\lambda_m)k$ ,  $k = \Delta t$ , that  $-\frac{C_*^2}{h^2}k \geq -2$ , i.e.  $\frac{k}{h^2} \leq \frac{2}{C_*^2}$ , which is precisely the restriction (6.40) found by the energy method. For the implicit Euler or the Crank-Nicolson (i.e. the trapezoidal) scheme for which  $\alpha = +\infty$ , there is no restriction.

# Chapter 7

## The Galerkin Finite Element Method for the Wave Equation

### 7.1 Introduction

In this chapter we shall consider Galerkin finite element methods for the following model *second-order hyperbolic* initial-boundary-value problem. Let  $\Omega$  be a bounded domain in  $\mathbb{R}^d$ ,  $d = 1, 2$ , or  $3$ . We seek a real function  $u = u(x, t)$ ,  $x \in \bar{\Omega}$ ,  $t \geq 0$ , such that

$$\begin{cases} u_{tt} - \Delta u = f, & x \in \Omega, t \geq 0, \\ u = 0, & x \in \partial\Omega, t \geq 0, \\ u(x, 0) = u^0(x), & u_t(x, 0) = u_t^0(x), x \in \bar{\Omega}. \end{cases} \quad (7.1)$$

Here  $f = f(x, t)$  and  $u^0, u_t^0$  are given real functions on  $\Omega \times [0, \infty)$ , and  $\bar{\Omega}$  respectively. We shall assume that the ibvp (7.1) has a unique solution, which is sufficiently smooth for the purposes of the analysis of its numerical approximation. For the theory of existence-uniqueness and regularity of (7.1) see [2.2]-[2.5]. We just make two remarks here for the homogeneous problem, i.e. when  $f = 0$  in (7.1).

i. Multiplying both sides of the pde in (7.1) by  $u_t$ , integrating over  $\Omega$  using Gauss's theorem and the boundary and initial conditions, we easily get the *energy* conservation identity

$$\|u_t\|^2 + \|\nabla u\|^2 = \|u_t^0\|^2 + \|\nabla u^0\|^2, \quad t \geq 0. \quad (7.2)$$

ii. The regularity theory for (7.1) requires smoothness and compatibility conditions at  $\partial\Omega$  of the initial data  $u^0, u_t^0$ , and establishes corresponding smoothness and compatibility properties in the same spaces of the solutions  $u$  and  $u_t$  for  $t > 0$ . Thus, in the case of (7.1) we do not have any smoothing effect on the solutions for  $t > 0$  as in the case of the heat equation, cf. e.g [2.4].

## 7.2 Standard Galerkin semidiscretization

Multiplying the pde in (7.1) by  $v \in \overset{\circ}{H}^1$  and integrating over  $\Omega$  using Gauss's theorem we obtain

$$(u_{tt}, v) + (\nabla u, \nabla v) = (f, v), \quad t \geq 0. \quad (7.3)$$

Let  $S_h$  be a finite-dimensional subspace of  $\overset{\circ}{H}^1$  satisfying (6.2). Motivated by (7.3) we define the (*standard Galerkin*) *semidiscrete approximation*  $u_h$  of  $u$  as a function  $u_h(t)$  in  $S_h, t \geq 0$ , such that

$$\begin{cases} (u_{htt}, \varphi) + (\nabla u_h, \nabla \varphi) = (f, \varphi), \quad \forall \varphi \in S_h, \quad t \geq 0, \\ u_h(0) = u_h^0, \\ u_{ht}(0) = u_{t,h}^0, \end{cases} \quad (7.4)$$

where  $u_h^0, u_{t,h}^0$  are given elements of  $S_h$  approximating  $u^0, u_t^0$ , respectively. If  $\{\varphi_j\}_{j=1}^{N_h}$  is a basis of  $S_h$  and we let

$$u_h(x, t) = \sum_{j=1}^{N_h} \alpha_j(t) \varphi_j(x),$$

we see that

$$\begin{cases} G\ddot{\alpha} + S\alpha = F(t), \quad t \geq 0, \\ \alpha(0) = \beta, \\ \dot{\alpha}(0) = \gamma, \end{cases} \quad (7.5)$$

where  $G_{ij} = (\varphi_j, \varphi_i)$ ,  $S_{ij} = (\nabla \varphi_j, \nabla \varphi_i)$ ,  $1 \leq i, j \leq N_h$ ,  $F_i = (f, \varphi_i)$ ,  $1 \leq i \leq N_h$ ,  $\alpha = \alpha(t) = [\alpha_1, \dots, \alpha_{N_h}]^T$ , and  $\beta = (\beta_i)$ ,  $\gamma = (\gamma_i)$  are the coefficients of  $u_h^0, u_{t,h}^0$ , with

respect to the basis  $\{\varphi_i\}$ . The ivp (7.5) may also be written as

$$\begin{cases} \begin{pmatrix} I & 0 \\ 0 & G \end{pmatrix} \frac{d}{dt} \begin{pmatrix} \alpha \\ \dot{\alpha} \end{pmatrix} = \begin{pmatrix} 0 & I \\ -S & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \dot{\alpha} \end{pmatrix} + \begin{pmatrix} 0 \\ F \end{pmatrix}, & t \geq 0, \\ (\alpha(0), \dot{\alpha}(0))^T = (\beta, \gamma)^T, \end{cases}$$

i.e. as an ivp for a first-order system of ode' s; it is then evident that (7.5) has a unique solution for  $t \geq 0$ . If we put  $\varphi = u_h$  in (7.4) with  $f = 0$  it is straightforward to prove that

$$\|u_{ht}(t)\|^2 + \|\nabla u_h(t)\|^2 = \|u_{t,h}^0\|^2 + \|\nabla u_h^0\|^2, \quad t \geq 0, \quad (7.6)$$

which is the discrete analog of (7.2) and expresses the stability (conservation) of  $(u_h, u_{ht})$  in  $\overset{\circ}{H}^1 \times L^2$ .

We recall now the definition of the *elliptic projection* of a function  $v \in \overset{\circ}{H}^1$  as the element  $R_h v \in S_h$  such that  $(\nabla R_h v, \nabla \chi) = (\nabla v, \nabla \chi)$  for all  $\chi \in S_h$ . Following Dupont, SIAM J. Numer. Anal., 10(1973), 880-889, we have

**Theorem 7.1.** *Let  $u, u_h$  be the solutions of the ivp' s (7.1) and (7.4) respectively. Then, given  $T > 0$ , there exists a positive constant  $C = C(T)$  such that*

$$\begin{aligned} \|u_h(t) - u(t)\| \leq C \left\{ \|\nabla(u_h^0 - R_h u^0)\| + \|u_{t,h}^0 - R_h u_t^0\| \right. \\ \left. + h^r \left[ \|u(t)\|_r + \left( \int_0^t \|u_{tt}\|_r^2 ds \right)^{1/2} \right] \right\}, \quad 0 \leq t \leq T. \end{aligned} \quad (7.7)$$

*Proof.* We write as usual  $u_h - u = \theta + \varrho$ , where  $\theta = u_h - R_h u$ ,  $\varrho = R_h u - u$ . We have as before

$$\|\varrho(t)\| \leq Ch^r \|u(t)\|_r, \quad t \geq 0. \quad (7.8)$$

Now, for  $t \geq 0$  and any  $\chi \in S_h$  we have using (7.4) and (7.3)

$$\begin{aligned} (\theta_{tt}, \chi) + (\nabla \theta, \nabla \chi) &= (u_{htt}, \chi) + (\nabla u_h, \nabla \chi) - (R_h u_{tt}, \chi) - (\nabla R_h u, \nabla \chi) \\ &= (f, \chi) - (R_h u_{tt}, \chi) - (\nabla u, \nabla \chi) \\ &= (u_{tt} - R_h u_{tt}, \chi) = -(\varrho_{tt}, \chi). \end{aligned}$$

Given  $t > 0$  take  $\chi = \theta_t$  in the above to obtain

$$\frac{1}{2} \frac{d}{dt} (\|\theta_t\|^2 + \|\nabla \theta\|^2) = -(\varrho_{tt}, \theta_t) \leq \frac{1}{2} \|\varrho_{tt}\|^2 + \frac{1}{2} \|\theta_t\|^2.$$



Hence

$$\frac{d}{dt}(\|\theta_t\|^2 + \|\nabla\theta\|^2) \leq \|\varrho_{tt}\|^2 + (\|\theta_t\|^2 + \|\nabla\theta\|^2), \quad t \geq 0. \quad (7.9)$$

We recall now *Gronwall's lemma*: If  $\sigma'(t) \leq \alpha(t) + \sigma(t)$ ,  $t \geq 0$ , then  $\sigma(t) \leq e^t \sigma(0) + \int_0^t e^{t-s} \alpha(s) ds$ . To see this, note that for  $t \geq 0$

$$e^{-t} \sigma'(t) - e^{-t} \sigma(t) \leq e^{-t} \alpha(t).$$

Therefore

$$\begin{aligned} \frac{d}{dt}(e^{-t} \sigma(t)) &\leq e^{-t} \alpha(t), \quad \text{i.e.} \\ e^{-t} \sigma(t) - \sigma(0) &\leq \int_0^t e^{-s} \alpha(s) ds, \end{aligned}$$

from which the conclusion follows. Using this result in (7.9) for  $\sigma = \|\theta_t\|^2 + \|\nabla\theta\|^2$  we have for  $0 \leq t \leq T$

$$\|\theta_t\|^2 + \|\nabla\theta\|^2 \leq C(T) \left( \|\theta_t(0)\|^2 + \|\nabla\theta(0)\|^2 + \int_0^t \|\varrho_{tt}\|^2 ds \right),$$

where  $C(T) = e^T$ . By the definition of  $\theta$  and  $\varrho$  it follows that for  $0 \leq t \leq T$

$$\|\theta_t\|^2 + \|\nabla\theta\|^2 \leq C(T) \left[ \|u_{t,h}^0 - R_h u_t^0\|^2 + \|\nabla(u_h^0 - R_h u^0)\|^2 + Ch^{2r} \int_0^t \|u_{tt}\|^2 ds \right].$$

Using now the inequalities  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \leq (\sum_{i=1}^n \alpha_i^2)^{1/2} \leq \sum_{i=1}^n \alpha_i$ , valid for  $\alpha_i \geq 0$ , we conclude that for some constant  $C = C(T)$

$$\begin{aligned} \|\theta_t\| + \|\nabla\theta\| &\leq C[\|u_{t,h}^0 - R_h u_t^0\| + \|\nabla(u_h^0 - R_h u^0)\| \\ &\quad + h^r \left( \int_0^t \|u_{tt}\|^2 ds \right)^{1/2}], \quad 0 \leq t \leq T. \end{aligned} \quad (7.10)$$

We note now that  $\theta(t) = \int_0^t \theta_t(s) ds + \theta(0)$ , from which for  $t \geq 0$

$$\|\theta(t)\| \leq \|\theta(0)\| + \int_0^t \|\theta_t\| ds \leq \|\theta(0)\| + t \max_{0 \leq s \leq t} \|\theta_t(s)\|.$$

Therefore, using (7.10) we have for  $0 \leq t \leq T$

$$\begin{aligned} \|u_h(t) - u(t)\| &\leq \|\varrho(t)\| + \|\theta(t)\| \\ &\leq Ch^r \|u(t)\|_r + \|\theta(0)\| + t \max_{0 \leq s \leq t} \|\theta_t(s)\| \\ &\leq C(T) \{ \|u_{t,h}^0 - R_h u_t^0\| + \|\nabla(u_h^0 - R_h u^0)\| \\ &\quad + h^r [\|u(t)\|_r + \left( \int_0^t \|u_{tt}\|_r^2 ds \right)^{1/2}] \}, \end{aligned}$$

which is the desired estimate (7.7). (Note that we used *Poincaré's inequality*  $\|\theta(0)\| \leq C\|\nabla\theta(0)\|$  in the last inequality.)

□

### Remarks

**a.** The estimate (7.7) implies that if we choose  $u_h^0 = R_h u^0$ ,  $u_{t,h}^0 = R_h u_t^0$ , then we have the optimal-order  $L^2$ -estimate  $\|u_h(t) - u(t)\| = O(h^r)$ . We may also choose  $u_{t,h}^0$  as any optimal-order (in  $L^2$ ) approximation to  $u_t^0$  in  $S_h$ , e.g. as  $Pu_t$ , where  $P$  is the  $L^2$ -projection onto  $S_h$ . However we cannot do the same for  $u_h^0$ , which must be close to  $R_h u^0$  in  $H^1$  to  $O(h^r)$  to guarantee the optimal-order bound in (7.7).

**b.** From (7.10) it also follows that for  $0 \leq t \leq T$

$$\begin{aligned} \|u_{ht} - u_t\| &\leq \|\theta_t\| + \|\varrho_t\| \leq C \left\{ \|u_{t,h}^0 - R_h u_t^0\| \right. \\ &\quad \left. + \|\nabla(u_h^0 - R_h u^0)\| + h^r \left[ \|u_t(t)\|_r + \left( \int_0^t \|u_{tt}\|_r^2 ds \right)^{1/2} \right] \right\}, \end{aligned}$$

and

$$\begin{aligned} \|\nabla(u_h - u)\| &\leq \|\nabla\theta\| + \|\nabla\varrho\| \leq C \left\{ \|u_{t,h}^0 - R_h u_t^0\| \right. \\ &\quad \left. + \|\nabla(u_h^0 - R_h u^0)\| + h^{r-1} \left[ \|u\|_{r-1} + \left( \int_0^t \|u_{tt}\|_r^2 ds \right)^{1/2} \right] \right\}. \end{aligned}$$

The latter inequality implies that initial conditions e.g. of the type  $u_h^0 = Pu^0$ ,  $u_{t,h}^0 = Pu_t^0$  will give an optimal-order estimate for  $\|\nabla(u_h - u)\|$ .

The following result, due to G. Baker, SIAM J. Numer. Anal. 13(1976), 564-576, relies on a ‘duality’ argument in time and shows that one may after all get an  $L^2$  estimate of optimal order starting with any optimal-order  $L^2$  approximation of  $u^0$  and  $u_t^0$ . We state it in a form that also improves Theorem 7.1 with respect to the required regularity of the solution and the dependence of  $C$  on  $T$ .

**Theorem 7.2.** *Let  $u$ ,  $u_h$  be the solutions of the ivp's (7.1), (7.4), respectively. Then there exists a constant  $C$ , independent of  $t$  and  $h$ , such that*

$$\begin{aligned} \|u_h(t) - u(t)\| &\leq C \left\{ \|u_h^0 - Pu^0\| + t \|u_{t,h}^0 - Pu_t^0\| \right. \\ &\quad \left. + h^r [\|u^0\|_r + \int_0^t \|u_t\|_r ds] \right\}. \end{aligned} \tag{7.11}$$

*Proof.* We put  $e := u_h - u = \theta + \varrho$ , where  $\theta = u_h - R_h u$ ,  $\varrho = R_h u - u$ . As in the proof of Theorem 7.1 we have  $(\theta_{tt}, \chi) + (\nabla\theta, \nabla\chi) = -(\varrho_{tt}, \chi)$  for all  $\chi \in S_h$ ,  $t \geq 0$ . Since  $(\theta_{tt}, \chi) = \frac{d}{dt}(\theta_t, \chi) - (\theta_t, \chi_t)$  we have for any  $\chi \in S_h$ ,  $t > 0$

$$\begin{aligned} -(\theta_t, \chi_t) + (\nabla\theta, \nabla\chi) &= -\frac{d}{dt}(\theta_t, \chi) - \frac{d}{dt}(\varrho_t, \chi) + (\varrho_t, \chi_t) \\ &= -\frac{d}{dt}(e_t, \chi) + (\varrho_t, \chi_t). \end{aligned} \quad (7.12)$$

Fix  $\xi > 0$  and let  $\hat{\chi}(t) := \int_t^\xi \theta(s) ds$ ,  $t \geq 0$ . Then  $\hat{\chi}(t) \in S_h$  for  $t \geq 0$ ,  $\hat{\chi}(\xi) = 0$ , and  $\hat{\chi}_t = -\theta$ . Select  $\chi = \hat{\chi}$  in (7.12) to obtain

$$(\theta_t, \theta) - (\nabla\hat{\chi}_t, \nabla\hat{\chi}) = -\frac{d}{dt}(e_t, \hat{\chi}) - (\varrho_t, \theta),$$

i.e.

$$\frac{1}{2} \frac{d}{dt} \|\theta\|^2 - \frac{1}{2} \frac{d}{dt} \|\nabla\hat{\chi}\|^2 = -\frac{d}{dt}(e_t, \hat{\chi}) - (\varrho_t, \theta), \quad t \geq 0.$$

Integrating both sides of the above with respect to  $t$  from 0 to  $\xi$ , we obtain

$$\begin{aligned} \|\theta(\xi)\|^2 - \|\theta(0)\|^2 - \|\nabla\hat{\chi}(\xi)\|^2 + \|\nabla\hat{\chi}(0)\|^2 \\ = -2(e_t(\xi), \hat{\chi}(\xi)) + 2(e_t(0), \hat{\chi}(0)) - 2 \int_0^\xi (\varrho_t, \theta) dt. \end{aligned}$$

Since  $\hat{\chi}(\xi) = 0$ ,  $\hat{\chi} \in S_h$ , we have for any  $\xi > 0$ :

$$\|\theta(\xi)\|^2 \leq \|\theta(0)\|^2 + 2(Pe_t(0), \hat{\chi}(0)) - 2 \int_0^\xi (\varrho_t, \theta) dt.$$

Hence

$$\|\theta(\xi)\|^2 \leq \|\theta(0)\|^2 + 2\|Pe_t(0)\| \|\hat{\chi}(0)\| + 2 \int_0^\xi \|\varrho_t\| \|\theta\| dt.$$

We recall that  $\hat{\chi}(0) = \int_0^\xi \theta(t) dt$ . Therefore  $\|\hat{\chi}(0)\| \leq \int_0^\xi \|\theta\| dt$  and therefore

$$\begin{aligned} \|\theta(\xi)\|^2 &\leq \|\theta(0)\|^2 + 2\|Pe_t(0)\| \int_0^\xi \|\theta\| dt + 2 \int_0^\xi \|\varrho_t\| \|\theta\| dt \\ &\leq \sup_{0 \leq s \leq \xi} \|\theta(s)\| (\|\theta(0)\| + 2\xi \|Pe_t(0)\|) + 2 \int_0^\xi \|\varrho_t\| \|\theta\| dt. \end{aligned}$$

Since  $\xi$  was arbitrary, the inequality above holds for all  $\xi' \in [0, \xi]$ . Let  $\tau \in [0, \xi]$  be a point where  $\|\theta(\tau)\| = \sup_{0 \leq s \leq \xi} \|\theta(s)\|$ . Applying the inequality for  $\xi' = \tau$  we get

$$\|\theta(\tau)\|^2 \leq \|\theta(\tau)\| \left( \|\theta(0)\| + 2\xi \|Pe_t(0)\| + 2 \int_0^\xi \|\varrho_t\| dt \right).$$

Hence, since  $\|\theta(\xi)\| \leq \|\theta(\tau)\|$

$$\|\theta(\xi)\| \leq \|\theta(0)\| + 2\xi\|Pe_t(0)\| + 2 \int_0^\xi \|\varrho_t\| dt,$$

for any  $\xi > 0$ . Therefore, for  $t \geq 0$ , using the definition of  $\varrho$

$$\|\theta(t)\| \leq \|\theta(0)\| + 2t\|Pe_t(0)\| + Ch^r \int_0^t \|u_t\|_r ds, \quad (7.13)$$

and

$$\begin{aligned} \|u_h(t) - u(t)\| &\leq \|\varrho(t)\| + \|\theta(t)\| \leq \|\varrho(0)\| + \int_0^t \|\varrho_t\| ds + \|\theta(t)\| \\ &\leq C \left\{ \|u_h^0 - Pu^0\| + t\|u_{t,h}^0 - Pu_t^0\| + h^r [\|u^0\|_r + \int_0^t \|u_t\|_r ds] \right\}, \end{aligned}$$

which is (7.11). (To get the last inequality of the right-hand side we used (7.13) and that  $\|\varrho(0)\| = \|R_h u^0 - u^0\| \leq Ch^r \|u^0\|_r$ ,  $\|\theta(0)\| \leq \|u_h^0 - Pu^0\| + Ch^r \|u^0\|_r$ , and  $Pe_t(0) = P(u_{t,h}^0 - u_t^0) = u_{t,h}^0 - Pu_t^0$ .)  $\square$

It is evident that (7.11) implies that  $\|u(t) - u_h(t)\| = O(h^r)$  if  $u_h^0, u_{t,h}^0$  are any optimal-order  $L^2$ -approximations of  $u^0, u_t^0$  in  $S_h$ , respectively.

### Remark

It is straightforward to check that analogs of Theorems 7.1 and 7.2 hold for the ibvp

$$\begin{cases} u_{tt} - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} (a_{ij}(x) \frac{\partial u}{\partial x_j}) + a_0(x)u = f(x, t), & x \in \Omega, t \geq 0, \\ u = 0 & \text{on } \partial\Omega, t \geq 0, \\ u(x, 0) = u^0(x), \quad u_t(x, 0) = u_t^0(x), & x \in \bar{\Omega}, \end{cases}$$

where the coefficients  $a_{ij}, a_0$  satisfy the conditions set forth in Exercise 3 of section 6.2. L.A. Bales has proved (Math. Comp. 43(1984), 383-414) that similar results hold when  $a_{ij}$  and  $a_0$  depend also on  $t$ .

## 7.3 Fully discrete schemes

In this section we shall examine some simple methods for discretizing in time the semidiscrete Galerkin equations (7.4). As usual we put  $t^n = nk$ ,  $n = 0, \dots, M$ ,  $Mk =$

$T > 0$ . We seek  $U^n \in S_h$ ,  $0 \leq n \leq M$ , approximations of the solution  $u$  of (7.1) at  $t^n$ , satisfying

$$\begin{cases} \frac{1}{k^2}(U^{n+1} - 2U^n + U^{n-1}, \chi) + (\nabla \hat{U}_\beta^n, \nabla \chi) = (\hat{f}_\beta^n, \chi), & \forall \chi \in S_h, \quad 1 \leq n \leq M-1, \\ U^0, U^1 \text{ given in } S_h, \end{cases} \quad (7.14)$$

where, for  $1 \leq n \leq M-1$  and  $\beta \geq 0$ ,  $\hat{U}_\beta^n := \beta U^{n+1} + (1-2\beta)U^n + \beta U^{n-1}$ ,  $\hat{f}_\beta^n = \beta f^{n+1} + (1-2\beta)f^n + \beta f^{n-1}$ ,  $f^n = f(\cdot, t^n)$ . Given  $U^{n-1}, U^n$ , (7.14) defines uniquely  $U^{n+1} \in S_h$  as solution of the linear system of equations

$$(G + \beta k^2 S) \alpha^{n+1} = (2G - (1-2\beta)k^2 S) \alpha^n - (G + \beta k^2 S) \alpha^{n-1} + k^2 \hat{F}_\beta^n,$$

where  $U^n = \sum_{i=1}^{N_h} \alpha_i^n \varphi_i$ ,  $\{\varphi_i\}$ ,  $G$ ,  $S$ , were defined in section 7.2, and  $\hat{F}_\beta^n$  is the  $N_h$ -vector with components  $(\hat{f}_\beta^n, \varphi_i)$ . If  $\beta = 0$  (7.14) corresponds to the classical Courant-Friedrichs-Lewy two-step scheme for the wave equation.

Taking for simplicity  $f = 0$  in (7.14) and putting  $\chi = U^{n+1} - U^{n-1}$  we obtain for  $1 \leq n \leq M-1$

$$(U^{n+1} - 2U^n + U^{n-1}, U^{n+1} - U^{n-1}) + k^2 (\nabla \hat{U}_\beta^n, \nabla (U^{n+1} - U^{n-1})) = 0. \quad (7.15)$$

Now

$$\begin{aligned} & (U^{n+1} - 2U^n + U^{n-1}, U^{n+1} - U^{n-1}) \\ &= ((U^{n+1} - U^n) - (U^n - U^{n-1}), (U^{n+1} - U^n) + (U^n - U^{n-1})) \\ &= \|U^{n+1} - U^n\|^2 - \|U^n - U^{n-1}\|^2. \end{aligned}$$

In addition

$$\begin{aligned} & (\nabla \hat{U}_\beta^n, \nabla (U^{n+1} - U^{n-1})) = (\nabla (\beta U^{n+1} + (1-2\beta)U^n + \beta U^{n-1}), \nabla (U^{n+1} - U^{n-1})) \\ &= \beta ((\nabla U^{n+1}, \nabla U^{n+1}) - (\nabla U^{n-1}, \nabla U^{n-1})) \\ &\quad + (1-2\beta) ((\nabla U^{n+1}, \nabla U^n) - (\nabla U^n, \nabla U^{n-1})). \end{aligned}$$

Hence, summing in (7.15) with respect to  $n$  from  $n = 1$  to  $l$ , where  $1 \leq l \leq M-1$ , we see that

$$\begin{aligned} & \|U^{l+1} - U^l\|^2 + \beta k^2 (\|\nabla U^{l+1}\|^2 + \|\nabla U^l\|^2) + (1-2\beta)k^2 (\nabla U^{l+1}, \nabla U^l) \\ &= \|U^1 - U^0\|^2 + \beta k^2 (\|\nabla U^1\|^2 + \|\nabla U^0\|^2) + (1-2\beta)k^2 (\nabla U^1, \nabla U^0), \end{aligned}$$

which, motivated by the identity

$$\beta(x^2 + y^2) + (1 - 2\beta)xy = \frac{(x + y)^2}{4} + \left(\beta - \frac{1}{4}\right)(x - y)^2,$$

we rewrite as

$$\begin{aligned} & \|U^{l+1} - U^l\|^2 + \frac{k^2}{4} \|\nabla(U^{l+1} + U^l)\|^2 + k^2\left(\beta - \frac{1}{4}\right) \|\nabla(U^{l+1} - U^l)\|^2 \\ &= \|U^1 - U^0\|^2 + \frac{k^2}{4} \|\nabla(U^1 + U^0)\|^2 + k^2\left(\beta - \frac{1}{4}\right) \|\nabla(U^1 - U^0)\|^2, \end{aligned} \quad (7.16)$$

which is valid for  $0 \leq l \leq M - 1$ . With the aid of the identity (7.16) we may prove the following *stability* result for the fully discrete scheme (7.14).

**Proposition 7.1.** *Suppose that there exists a constant  $c$ , independent of  $h$  and  $k$ , such that*

$$\|U^1 - U^0\| \leq ck \quad (7.17)$$

and

$$\|\nabla(U^1 \pm U^0)\| \leq c. \quad (7.18)$$

Then, there exists a constant  $C$ , independent of  $h$  and  $k$ , such that for all  $\beta \geq \frac{1}{4}$  the solution of (7.14) with  $f = 0$  satisfies

$$\max_{0 \leq n \leq M} \|U^n\| \leq C. \quad (7.19)$$

If  $0 \leq \beta < \frac{1}{4}$  and the inverse inequality (6.39) is valid in  $S_h$ , then (7.19) holds provided  $\frac{k}{h} \leq \alpha$ , where  $\alpha$  is some constant depending on  $C_*$  and  $\beta$ .

*Proof.* If  $\beta \geq \frac{1}{4}$ , (7.16) and (7.17) - (7.18) give, for  $l \geq 0$  and  $k \leq 1$ ,  $\|U^{l+1} - U^l\| \leq C$  and  $\|\nabla(U^{l+1} - U^l)\| \leq C$ . Hence, by Poincaré's inequality,  $\|U^{l+1} \pm U^l\| \leq C$ ,  $l \geq 0$ , and (7.19) follows since  $\|U^n\| = \left\| \left( \frac{U^{n+1} + U^n}{2} \right) - \left( \frac{U^{n+1} - U^n}{2} \right) \right\|$ .

If  $0 \leq \beta < \frac{1}{4}$ , (7.16) and (7.17) - (7.18) give for  $l \geq 0$

$$\|U^{l+1} - U^l\|^2 + \frac{k^2}{4} \|\nabla(U^{l+1} + U^l)\|^2 \leq k^2\left(\frac{1}{4} - \beta\right) \|\nabla(U^{l+1} - U^l)\|^2 + Ck^2.$$

Therefore, by the inverse inequality (6.39) we have

$$\|U^{l+1} - U^l\|^2 + \frac{k^2}{4} \|\nabla(U^{l+1} - U^l)\|^2 \leq \frac{k^2}{h^2} \left(\frac{1}{4} - \beta\right) C_*^2 \|U^{l+1} - U^l\|^2 + Ck^2,$$

i.e.

$$[1 - C_*^2 \frac{k^2}{h^2} (\frac{1}{4} - \beta)] \|U^{l+1} - U^l\|^2 + \frac{k^2}{4} \|\nabla(U^{l+1} + U^l)\|^2 \leq Ck^2, \quad l \geq 0.$$

Hence, if

$$\frac{k}{h} \leq \alpha < \frac{2}{C_* \sqrt{1 - 4\beta}} \quad (7.20)$$

we have that  $\|U^{l+1} \pm U^l\| \leq C(\alpha)$  and (7.19) follows.  $\square$

As we remarked in Section 6.4, a more general sufficient stability condition is

$$k [\lambda_{\max}(G^{-1}S)]^{1/2} < \frac{2}{\sqrt{1 - 4\beta}}. \quad (7.21)$$

The inverse inequality (6.39) and (7.20) imply (7.21).

We proceed now to derive  $L^2$ -error estimates for the scheme (7.14), following e.g. Dupont, op.cit. For simplicity we treat only the case  $\beta = \frac{1}{4}$ , i.e. the case of smallest value of  $\beta$  for which (7.19) holds unconditionally. The proof for the other  $\beta \geq 0$  follows along similar lines. The basic step of the proof is the following result.

**Proposition 7.2.** *Let  $u$  be the solution of (7.1),  $U^n$  the solution of the scheme (7.14) for  $\beta = \frac{1}{4}$ , and  $\theta^n = U^n - R_h u^n$ , where  $u^n = u(t^n)$ . Then, there exists a constant  $C$  independent of  $h$ ,  $k$  and  $T$  such that for  $0 \leq n \leq M - 1$*

$$\begin{aligned} \max_{0 \leq l \leq n} \left( \frac{1}{k} \|\theta^{l+1} - \theta^l\| + \|\nabla(\theta^{l+1} + \theta^l)\| \right) \leq C \left\{ \frac{1}{k} \|\theta^1 - \theta^0\| + \|\nabla(\theta^1 + \theta^0)\| \right. \\ \left. + \sqrt{T} \left[ h^r \left( \int_0^{t^{n+1}} \|u_{tt}\|_r^2 ds \right)^{1/2} + k^2 \left( \int_0^{t^{n+1}} \|\partial_t^4 u\|^2 ds \right)^{1/2} \right] \right\}. \quad (7.22) \end{aligned}$$

*Proof.* For  $1 \leq n \leq M - 1$  we denote  $\partial_\tau^2 g^n = \frac{1}{k^2}(g^{n+1} - 2g^n + g^{n-1})$ ,  $\hat{g}^n = \hat{g}_{1/4}^n = \frac{1}{4}(g^{n+1} + 2g^n + g^{n-1})$ . If  $U^n - u^n = (U^n - R_h u^n) + (R_h u^n - u^n) = \theta^n + \varrho^n$ , using (7.14) with  $\beta = 1/4$  and (7.1), we have for  $1 \leq n \leq M - 1$ ,  $\chi \in S_h$

$$\begin{aligned} (\partial_\tau^2 \theta^n, \chi) + (\nabla \hat{\theta}^n, \nabla \chi) &= (\hat{f}^n, \chi) - (\partial_\tau^2 (R_h u^n), \chi) - (\nabla \hat{u}^n, \nabla \chi) \\ &= (\hat{u}_{tt}^n - \partial_\tau^2 (R_h u^n), \chi) \\ &= (\partial_\tau^2 u^n - \partial_\tau^2 (R_h u^n) + \hat{u}_{tt}^n - \partial_\tau^2 u^n, \chi) \\ &= (-\partial_\tau^2 \varrho^n + \omega^n, \chi), \end{aligned}$$

where  $\omega^n := \hat{u}_{tt}^n - \partial_\tau^2 u^n$ . Therefore, for any  $\chi \in S_h$

$$(\theta^{n+1} - 2\theta^n + \theta^{n-1}, \chi) + k^2(\nabla \hat{\theta}^n, \nabla \chi) = k^2(-\partial_\tau^2 \varrho^n + \omega^n, \chi), \quad 1 \leq n \leq M-1. \quad (7.23)$$

We take now  $\chi = U^{n+1} - U^{n-1}$  and sum both sides of (7.23) with respect to  $n$  from  $n = 1$  to  $l$ , for  $1 \leq l \leq M-1$ . As in the proof of Proposition 7.1 we obtain

$$\begin{aligned} \|\theta^{l+1} - \theta^l\|^2 + \frac{k^2}{4} \|\nabla(\theta^{l+1} + \theta^l)\|^2 &= \|\theta^1 - \theta^0\|^2 + \frac{k^2}{4} \|\nabla(\theta^1 + \theta^0)\|^2 \\ &+ k^2 \sum_{n=1}^l (-\partial_\tau^2 \varrho^n + \omega^n, \theta^{n+1} - \theta^{n-1}). \end{aligned} \quad (7.24)$$

For any  $\varepsilon > 0$  we have

$$\begin{aligned} k^2 \sum_{n=1}^l (-\partial_\tau^2 \varrho^n + \omega^n, \theta^{n+1} - \theta^{n-1}) &\leq k^2 \sum_{n=1}^l \|-\partial_\tau^2 \varrho^n + \omega^n\| \|\theta^{n+1} - \theta^{n-1}\| \\ &\leq \frac{k^2}{2\varepsilon} \sum_{n=1}^l \|-\partial_\tau^2 \varrho^n + \omega^n\|^2 + \frac{\varepsilon k^2}{2} \sum_{n=1}^l \|\theta^{n+1} - \theta^{n-1}\|^2 \\ &\leq \frac{k^2}{\varepsilon} \sum_{n=1}^l \|\partial_\tau^2 \varrho^n\|^2 + \frac{k^2}{\varepsilon} \sum_{n=1}^l \|\omega^n\|^2 + \frac{\varepsilon k^2}{2} \sum_{n=1}^l (\|\theta^{n+1} - \theta^n\| + \|\theta^n - \theta^{n-1}\|)^2 \\ &\leq \frac{k^2}{\varepsilon} \sum_{n=1}^l \|\partial_\tau^2 \varrho^n\|^2 + \frac{k^2}{\varepsilon} \sum_{n=1}^l \|\omega^n\|^2 + 2\varepsilon k^2 \sum_{n=0}^l \|\theta^{n+1} - \theta^n\|^2. \end{aligned} \quad (7.25)$$

We estimate now the first two terms in the right-hand side of the above. By Taylor's theorem we have

$$\begin{aligned} \varrho^{n+1} &= \varrho^n + k \varrho_t^n + \int_{t^n}^{t^{n+1}} (t^{n+1} - \tau) \varrho_{tt}(\tau) d\tau, \\ \varrho^{n-1} &= \varrho^n - k \varrho_t^n + \int_{t^n}^{t^{n-1}} (t^{n-1} - \tau) \varrho_{tt}(\tau) d\tau. \end{aligned}$$

Therefore

$$\partial_\tau^2 \varrho^n = \frac{1}{k^2} \left[ \int_{t^n}^{t^{n+1}} (t^{n+1} - \tau) \varrho_{tt} d\tau - \int_{t^{n-1}}^{t^n} (t^{n-1} - \tau) \varrho_{tt} d\tau \right].$$

Since

$$\begin{aligned} \left( \int_{t^n}^{t^{n+1}} (t^{n+1} - \tau) \varrho_{tt} d\tau \right)^2 &\leq \int_{t^n}^{t^{n+1}} (t^{n+1} - \tau)^2 d\tau \int_{t^n}^{t^{n+1}} (\varrho_{tt})^2 d\tau \\ &\leq \frac{k^3}{3} \int_{t^n}^{t^{n+1}} (\varrho_{tt})^2 d\tau, \end{aligned}$$



we have

$$(\partial_\tau^2 \varrho^n)^2 \leq \frac{C}{k^4} k^3 \int_{t^{n-1}}^{t^{n+1}} (\varrho_{tt})^2 ds,$$

and we conclude by the definition of  $\varrho$  that

$$\sum_{n=1}^l \|\partial_\tau \varrho^n\|^2 \leq \frac{C}{k} \int_0^{t^{l+1}} \|\varrho_{tt}\|^2 ds \leq Ck^{-1}h^{2r} \int_0^{t^{l+1}} \|u_{tt}\|_r^2 ds.$$

Now

$$\omega^n = \hat{u}_{tt}^n - \partial_\tau^2 u^n = (\hat{u}_{tt}^n - u_{tt}^n) + (u_{tt}^n - \partial_\tau^2 u^n) =: \omega_1^n + \omega_2^n.$$

For  $\omega_1^n$  by Taylor's theorem we have

$$\omega_1^n = \frac{1}{4} \left[ \int_{t^n}^{t^{n+1}} (t^{n+1} - s) \partial_t^4 u ds + \int_{t^n}^{t^{n-1}} (t^{n-1} - s) \partial_t^4 u ds \right].$$

Hence  $(\omega_1^n)^2 \leq ck^3 \int_{t^{n-1}}^{t^{n+1}} (\partial_t^4 u)^2 ds$  and

$$\sum_{n=1}^l \|\omega_1^n\|^2 \leq ck^3 \int_0^{t^{l+1}} \|\partial_t^4 u\|^2 ds.$$

Similarly,

$$\sum_{n=1}^l \|\omega_2^n\|^2 \leq ck^3 \int_0^{t^{l+1}} \|\partial_t^4 u\|^2 ds.$$

We conclude that

$$\begin{aligned} \frac{k^2}{\varepsilon} \left( \sum_{n=1}^l \|\partial_\tau^2 \varrho^n\|^2 + \sum_{n=1}^l \|\omega^n\|^2 \right) &\leq \frac{ck}{\varepsilon} \left[ h^{2r} \int_0^{t^{l+1}} \|u_{tt}\|_r^2 ds + k^4 \int_0^{t^{l+1}} \|\partial_t^4 u\|^2 ds \right] \\ &=: \frac{ck}{\varepsilon} \sigma_l, \quad 1 \leq l \leq M-1. \end{aligned} \quad (7.26)$$

Therefore, by (7.24) - (7.26) we obtain for  $0 \leq l \leq M-1$

$$\begin{aligned} \|\theta^{l+1} - \theta^l\|^2 + \frac{k^2}{4} \|\nabla(\theta^{l+1} + \theta^l)\|^2 &\leq \|\theta^1 - \theta^0\|^2 + \frac{k^2}{4} \|\nabla(\theta^1 + \theta^0)\|^2 \\ &\quad + \frac{ck}{\varepsilon} \sigma_l + 2\varepsilon k^2 \sum_{n=0}^l \left( \|\theta^{n+1} - \theta^n\|^2 + \frac{k^2}{4} \|\nabla(\theta^{n+1} + \theta^n)\|^2 \right) \\ &\leq \|\theta^1 - \theta^0\|^2 + \frac{k^2}{4} \|\nabla(\theta^1 + \theta^0)\|^2 + \frac{ck}{\varepsilon} \sigma_l \\ &\quad + 2\varepsilon kT \max_{0 \leq n \leq l} \left( \|\theta^{n+1} - \theta^n\|^2 + \frac{k^2}{4} \|\nabla(\theta^{n+1} + \theta^n)\|^2 \right), \end{aligned}$$

since  $(l+1)k \leq T$ . Choose now  $\varepsilon = \frac{1}{4Tk}$ . Then, for  $0 \leq l \leq M-1$

$$\begin{aligned} \|\theta^{l+1} - \theta^l\|^2 + \frac{k^2}{4} \|\nabla(\theta^{l+1} + \theta^l)\|^2 &\leq \|\theta^1 - \theta^0\|^2 + \frac{k^2}{4} \|\nabla(\theta^1 + \theta^0)\|^2 \\ &\quad + CTk^2 \sigma_l + \frac{1}{2} \max_{0 \leq n \leq l} \left( \|\theta^{n+1} - \theta^n\|^2 + \frac{k^2}{4} \|\nabla(\theta^{n+1} + \theta^n)\|^2 \right). \end{aligned} \quad (7.27)$$

Fix  $l$  for the moment and let  $m$ ,  $0 \leq m \leq l$ , be an integer for which

$$\max_{0 \leq n \leq l} \left( \|\theta^{n+1} - \theta^n\|^2 + \frac{k^2}{4} \|\nabla(\theta^{n+1} + \theta^n)\|^2 \right) = \|\theta^{m+1} - \theta^m\|^2 + \frac{k^2}{4} \|\nabla(\theta^{m+1} + \theta^m)\|^2.$$

Then, from (7.27), since  $\sigma_n$  is increasing with  $n$

$$\begin{aligned} \|\theta^{m+1} - \theta^m\|^2 + \frac{k^2}{4} \|\nabla(\theta^{m+1} - \theta^m)\|^2 &\leq \|\theta^1 - \theta^0\|^2 + \frac{k^2}{4} \|\nabla(\theta^1 + \theta^0)\|^2 \\ &\quad + CTk^2\sigma_l + \frac{1}{2} \left( \|\theta^{m+1} - \theta^m\|^2 + \frac{k^2}{4} \|\nabla(\theta^{m+1} + \theta^m)\|^2 \right). \end{aligned}$$

Therefore

$$\|\theta^{m+1} - \theta^m\|^2 + \frac{k^2}{4} \|\nabla(\theta^{m+1} - \theta^m)\|^2 \leq 2[\|\theta^1 - \theta^0\|^2 + \frac{k^2}{4} \|\nabla(\theta^1 + \theta^0)\|^2] + CTk^2\sigma_l.$$

But

$$\|\theta^{l+1} - \theta^l\|^2 + \frac{k^2}{4} \|\nabla(\theta^{l+1} + \theta^l)\|^2 \leq \|\theta^{m+1} - \theta^m\|^2 + \frac{k^2}{4} \|\nabla(\theta^{m+1} + \theta^m)\|^2.$$

Hence,

$$\|\theta^{l+1} - \theta^l\|^2 + \frac{k^2}{4} \|\nabla(\theta^{l+1} + \theta^l)\|^2 \leq C[\|\theta^1 - \theta^0\|^2 + \frac{k^2}{4} \|\nabla(\theta^1 + \theta^0)\|^2 + Tk^2\sigma_l].$$

Dividing both sides of the above by  $k^2$  and taking square roots we obtain (7.22) in view of (7.26).  $\square$

**Exercise 1.** Using the technique of the proof of Proposition 7.1 show that an estimate of the form (7.22) holds for the solution of (7.14) for any  $\beta \geq 0$ , provided the stability condition (7.20) holds if  $0 \leq \beta < \frac{1}{4}$ .

**Exercise 2.** The scheme (7.14) in the case  $\beta = \frac{1}{12}$  is known as the *Störmer-Numerov method*. Show that this scheme is fourth-order accurate in time: Specifically prove that for  $\beta = \frac{1}{12}$

$$\sum_{n=1}^l \|\omega^n\|^2 \leq ck^7 \int_0^{t^{l+1}} \|\partial_t^8 u\|^2 ds,$$

and consequently that the last term in the right-hand side of (7.22) is of  $O(k^4)$ , provided (7.20) holds with  $\beta = \frac{1}{12}$ . (The Störmer-Numerov scheme is the only fourth-order accurate in  $k$  scheme of the family (7.14). For all other  $\beta \geq 0$  the temporal truncation error is of  $O(k^2)$ .)

We present now some straightforward implications of Proposition 7.2.

**Proposition 7.3.** *Let the hypotheses of Proposition 7.2 hold and assume in addition that for some constant  $C$  independent of  $h$  and  $k$  we have*

$$\frac{1}{k}\|\theta^1 - \theta^0\| + \|\nabla(\theta^1 + \theta^0)\| \leq C(k^2 + h^r). \quad (7.28)$$

Then, there is a constant  $C = C(u, T)$  such that

- (i)  $\|\frac{1}{k}(U^{n+1} - U^n) - u_t(t^{n+1/2})\| \leq C(k^2 + h^r), \quad t^{n+1/2} = t^n + \frac{k}{2}, \quad 0 \leq n \leq M - 1,$
- (ii)  $\|\frac{1}{2}\nabla(U^{n+1} + U^n) - \nabla u(t^{n+1/2})\| \leq C(k^2 + h^{r-1}), \quad 0 \leq n \leq M - 1,$
- (iii)  $\max_{0 \leq n \leq M} \|U^n - u^n\| \leq C(k^2 + h^r).$

*Proof.* (i). Let  $0 \leq n \leq M - 1$ . Since

$$\begin{aligned} \frac{1}{k}(U^{n+1} - U^n) - u_t(t^{n+1/2}) &= \frac{1}{k}(\theta^{n+1} - \theta^n) + \frac{1}{k}R_h(u^{n+1} - u^n) - u_t(t^{n+1/2}) \\ &= \frac{1}{k}(\theta^{n+1} - \theta^n) + R_h\left(\frac{u^{n+1} - u^n}{k}\right) - \left(\frac{u^{n+1} - u^n}{k}\right) + \left(\frac{u^{n+1} - u^n}{k} - u_t(t^{n+1/2})\right) \\ &= \frac{1}{k}(\theta^{n+1} - \theta^n) + \frac{1}{k} \int_{t^n}^{t^{n+1}} (R_h u_t - u_t) ds + \left(\frac{1}{k}(u^{n+1} - u^n) - u_t(t^{n+1/2})\right), \end{aligned}$$

we have

$$\begin{aligned} \left\| \frac{1}{k}(U^{n+1} - U^n) - u_t(t^{n+1/2}) \right\| &\leq \frac{1}{k} \|\theta^{n+1} - \theta^n\| + Ch^r \max_{t^n \leq s \leq t^{n+1}} \|u_t(s)\|_r \\ &\quad + Ck^2 \max_{t^n \leq s \leq t^{n+1}} \|u_{ttt}(s)\|, \end{aligned}$$

and (i) follows from (7.22) and (7.28).

(ii). Let  $0 \leq n \leq M - 1$ . Since

$$\begin{aligned} \frac{1}{2}\nabla(U^{n+1} + U^n) - \nabla u(t^{n+1/2}) &= \frac{1}{2}\nabla(\theta^{n+1} + \theta^n) + \frac{1}{2}\nabla(R_h(u^{n+1} + u^n) - (u^{n+1} + u^n)) \\ &\quad + \nabla\left(\frac{1}{2}(u^{n+1} + u^n) - u(t^{n+1/2})\right), \end{aligned}$$

we have

$$\begin{aligned} \left\| \frac{1}{2}\nabla(U^{n+1} + U^n) - \nabla u(t^{n+1/2}) \right\| &\leq \frac{1}{2} \|\nabla(\theta^{n+1} + \theta^n)\| + ch^{r-1}(\|u^{n+1}\|_r + \|u^n\|_r) \\ &\quad + Ck^2 \max_{t^n \leq s \leq t^{n+1}} \|\nabla u_{tt}(s)\|, \end{aligned}$$

and (ii) follows from (7.22) and (7.28).

(iii). Since

$$U^l - u^l = \theta^l + \varrho^l, \quad 0 \leq l \leq M,$$

we have

$$\|U^l - u^l\| \leq \|\theta^l\| + Ch^r \|u^l\|_r, \quad 0 \leq l \leq M. \quad (7.29)$$

But for  $0 \leq n \leq M - 1$ ,  $\theta^{n+1} = \frac{\theta^{n+1} + \theta^n}{2} + \frac{k}{2} \left( \frac{\theta^{n+1} - \theta^n}{k} \right)$ . Hence by (7.22), (7.28) and Poincaré's inequality

$$\begin{aligned} \|\theta^{n+1}\| &\leq \frac{1}{2} \|\theta^{n+1} + \theta^n\| + \frac{k}{2} \frac{1}{k} \|\theta^{n+1} - \theta^n\| \\ &\leq C \|\nabla(\theta^{n+1} + \theta^n)\| + \frac{k}{2} \left\| \frac{\theta^{n+1} - \theta^n}{k} \right\| \leq C(k^2 + h^r), \end{aligned}$$

for  $0 \leq n \leq M - 1$ . In addition, since  $\theta^0 = \frac{\theta^1 + \theta^0}{2} - \frac{k}{2} \left( \frac{\theta^1 - \theta^0}{k} \right)$ , we have by the Poincaré inequality and (7.28)

$$\|\theta^0\| \leq \frac{1}{2} \|\theta^1 + \theta^0\| + \frac{k}{2} \left\| \frac{\theta^1 - \theta^0}{k} \right\| \leq \frac{C}{2} \|\nabla(\theta^1 + \theta^0)\| + \frac{k}{2} \frac{\|\theta^1 - \theta^0\|}{k} \leq C(k^2 + h^r).$$

Therefore (iii) follows from (7.29) and these estimates. □

We turn now to the matter of choosing  $U^0$  and  $U^1$  in  $S_h$  so that (7.28) holds. We choose first

$$U^0 = R_h u^0. \quad (7.30)$$

This implies that  $\theta^0 = 0$ ; hence,  $U^1$  must be chosen so that

$$\frac{1}{k} \|\theta^1\| + \|\nabla \theta^1\| \leq C(k^2 + h^r). \quad (7.31)$$

A straightforward way of doing this is using Taylor expansions:

Let

$$u^*(x) = u^0(x) + k u_t^0(x) + \frac{k^2}{2} (\Delta u^0(x) + f(x, 0)), \quad x \in \bar{\Omega}. \quad (7.32)$$

Since by (7.1)  $u_{tt} = \Delta u + f$ , it is clear that  $u^* \upharpoonright_{\partial\Omega} = 0$  and that we have

$$u^* = u(k) + O(k^3), \quad \nabla u^* = \nabla u(k) + O(k^3).$$

We let

$$U^1 = R_h u^*. \quad (7.33)$$

Then by Poincaré's inequality

$$\begin{aligned}\|\theta^1\| &= \|U^1 - R_h u(k)\| = \|R_h(u^* - u(k))\| \\ &\leq C\|\nabla R_h(u^* - u(k))\| \leq C\|\nabla(u^* - u(k))\| \leq Ck^3,\end{aligned}$$

and

$$\|\nabla\theta^1\| = \|\nabla R_h(u^* - u(k))\| \leq \|\nabla(u^* - u(k))\| \leq Ck^3.$$

Therefore  $\frac{1}{k}\|\theta^1\| + \|\nabla\theta^1\| \leq Ck^2$  and (7.31) holds. It is then straightforward to check that the initial conditions (7.30) and (7.33) satisfy the hypotheses (7.17) and (7.18) for the  $L^2$ -stability of the scheme (7.14). Indeed, we have by Poincaré's inequality, that

$$\|U^1 - U^0\| = \|R_h(u^* - u^0)\| \leq C\|\nabla(u^* - u^0)\| \leq Ck,$$

and

$$\|\nabla(U^1 \pm U^0)\| = \|\nabla R_h(u^* \pm u_0)\| \leq C\|\nabla(u^* \pm u_0)\| \leq C.$$

**Exercise 3.** Consider the Störmer-Numerov method (see Exercise 2.). Take  $U^0 = R_h u^0$  and  $U^1 = R_h u^{**}$ , where  $u^{**} = u^0 + ku_t^0 + \frac{k^2}{2}u_{tt}(0) + \frac{k^3}{3!}\partial_t^3 u(0) + \frac{k^4}{4!}\partial_t^4 u(0)$ . Prove that with these choices one has

$$\frac{1}{k}\|\theta^1 - \theta^0\| + \|\nabla(\theta^1 + \theta^0)\| \leq Ck^4, \quad (7.34)$$

so that by Exercises 1 and 2 and Proposition 7.3 (iii) one obtains for the Störmer-Numerov scheme the error estimate

$$\max_{0 \leq n \leq M} \|U^n - u^n\| \leq C(k^4 + h^r),$$

provided (7.17) holds for  $\beta = \frac{1}{12}$ . (Note that for the wave equation,

$$\begin{aligned}\partial_t^2 u(0) &= \Delta u^0 + f(0), & \partial_t^3 u(0) &= \Delta u_t^0 + f_t(0), \\ \partial_t^4 u(0) &= \partial_t^2(\Delta u + f) \upharpoonright_{t=0} = \Delta^2 u^0 + \Delta f(0) + f_{tt}(0),\end{aligned}$$

so that, in principle,  $u^{**}$  can be computed by the data of (7.1).) Finally show that these choices in  $U^1$  and  $U^0$  also satisfy the estimates (7.17) and (7.18).

### Remarks

a. The choice  $U^1 = R_h u^*$  has the disadvantage that it needs the computation of  $\Delta u^0$ .

Alternatively (see Dougalis and Serbin, *Comput. Math. Appl.* 7(1981), 261-279)), one may compute initial conditions for the scheme (7.14) for  $\beta \neq \frac{1}{12}$  as follows: (For simplicity we consider only the homogeneous equation,  $f = 0$ ). Take again  $U^0 = R_h u^0$  and compute  $U^1 \in S_h$  as solution of the problem

$$(U^1, \chi) + \beta k^2 (\nabla U^1, \nabla \chi) = (U^0, \chi) + (\beta - \frac{1}{2}) k^2 (\nabla U^0, \nabla \chi) + k(u_t^0, \chi), \quad \forall \chi \in S_h.$$

It can be shown that for this choice of  $U^0$  and  $U^1$ , (7.28) and (7.17)-(7.18) hold.

**b.** In the case of the Störmer-Numerov method ( $\beta = \frac{1}{12}$ ) the choice  $U^1 = R_h u^{**}$  (see Exercise 3 above) requires computing  $\Delta u^0$ ,  $\Delta u_t^0$ ,  $\Delta^2 u^0$ ,  $f_t(0)$ ,  $f_{tt}(0)$ ,  $\Delta f(0)$ , which may be difficult or impossible in practice. A more efficient way (see Dougalis and Serbin, *op. cit.*) of computing  $U^0$ ,  $U^1$ , so that (7.34) and (7.17)-(7.18) hold, is the following. (We take  $f = 0$  for simplicity.) Define again  $U^0 = R_h u^0$  and compute successively  $U^{0,1}$ ,  $U^{0,2}$ , and  $U^1$  in  $S_h$  by the equations

$$(U^{0,1}, \chi) + \frac{k^2}{12} (\nabla U^{0,1}, \nabla \chi) = 6(U^0, \chi) + \frac{k^2}{2} (\nabla U^0, \nabla \chi) + k(u_t^0, \chi), \quad \forall \chi \in S_h,$$

$$(U^{0,2}, \chi) + \frac{k^2}{12} (\nabla U^{0,2}, \nabla \chi) = (U^{0,1}, \chi), \quad \forall \chi \in S_h,$$

$$U^1 = U^{0,2} - 5U^0.$$

**c.** For higher-order accurate in time full discretizations of the second-order semidiscrete problem (7.4) one may use e.g. the so-called *cosine methods* (cf. e.g. Baker et al., *Numer. Math.* 35(1980), 127-142, *RAIRO Anal. Numer.* 13 (1979), 201-226), extended to problems with time-dependent coefficients by Bales et al., *Math. Comp.* 45(1985), 65-89, and to nonlinear problems by Bales and Dougalis 52(1989) 299-319, and Makridakis, *Comput. Math. Appl.* 19(1990), 19-34, or *linear multistep methods* (cf. e.g. Dougalis, *Math. Comp.* 33(1979), 563-584), etc.

Another class of fully discrete Galerkin methods for the wave equation is obtained by writing (7.1) as a *first-order system* in the temporal variable:

$$\begin{cases} q_t = \Delta u + f, & x \in \Omega, t \geq 0, \\ u_t = q, & x \in \Omega, t \geq 0, \\ u = 0, q = 0, & x \in \partial\Omega, t \geq 0, \\ u(x, 0) = u^0(x), q(x, 0) = u_t^0, & x \in \bar{\Omega}. \end{cases} \quad (7.35)$$

We may then consider the Galerkin semidiscretization of (7.35) and its temporal discretization by single-step or multistep schemes. For example, consider the *trapezoidal method* in which we seek  $U^n, Q^n$  in  $S_h$  for  $0 \leq n \leq M$  satisfying

$$U^0 = Pu^0, \quad Q^0 = Pu_t^0,$$

and for  $n = 0, 1, \dots, M - 1$ :

$$\begin{cases} \left( \frac{Q^{n+1} - Q^n}{k}, \chi \right) + \frac{1}{2} \left( \frac{U^{n+1} + U^n}{2}, \nabla \chi \right) = \frac{1}{2} (f^{n+1} + f^n, \chi), & \forall \chi \in S_h, \\ \frac{1}{k} (U^{n+1} - U^n) = \frac{1}{2} (Q^{n+1} + Q^n). \end{cases}$$

(Here  $P$  is the  $L^2$ -projection onto  $S_h$ .) The scheme is easily solvable for  $Q^{n+1}$  by substituting  $U^{n+1}$  from the second equation into the first one. Taking  $\chi = U^{n+1} - U^n$  in the first equation and the  $L^2$ -inner product of both sides of the second equation by  $Q^{n+1} - Q^n$  we easily obtain the stability estimate

$$\|Q^n\|^2 + \frac{k}{2} \|\nabla U^n\|^2 = \|Q^0\|^2 + \frac{k}{2} \|\nabla U^0\|^2,$$

that may be viewed as a discrete analog of (7.2). Baker, SIAM J. Numer. Anal., 13(1976), 564-576, has analyzed the convergence of the scheme and shown that  $\max_n \|U^n - u^n\| = O(k^2 + h^r)$ . Higher-order temporal discretizations for the Galerkin semidiscretization of (7.35) have been studied by Baker and Bramble, RAIRO Anal. Numer. 13(1979), 75-100, and extended by Bales (Math. Comp. 43(1984), 383-414, SIAM J. Numer. Anal., 23(1986), 27-43, Comput. Math. Appl. 12A(1986), 581-604) to the case of equations with time-dependent coefficients and nonlinear terms.

# References

## INTRODUCTORY

- [1.1] G. Strang and G.J. Fix, *An Analysis of the Finite Element Method*, Prentice Hall 1973.
- [1.2] M.H. Schultz, *Spline Analysis*, Prentice Hall 1973.
- [1.3] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press 1989.

## SOBOLEV SPACES AND PDE's

- [2.1] R.A. Adams, *Sobolev Spaces*, Academic Press 1975.
- [2.2] H. Brezis, *Analyse fonctionnelle, théorie et applications*, Masson, Paris, 1983.  
(Greek translation, University Press N.T.U.A., 1997)
- [2.3] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer 2011.
- [2.4] L.C. Evans, *Partial Differential Equations*, American Math. Society, Providence 1998.
- [2.5] H. Triebel, *Höhere Analysis*, VEB, Berlin 1972.

## BOOKS EMPHASIZING THE MATHEMATICAL THEORY

- [3.1] A.K. Aziz (editor), *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press 1972.



- [3.2] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York 1994, 2<sup>nd</sup> ed 2002.
- [3.3] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North Holland 1978. (Reprinted SIAM, 2002)
- [3.4] P.G. Ciarlet and J.L. Lions, editors. *Handbook of Numerical Analysis* Vol. 2, Pt. 1, Elsevier 1991.
- [3.5] P.A. Raviart et J.M. Thomas, *Introduction à l'Analyse Numérique des équations aux dérivées partielles*, Masson, Paris 1983
- [3.6] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag 1997.
- [3.7] S. Larsson and V. Thomée, *Partial Differential Equations with Numerical Methods*, Springer-Verlag 2009.

### **FINITE ELEMENTS IN ENGINEERING**

- [4.1] T.J.R. Hughes, *The Finite Element Method*, Dover 2000.
- [4.2] O.C. Zienkiewicz, *The Finite Element Method*, 3rd ed., McGraw Hill 1977.
- [4.3] M. Bernadou et al., *MODULEF, a Modular Library of Finite Elements*, INRIA, France 1988.

### **SPLINES**

- [5.1] C. de Boor, *A Practical Guide to Splines*, Springer–Verlag 1978.

### **NUMERICAL ANALYSIS AND ODE's (In Greek)**

- [6.1] G.D. Akrivis and V.A. Dougalis, *Introduction to Numerical Analysis*, Crete University Press, 4<sup>th</sup> revised edition 2010. (In Greek).
- [6.2] G.D. Akrivis and V.A. Dougalis, *Numerical Methods for Ordinary Differential Equations*, Crete University Press 2006. (In Greek).

- [6.3] V.A. Dougalis, *Numerical Analysis: Lecture Notes for a graduate-level course*, 1987. (In Greek).