Chapter 5

THE FINITE ELEMENT METHOD

The finite element method is a rather general technique for the construction of approximate solutions to boundary value problems. As in the finite difference method the domain is covered by a grid. This grid creates small sub-domains, called *finite elements*. The idea is to approximate the governing equations within these elements by functions defined in finite dimensional functions spaces. We will begin this chapter with a simple one dimensional problem.

5.1 The model problem

Suppose we want to solve the equation

$$-\frac{d^2u}{dx^2} + u = x \quad x \in (0,1)\ , \tag{5.1}$$

with $u(0) = u(1) = 0$. This is a one dimensional *Helmholtz equation* and this type of equation models the deflection of a string on an elastic foundation or the temperature distribution in a rod.

The data for this problem is all the additional information apart from the differential equation. So the domain over which the problem is stated, in our case the open interval $(0,1)$, the right-hand-side vector, in our case $f = x$ and the boundary conditions, $u(0) = u(1) = 0$. The boundary conditions are called *homogeneous*. The advantage of homogeneous boundary conditions is that all the functions which satisfy the homogeneous boundary conditions form a linear space, i.e. if v_1 and v_2 satisfy the homogeneous boundary conditions, then so will any linear combination of these two functions. The

right-hand-side constains a smooth function, $f = x$. In fact, the function $f(x) = x$ is infinitely many times differentiable. In constrast to a smooth right-hand-side function, we can also define a right-hand-side function which is not a function in the classical sense. An example of such a differential equation would be

$$-\frac{d^2u}{dx^2} + u = \delta(x - \frac{1}{2}) \quad x \in (0,1) \quad u(0) = u(1) = 0 \ . \tag{5.2}$$

Instead of trying to approximate differential equations in the form given by (5.1) and (5.2) we are going to convert them to the so-called *weak* or *variational* form. Whenever, the differential possesses a classical solution, i.e. a solution which is twice differentiable, the solution to the weak problem will coincide with the solution of the differential equation.

5.1.1 Variational formulation

One weak statement of the model problem (5.1) is given as follows: find the function u such that the differential equation, together with the boundary conditions, are satisfied in the sense of weighted averages. By the satisfaction of all "weighted averages" of the differential equation, we mean that we require

$$\int_0^1 \left(-\frac{d^2u}{dx^2} + u\right) v \, dx = \int_0^1 xv \, dx \ , \tag{5.3}$$

for all members v of a suitable class of functions. This class of *weight functions* or *test functions* is chosen such that the integrals in the weak formulation make sense. Instead of the weak formulation given above, the weak formulation is usually converted, using integration by parts to the first term.

$$\int_0^1 -\frac{d^2u}{dx^2} v \, dx = \int_0^1 \frac{du}{dx}\frac{dv}{dx} \, dx + \left[\frac{du}{dx} v\right]_0^1 \ . \tag{5.4}$$

If we require that the test functions vanish at the end-points the weak formulation can be written as

$$\int_0^1 \left(\frac{du}{dx}\frac{dv}{dx} + uv - xv\right) dx = 0 \ , \tag{5.5}$$

for all v which satisfy $v(0) = v(1) = 0$ and for which the integral makes sense. Alternatively we could chose the solution u from the space of functions for which this integral

makes sense. For instance, if we require, that u and v are elements opf $L^2(0,1)$, i.e. all the functions which are square integrable over the interval $(0,1)$ we have by the Cauchy-Schwartz inequality (see Section 5.5.1)

$$\left| \int_0^1 uv\,dx \right|^2 \leq \left(\int_0^1 u^2\,dx \right) \left(\int_0^1 v^2\,dx \right) < \infty \,. \tag{5.6}$$

So the second term in our variational formulation yields a finite value in the integral if we require that both u and v belong to $L^2(0,1)$. Likewise, the last term yields a finite value because $x \in L^2(0,1)$. However, the requirement that $u, v \in L^2(0,1)$ is necessary but not sufficient, due to the presence of the first term. Take for example $u = v = \sqrt{x} \in L^2(0,1)$, then

$$\int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx = \frac{1}{4}\int_0^1 x^{-1}\,dx = \infty \,. \tag{5.7}$$

An additional requirement would be, that also the first derivative of the test functions belongs to $L^2(0,1)$. So the space of admissable functions becomes

$$H^1(0,1) = \left\{ u \,|\, u \in L^2(0,1) \text{ and } \frac{du}{dx} \in L^2(0,1) \right\} \,. \tag{5.8}$$

In order to indicate that both u and v vanish at the end-points we define

$$H_0^1(0,1) = \left\{ u \in H^1(0,1) \,|\, u(0) = u(1) = 0 \right\} \,. \tag{5.9}$$

The space $H^1(0,1)$ is called a *Sobolev* space. Sobolev spaces were introduced by S.L. Sobolev around 1950. The superscript '1' refers to the fact that the first derivative has to be square integrable. Analogously, one can define $H^2(0,1)$, i.e. the space of function in which the function itself, the first derivative and the second derivative are square integrable, and so on. It suffices for our example to require that $u, v \in H_0^1(0,1)$.

With the Sobolev space $H^k(0,1)$, where k is a non-negative integer, we can associate an inner-product. This inner-product is defined as

$$(u,v)_{H^k(0,1)} := \int_0^1 \sum_{i=0}^{k} \frac{d^i u}{dx^i}\frac{d^i v}{dx^i}\,dx \,. \tag{5.10}$$

In order to see that this is an inner product we have to look at the definition of the inner-product

Definition 3 *An inner product space (or a pre-Hilbert space) is a vector space X with an inner product defined on X. The inner-product on X is a mapping from $X \times X$ into the scalar field K of X (K can be either the space of real numbers $\mathcal{R}$ or the space of complex numbers $\mathcal{R}$; that is, with every pair x and y from X there is associated a scalar which is written as*

$$(x, y)_X \ ,$$

and is called the inner product (or scalar product) of x and y, such that for all elements x, y and z and scalars α we have

$$(x + y, z)_X = (x, z)_X + (y, z)_X \ , \tag{5.11}$$

$$(\alpha x, y)_X = \alpha (x, y)_X \ , \tag{5.12}$$

$$(x, y)_X = \overline{(y, x)}_X \ , \tag{5.13}$$

and

$$(x, x)_X \geq 0 \ \ , \ \ (x, x)_X = 0 \iff x = 0 \ . \tag{5.14}$$

An inner product defines a norm on X by

$$\|x\|_X = \sqrt{(x, x)_X} \ . \tag{5.15}$$

A norm is defined as follows

Definition 4 *A norm is a functional on a vector space X, i.e. a mapping from elements of X to the space of scalars (real or complex), satifying*

$$\|x\| \geq 0 \ , \tag{5.16}$$

$$\|x\| = 0 \iff x = 0 \ , \tag{5.17}$$

$$\|\alpha x\| = |\alpha| \|x\| \ , \tag{5.18}$$

and

$$\|x+y\| \leq \|x\| + \|y\| \ , \quad \textit{The triangle inequality.} \tag{5.19}$$

In some cases it is convenient to drop the second requirement (5.17) in which case the functional is called *a semi-norm*. So for a semi-norm $\|x\| = 0$ does not imply that $x = 0$. in which case the functional is called *a semi-norm*. So for a semi-norm $\|x\| = 0$ does not imply that $x = 0$.

So the weak variational formulation can also be written as

$$(u,v)_{H_0^1(0,1)} = \int_0^1 xv\,dx \quad \forall v \in H_0^1(0,1) \ . \tag{5.20}$$

Another concept that we will use frequently, is the concept of the dual space. We have seen that we can define a norm on the function space $H_0^1(0,1)$ by (5.15). This means that $H_0^1(0,1)$ is a *normed space*. Note that all inner-product spaces are normed spaces, but the converse is not necessarily true. We define the dual space as

Definition 5 *Let X be normed space. Then the set of all bounded linear functionals on X constitute a space with norm defined by*

$$\|f\| = \sup_{x\in X, x\neq 0} \frac{|f(x)|}{\|x\|_X} = \sup_{x\in X, \|x\|_X=1} |f(x)| = \sup_{x\in X, \|x\|_X=1} f(x) < \infty \ , \tag{5.21}$$

which is called the dual space of X and is denoted by X'.

The boundedness of the elements of X' is expressed by (5.21) and the linearity implies that for all $x, y \in X$ and scalars α_1 and α_2 all elements $f \in X'$ satisfy

$$f(\alpha_1 x + \alpha_2 y) = \alpha_1 f(x) + \alpha_2 f(y) \ . \tag{5.22}$$

Instead of the notation $f(x)$, this operators is usually denoted by $< f, x >_X$ and is called *duality pairing on X*. The dual space of the Sobolev space $H^k(0,1)$ is denoted by $H^{-k}(0,1)$ So in our model problem the right-hand-side can be written as

$$< x, v >_{H^1(0,1)} = \int_0^1 xv\,dx \quad \forall v \in H_0^1(0,1) \ . \tag{5.23}$$

So, in the new notation the weak formulation reads

Find $u \in H_0^1(0,1)$ such that for $x \in H^{-1}(0,1)$ such that

$$(u,v)_{H^1(0,1)} =< x, v >_{H^1(0,1)} \quad \forall v \in H_0^1(0,1) \; . \tag{5.24}$$

5.1.2 Lax-Milgram Theorem

Before diving into methods to approximate the solution of the weak problem stated before, it is necessary to establish under which conditions we have a unique solution. One cannot expect to find approximate solutions to problems which don't even possess a solution. The necessary ingredients for existence and uniqueness are provided by the Lax-Milgram Theorem. In an abstract setting the question is:

Find $u \in X$ such that

$$a(u,v) = F(v) \quad \forall v \in V \; , \tag{5.25}$$

where X is a complete inner-product space (a Hilbert space), $F \in X'$ and $a(\cdot,\cdot)$ is a bilinear form. In our sample problem this bilinear form was given by

$$a(u,v) = \int_0^1 (u'v' + uv)\, dx \; . \tag{5.26}$$

If this bilinear form is continuous and coercive then the Lax-Milgram Theorem states that this variational form has a unique solution. Let us therefore define the concepts of continuity and coercivity.

Definition 6 *A bilinear form $a(\cdot,\cdot)$ on a normed linear space X is said to be* **bounded** *(or* **continuous***) if there exists a constant $c_1 < \infty$ such that*

$$a(u,v) \leq c_1 \|u\|_X \|v\|_X \quad \forall u, v \in X \tag{5.27}$$

The smallest constant c_1 for which this inequality holds, for all functions u and v is usually denoted by $\|a\|$.

Note that in this definition no distinction is made between continuity and boundedness. In fact, the definition only defines boundedness for the bilinear form, but it is known from functional analysis that boundedness is equivalent to continuity, see Section 5.5.2. The second concept to be defined is *coercivity*.

Definition 7 *A bilinear form $a(\cdot,\cdot)$ on a normed linear space X is said to be coercive if there exists a constant $c_2 > 0$ such that*

$$a(v,v) \geq c_2 \|v\|_X \quad \forall v \in X \ . \tag{5.28}$$

An immediate consequence of coercivity is that $a(v,v) = 0$ implies that $v = 0$. If, in addition, $a(\cdot,\cdot)$ is symmetric the bilinear form defines an inner-product on the space X. Coercivity implies that the bilinear operator has a continuous (bounded) inverse.

Now that we are familiar with the ingredients, let's turn to the celebrated Lax-Milgram Theorem

Theorem 3 *Given an Hilbert space X and a continuous, coercive, bilinear form $a(\cdot,\cdot)$ and a continuous linear functional $F \in X'$, there exists a unique $u \in X$ such that*

$$a(u,v) = F(v) \quad \forall v \in X \ . \tag{5.29}$$

In order to prove this Theorem we need the following Lemma

Lemma 1 *Given a Banach space (a complete normed space) X and a linear mapping $T : X \to X$, satisfying*

$$\|Tv_1 - Tv_2\|_X \leq M \|v_1 - v_2\|_X \ , \tag{5.30}$$

for all $v_1, v_2 \in X$ and fixed M, $0 \leq M < 1$, then there exists a unique $u \in X$ such that

$$u = Tu \ , \tag{5.31}$$

i.e. the **contraction mapping** *T has a unique* **fixed point** *u.*

Proof (Contraction Mapping) First we show uniqueness. Suppose $v_1 = Tv_1$ and $v_2 = Tv_2$. Since T is a contraction mapping we have

$$\|v_1 - v_2\|_X = \|Tv_1 - Tv_2\|_X \leq M\|v_1 - v_2\|_X \ . \tag{5.32}$$

So, either $M = 1$ or $\|v_1 - v_2\|_X = 0$. Since $0 \leq M < 1$ $v_1 = v_2$, so the fixed point is unique.

Next we show existence. Take an arbitrary $v_0 \in X$ and define

$$v_1 = Tv_0 \ , \quad v_2 = Tv_1 = T^2 v_0 \ , \quad \ldots \ , \quad v_k = Tv_{k-1} = T^k v_0 \ , \quad \ldots \tag{5.33}$$

Note that $\|v_{k+1} - v_k\| = \|Tv_k - Tv_{k-1}\| \leq M\|v_k - v_{k-1}\| \leq M^k\|v_1 - v_0\|$. So for $k \to \infty$ we have

$$\lim_{k\to\infty} \|v_{k+1} - v_k\| \leq \lim_{k\to\infty} M^k\|v_1 - v_0\| = \|v_1 - v_0\| \lim_{k\to\infty} M^k = 0 \ , \tag{5.34}$$

because $0 \leq M < 1$. Therefore the Cauchy sequence $\{v_k\}$ converges to a limit v and since the space X is complete this limit v will be a member of X, so we have $\lim_{k\to\infty} v_k = v \in X$. The only thing we need to establish now is that this limit point is actually a fixed point for T (and therefore the only fixed point due to uniqueness).

$$\begin{aligned} v &= \lim_{k\to\infty} v_k \\ &= \lim_{k\to\infty} Tv_{k-1} \\ &= T\left(\lim_{k\to\infty} v_{k-1}\right) \\ &= Tv \ , \end{aligned} \tag{5.35}$$

where interchanging the operator and the limit is allowed in the third line, because T is continuous. So the limit value v is indeed the fixed point. **End proof**

With this lemma we are able to prove the Lax-Milgram Theorem.

Proof (Lax-Milgram Theorem) For any $u \in X$, define a functional $Au \in X'$ by $< Au, v >= a(u, v)$ for all $v \in X$. Au is linear, because for arbitrary $v_1, v_2 \in X$ and scalars α_1 and α_2 we have

$$\begin{aligned} < A, \alpha_1 v_1 + \alpha_2 v_2 > &= a(u, \alpha_1 v_1 + \alpha_2 v_2) \\ &= \alpha_1 a(u, v_1) + \alpha_2 a(u, v_2) \\ &= \alpha_1 < Au, v_1 > + \alpha_2 < Au, v_2 > \end{aligned} \tag{5.36}$$

Au is also continuous, because for all $v \in X$ we have

$$| < Au, v > | = |a(u,v)| \leq \|a\| \|u\|_X \|v\|_X \ . \tag{5.37}$$

Therefore,

$$\|Au\|'_X = \sup_{v \neq 0} \frac{| < Au, v > |}{\|v\|_X} \leq \|a\| \|u\|_X \ . \tag{5.38}$$

Thus $Au \in X'$. So the weak formulation can now be reformulated as: Find $u \in X$ such that

$$< Au, v > = < F, v > \quad \forall v \in X \ . \tag{5.39}$$

In other words, we want to find the unique u such that

$$Au = F \quad \text{in } X' \ . \tag{5.40}$$

One can show then that the mapping $u \to Au$ is a linear map from $X \to X'$. Now Riesz's Representation Theorem (see section 5.5.3) states that for every $\phi \in X'$ there exists a unique element $\tau\phi \in X$ such that $< \phi, v > = (\tau\phi, v)_X$ for all $v \in X$. Remember that the square brackets denote duality pairing, whereas the round brackets denote the inner-product. This may be confusing, especially when you know that every element $\phi \in X'$ can be associated with a unique element $\tau\phi \in X$. Anyway, ..., due to the one-to-one correspondence we have

$$\tau Au = \tau F \quad \text{in } X \ . \tag{5.41}$$

We solve this equation using a contraction mapping. We want to find a $\rho \neq 0$ such that the mapping $T : X \to X$ is a contraction, where T is defined by

$$Tv := v - \rho\,(\tau Av - \tau F) \quad \forall v \in X \ . \tag{5.42}$$

If T is a contraction mapping, then there exists a unique element $u \in X$ such that

$$Tu = u - \rho\,(\tau Au - \tau F) = u \ , \tag{5.43}$$

which implies that $\tau Au = \tau F$.

It remains to show that such a $\rho \neq 0$ exists. For any $v_1, v_2 \in X$, let $v = v_1 - v_2$. Then

$$\begin{aligned}
\|Tv_1 - Tv_2\|_X^2 &= \|v_1 - v_2 - \rho\left(\tau Av_1 - \tau Av_2\right)\|_X^2 \\
&= \|v - \rho\left(\tau Av\right)\|_X^2 && (5.44)\\
&= \|v\|_X^2 - 2\rho(\tau Av, v) + \rho^2\|\tau Av\|_X^2 \\
&= \|v\|_X^2 - 2\rho < Av, v > +\rho^2 < Av, \tau Av > && (5.45)\\
&= \|v\|_X^2 - 2\rho a(v, v) + \rho^2 a(v, \tau Av) && (5.46)\\
&\leq \|v\|_X^2 - 2\rho\alpha\|v\|_X^2 + \rho^2\|a\|\|v\|_X\|\tau Au\|_X && (5.47)\\
&\leq \left(1 - 2\rho\alpha + \rho^2\|a\|^2\right)\|v\|_X^2 && (5.48)\\
&= \left(1 - 2\rho\alpha + \rho^2\|a\|^2\right)\|v_1 - v_2\|_X^2 \\
&= M^2\|v_1 - v_2\|_X^2 \ .
\end{aligned}$$

Here α is the coercivity constant of the bilinear form $a(\cdot, \cdot)$. In (5.44) we used the linearity of τ and A. In (5.45) we used the definition of τ. In (5.46) the definition of A, whereas in (5.47) we used coercivity and continuity of $a(\cdot, \cdot)$. In step (5.48) we use the fact that $\|\tau Av\| = \|Av\| \leq \|a\|\|v\|$. In order for T to be a contraction we should chose ρ such that

$$1 - 2\rho\alpha + \rho^2\|a\|^2 < 1 \quad \Longleftrightarrow \quad \rho\left(\rho\|a\|^2 - 2\alpha\right) < 0 \ . \tag{5.49}$$

So taking any $\rho \in (0, 2\alpha/\|a\|)$ yields $M < 1$ and therefore a unique solution. **End Proof**.

Looking back at the proof, all that was needed to establish this result were a lot of indentities and definitions from functional analysis and the fact that the bilinear is bounded and coercive, see (5.47). We can actually show that the solution to our continuous weak formulation exists and is unique, but it also follows that the solution is continuous with respect to the data, since we have $\|u\|_X \leq \|F\|_{X'}/\alpha$, see one of the exercises.

At this stage it is useful to check whether our one-dimensional model problem satifies the conditions for a unique solution. Remember that for our sample problem we had

$$a(u, v) = \int_0^1 \left(u'v' + uv\right) dx = (u, v)_{H^1(0,1)} \tag{5.50}$$

The Cauchy-Schwartz inequality gives us

$$|a(u, v)| = |(u, v)_{H^1(0,1)}| \leq \|u\|_{H^1(0,1)}\|v\|_{H^1(0,1)} \ , \tag{5.51}$$

and therefore the bilinear form is bounded and $\|a\| = 1$. Coercivity means that we should be able to find an α such that for all $v \in H^1(0,1)$ we have

$$\alpha\|v\|_{H^1(0,1)} \leq a(v,v) = \|v\|_{H^1(0,1)} \quad \forall v \in H^1(0,1) \; . \tag{5.52}$$

So taking $\alpha = 1$ does the job. The two required constants exist and therefore our model problem is well-posed and has a unique solution. Let is therefore concentrate on methods to approximate this solution.

5.1.3 Galerkin approximations

Reverting back to the integral form of the weak formulation, our problem can be stated as: find $u \in H_0^1(0,1)$ such that

$$\int_0^1 (u'v' + uv)\, dx = \int_0^1 xv\, dx \quad \forall v \in H_0^1(0,1) \; . \tag{5.53}$$

We now take up the question of determining approximate solutions to this equation and, once again, our approach centers on properties of the class $H_0^1(0,1)$ of admissable functions.

There are two fundamental properties which play a crucial role in the type of approximation we have in mind. First, $H_0^1(0,1)$ is a *linear space* of functions, and second, it is *infinite dimensional*.

By a linear space we simply mean that linear combinations of functions in $H_0^1(0,1)$ are also members of $H_0^1(0,1)$.

By "infinite dimensional" we mean that it is necessary to specify an infinity of parameters in order to define uniquely an arbitrary test function v in the space. The reader with some knowledge of Fourier series will have no difficulty in understanding this concept. Indeed, if we introduce the set of functions

$$\psi_n(x) = \sqrt{2}\sin n\pi x \; , \quad n = 1, 2, 3, \ldots \tag{5.54}$$

and v is a smooth test function in $H_0^1(0,1)$, then it is easily verified that v can be represented in the form

$$v(x) = \sum_{n=1}^{\infty} a_n \psi_n(x) \; , \tag{5.55}$$

where the scalar coefficients are given by

$$a_n = \int_0^1 v(x)\psi_n(x)\,dx \ . \tag{5.56}$$

Thus an infinity of coefficients a_n must be specified in order to define any function $v \in H_0^1(0,1)$. The space of admissable functions is therefore infinite-dimensional.

Let us suppose that we are given an infinite set of functions $\phi_1(x), \phi_2(x), \ldots$ in $H_0^1(0,1)$ which have the property that any $v \in H_0^1(0,1)$ can be represented as a linear combination of these functions, then we call the functions $\phi_i(x)$ the *basis functions*.

It is clear, that if we take a finite number N of terms in the series we will get an approximation v_N of v

$$v_N = \sum_{n=1}^{N} \beta_n \phi_n(x) \ . \tag{5.57}$$

The N basis functions $\{\phi_1, \ldots, \phi_N\}$ define an N-dimensional subspace of $H_0^1(0,1)$ we will denote by H^N for the time being. Thus, instead of tackling the infinite dimensional problem (5.53), we now seek a solution $u_N \in H^N$ of the form

$$u_N(x) = \sum_{n=1}^{N} \alpha_n \phi_n(x) \ , \tag{5.58}$$

which satisfies (5.53) with $H_0^1(0,1)$ replaced by H^N. In other words the variational formulation of the approximate problem is: Find $u_N \in H^N$ such that

$$\int_0^1 \left(u_N' v_N' + u_N v_N\right) dx = \int_0^1 x v_N \, dx \quad \forall v_N \in H^N \ . \tag{5.59}$$

Since the ϕ_i are known, u_N is completely determined by the N coefficients α_i. These coefficients are called the *degrees of freedom* of the approximation. Let us now insert this approximation in our weak formulation. We then get

$$\int_0^1 \left\{ \frac{d}{dx}\left[\sum_{i=1}^{N} \beta_i \phi_i(x)\right] \frac{d}{dx}\left[\sum_{j=1}^{N} \alpha_j \phi_j(x)\right] + \right.$$

$$\left[\sum_{i=1}^{N}\beta_i\phi_i(x)\right]\left[\sum_{j=1}^{N}\alpha_j\phi_j(x)\right]-x\left[\sum_{i=1}^{N}\beta_i\phi_i(x)\right]\Bigg\}\,dx=0\quad\forall\beta_i\ ,\ i=1,\ldots,N\ . \tag{5.60}$$

Rearranging this equation gives

$$\sum_{i=1}^{N}\beta_i\left(\sum_{j=1}^{N}\left\{\int_0^1\left[\phi_i'(x)\phi_j'(x)+\phi_i(x)\phi_j(x)\right]\,dx\right\}\alpha_i-\int_0^1 x\phi_i(x)\,dx\right)=0$$

$$\forall\beta_i\ ,\ i=1,\ldots,N\ . \tag{5.61}$$

This can be written compactly as

$$\sum_{i=1}^{N}\beta_i\left(\sum_{j=1}^{N}K_{ij}\alpha_j-F_i\right)=0\ , \tag{5.62}$$

for all possible choices of β_i, where

$$K_{ij}=\int_0^1\left[\phi_i'(x)\phi_j'(x)+\phi_i(x)\phi_j(x)\right]\,dx \tag{5.63}$$

and

$$F_i=\int_0^1 x\phi_i(x)\,dx\ , \tag{5.64}$$

for $i,j=1,\ldots,N$.

The $N\times N$ rectangular matrix $\boldsymbol{K}$ is usually called the *stiffness matrix* or the *system matrix*. The column vector $\vec{F}$ is simply referred to as the right-hand-side vector or the *load vector*.

We can satisfy (5.62) for all possible choices of β_i by setting

$$\boldsymbol{K}\vec{\alpha}=\vec{F}\quad\Longleftrightarrow\quad\sum_{j=1}^{N}K_{ij}\alpha_j=F_i\quad\text{for}\ \ i=1,\ldots,N\ . \tag{5.65}$$

Note that the truncated expansion of u_N in terms of only a finite number of basis functions, actually leads to the same variational formulation, but now in a finite dimensional subspace of $H_0^1(0,1)$, i.e. $H^N \subset H_0^1(0,1)$. This means that all the properties that hold for functions in $H_0^1(0,1)$ also hold for functions in H^N. In particular, when the bilinear form $a(\cdot,\cdot)$ is bounded and coercive for all functions in the space $H_0^1(0,1)$, it will also be bounded and coercive for all functions in the space H^N, with the same cosntants α and $\|a\|$. This measn that Lax-Milgram Theorem is also applicable to the linear subspace H^N and therefore (5.65) has a unique solution. The fact that $H^N \subset H_0^1(0,1)$ also allows us to say something about the error between the approximate solution u_N obtained by Galerkin's method and the exact solution. We do this by the following theorem.

Theorem 4 *(***Céa's Theorem)** *Suppose X is a Hilbert space and $a(\cdot,\cdot)$ is a bilinear form which is bounded and coercive on X. Let $X^h \subset X$ and let u be the solution of the infinite dimensional problem in X and u_N the Galerkin approximation in the finite dimensional subspace X^h, then we have*

$$\|u - u_N\|_X \leq \frac{\|a\|}{\alpha} \inf_{v \in X^h} \|u - v\|_X \ , \tag{5.66}$$

where $\|a\|$ is the norm of the bilinear form and α is the coercivity constant.

Proof (Céa's Theorem) Since $a(u,v) =< F, v >$ for all $v \in X$ and $a(u_N, v) =< F, v >$ for all $v \in X^h$ we have (by subtracting)

$$a(u - u_N, v) = 0 \quad \forall v \in X^h \ . \tag{5.67}$$

This important relation, essentially states that the error is orthogonal to all the elements in the subspace X^h with respect to the bilinear form $a(\cdot,\cdot)$. The equality (5.67) is called *Galerkin orthogonality.*

So for all $v \in X^h$ we have

$$\begin{aligned}
\alpha\|u - u_N\|^2 &\leq a(u - u_N, u - u_N) && (5.68)\\
&= a(u - u_N, u - v) + a(u - u_N, v - u_N) && (5.69)\\
&= a(u - u_N, u - v) && (5.70)\\
&= \|a\|\|u - u_N\|\|u - v\| \ . && (5.71)
\end{aligned}$$

In (5.68) we have used the fact that the bilinear form is coercive, in (5.69) we used the linearity of the bilinear form. In (5.70) note that $v - u_N \in X^h$ so we can use Galerkin

orthogonality, whereas in the last step (5.71) continuity of the bilinear form was used. Hence we have

$$\|u - u_N\|_X \leq \frac{\|a\|}{\alpha}\|u - v\|_X \quad \forall v \in X^h \ . \tag{5.72}$$

If this realtion holds for all $v \in X^h$ it also holds for that particular v which minimizes the right-hand-side, so

$$\begin{aligned} \|u - u_N\|_X &\leq \frac{\|a\|}{\alpha} \inf_{v \in X^h} \|u - v\|_X \\ &= \frac{\|a\|}{\alpha} \min_{v \in X^h} \|u - v\|_X \ , \end{aligned}$$

since X^h is a closed finite dimensional subspace.

Céa's Theorem shows that u_N is *quasi-optimal* in the sense that the error $\|u - u_N\|_X$ is proportional to the best it can be using the subspace X^h.

Integration by parts has allowed us to set up the current weak formulation and its Galerkin approximation. This weak form has several advantages

- The system has become symmetric. Interchanging the role of u and v on the continuous level does not alter the weak form. The discrete formulation obtained by approximating the solution by a truncated expansion also leads to a symmetric matrix $\boldsymbol{K}$. This allows us to only store half of the matrix and allows us to use numerical solvers designed to solve symmetric systems.
- Because the weak formulation and its approximation are symmetric we can use the same function space for u en v; hence one set of basis functions ϕ_i need to be constructed for such approximations.

The quality of the approximate solution is completely determined by the choice of the basis functions ϕ_i. Once these basis functions have been chosen, the determination of the approximate solution, i.e. the calculation of the coefficients α_i, reduces to a computational matter.

So it is important to chose the right basis functions. How do we chose these functions? Of course, we could chose polynomials with increasing polynomial degree in order to get better approximations. However the system will become ill-conditioned in that case. A better way to chose the basis functions is find functions which already possess as many

of the characteristics as the exact solution. Since we don't know the exact solution it is not straightforward how to do this. One could, for instance, take the first few eigenvectors of the differential equation, but one needs additional weighting factors in these eigenvectors to produce a good basis. Another way of chosing the basis vectors is to look at functions which are invariant under the operation of the differential equation. The exact solution will possess the same symmetries as the those invariant functions and therefore an expansions in terms of these invariants may prove to be a good method. These invariants may be found by means of Lie group theory, which is beyond the scope of these lectures.

Instead of persuing the search for good basis functions, we can also try to find convenient basis functions. Note that in setting up out discrete system we have to evaluate a large number of integrals. It would be nice if we knew from the ouset that a large number of the integrals would be zero, so we don't have to spend time evaluating these integrals. This will be case if the basis functions are local. If we define the support of a basis function by

$$\mathrm{supp}(\phi_i) = \{x \in (0,1) \mid \phi_i(x) \neq 0\} \ . \tag{5.73}$$

The idea of a local support means that $\mathrm{supp}(\phi_i)$ does not cover the whole domain and therefore two basis functions ϕ_i and ϕ_j with $\mathrm{supp}(\phi_i) \cap \mathrm{supp}(\phi_j) = \emptyset$ do not have to be taken into account in seting up the stiffness matrix. This is the idea which led to the finite element method

5.2 Finite Element Basis Functions

Apart from the fact that local basis functions decrease the effort of setting up the stiffness matrix, it is also convenient when we want to apply the Galerkin method in complex multi-dimensional problems.

Now that you are all convinced that basis functions with a local support are very handy, we will actually construct such basis functions. We do this by partitioning the interval $(0,1)$ into smaller, non-overlapping intervals, called the *elements*. Such a partitioning is shown in Fig. 5.1.

The nodes are given by $0 = x_0 < x_1 < \cdots < x_{K-1} < x_K = 1$ and the sub-intervals $\Omega_i = (x_i, x_{i-1})$ are called the finite elements. In Fig. 5.1 $K = 5$ and all elements have the same length $|\Omega_i| = x_i - x_{i-1} = h$. This does not have to be the case. The collection of nodes and elements is called the *finite element mesh*. The way to construct these meshes will be the topic of Chapter 7.

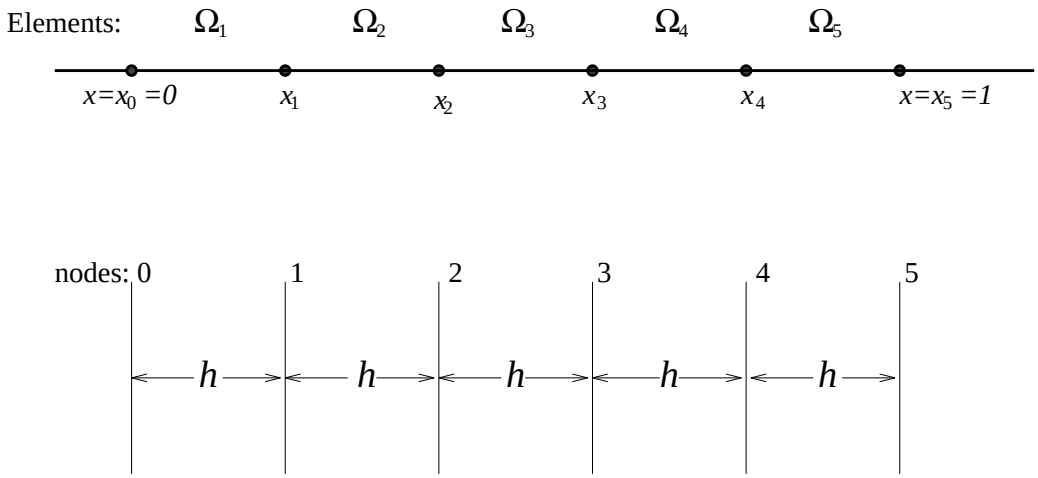

Figure 5.1: A finite element partitioning of the region $0 \leq x \leq 1$ comprising of five elements with the nodes at the element end points

Having constructed a finite element mesh for our model problem, we proceed to construct a corresponding set of basis functions using the following fundamental criteria.

- The basis functions are generated by simple functions defined peicewise - element by element - over the finite element mesh.
- The basis functions are smooth enough to be members of the class $H_0^1(0,1)$ of test functions.
- The basis functions are chosen in such a way that the parameters α_i coincide with value of the approximating function at the nodal points, i.e. $\alpha_i = u_N(x_i)$.

The first condition leads to efficient codes. If all the elements are treated in the same way, it is easy to write a loop over all elements. The second condition is practical, because our weak formulation only made sense for function $u, v \in H_0^1(0,1)$. If this condition is satisfied we call the approximation a *conforming approximation* and one speaks of an *interior approximation*. In case this condition is not satisfied we usually speak of a *non-conforming* or *exterior approximation*. Interior refers to the fact the we chose a subspace H^N of $H_0^1(0,1)$ in the first case, i.e. a function space within the space $H_0^1(0,1)$, whereas exterior means that we use functions that are not in $H_0^1(0,1)$ in the approximation of our problem. Some people refer to this latter type of approximating as *a variational crime* and since crime pays, these methods have become very popular. However, we will restrict ourselves in this introductory chapter to law-abiding methods, i.e. conforming methods. The last requirement is one of convenience. If the degrees of freedom of your problem coincide with the nodal values of your approximate solution, there is no need for the reconstruction of the approximating function.

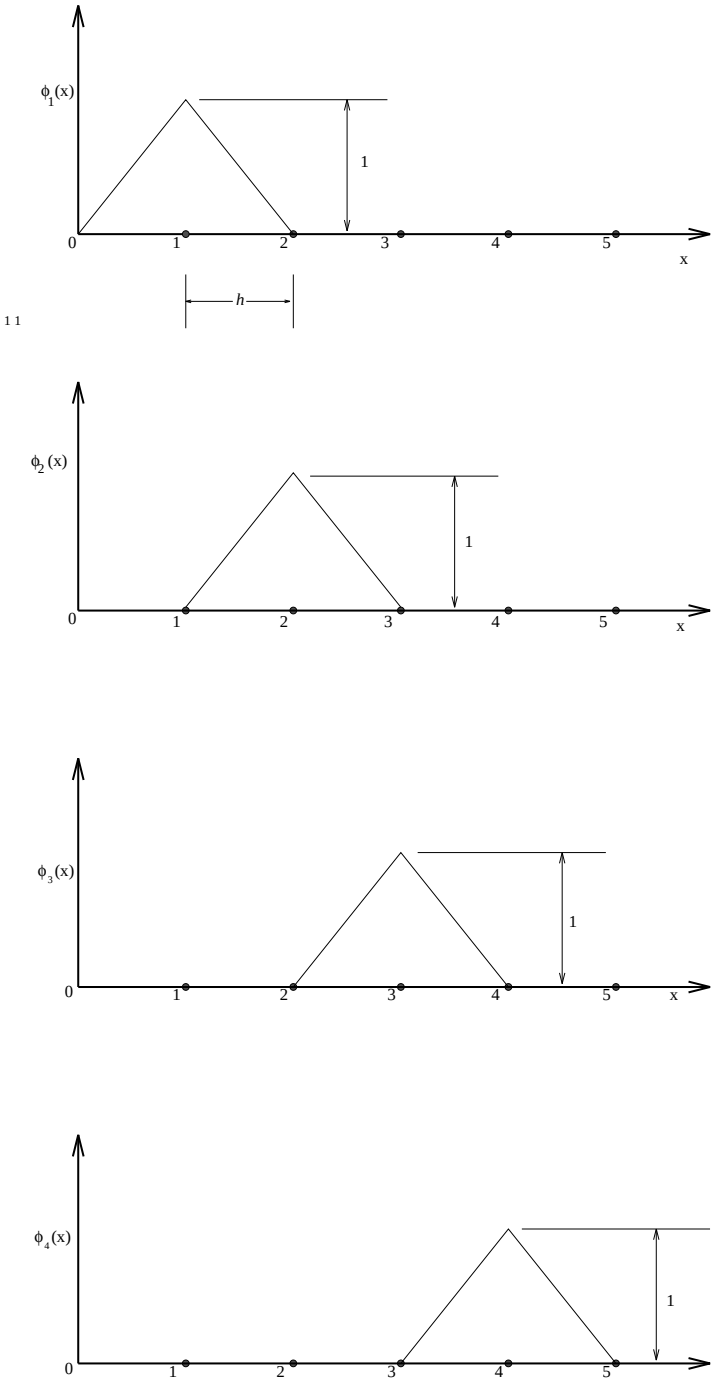

Figure 5.2: A finite element partitioning of the region $0 \leq x \leq 1$ comprising of five elements with the nodes at the element end points

One very simple, yet perfectly adequate, set of basis functions satisfying these three criteria is shown in Fig. 5.2. The basis functions for $i = 1, 2, 3$ and 4 are given by

$$\phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h_i} & \text{for } x_{i-1} \leq x \leq x_i \\ \frac{x_{i+1} - x}{h_{i+1}} & \text{for } x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{5.74}$$

The derivatives of these functions are given by

$$\phi_i'(x) = \begin{cases} \frac{1}{h_i} & \text{for } x_{i-1} \leq x \leq x_i \\ \frac{-1}{h_{i+1}} & \text{for } x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{5.75}$$

In order to demonstrate that these functions satisfy the three criteria mentioned above, first observe that the functions are continuous and therefore elements of $H_0^1(0,1)$ (see Exercise 28), so the second criterion is satisfied. Criterion 1 mentions that the basis function are generated in an element-by- element fashion. In order to do this we look at an arbitrary element Ω_i. We map this element $\Omega_i = (x_i, x_{i-1})$ onto a standard element $\Omega^P = (0,1)$, called the *parent element*. The mapping from the standard element Ω^P to our arbitrary element Ω_i is given by

$$x = (1 - \xi^i) * x_{i-1} + \xi^i * x_i \quad \text{for } 0 \leq \xi^i \leq 1 \, . \tag{5.76}$$

Here the superscript i refers to the fact that this mapping depends on the element Ω_i under consideration. So $\xi^i = 0$ is mapped onto x_{i-1} and $\xi^i = 1$ is mapped onto x_i. On this parent element we define two linear functions

$$\psi_I^P(\xi^i) = 1 - \xi^i \quad \text{and} \quad \psi_{II}^P(\xi^i) = \xi^i \, . \tag{5.77}$$

The parent element and the two linear basis functions are depicted in Fig. 5.3. If we map the parent element and the associated functions back to the computational domain and glue all elements together we retrieve the four basis functions shown in Fig. 5.2 plus to linear basis functions at the end of our computational domain. The contribution of the last two basis functions will be eliminated by incorporating the boundary conditions as we will see shortly.

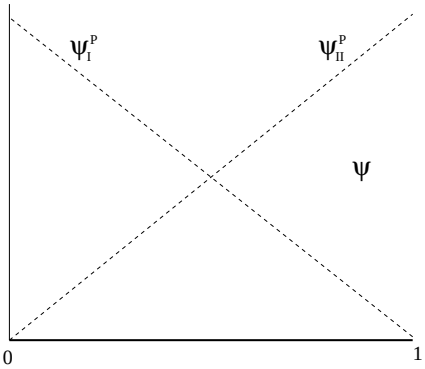

Figure 5.3: The two linear basis functions defined on the parent element

If we go back to our stiffness matrix $\boldsymbol{K}$ we have that

$$\begin{aligned} K_{ij} &= \int_0^1 \left(\phi_i'\phi_j' + \phi_i\phi_j\right)\, dx \\ &= \sum_{k=1}^{K} \int_{\Omega_k} \left(\frac{d\phi_i}{dx}\frac{d\phi_j}{dx} + \phi_i(x)\phi_j(x)\right)\, dx \\ &= \sum_{k=1}^{K} \int_{\Omega^P} \left(\frac{d\phi_i}{d\xi^k}\frac{d\xi^k}{dx}\frac{d\phi_j}{d\xi^k}\frac{d\xi^k}{dx} + \phi_i(\xi^k)\phi_j(\xi^k)\right)\frac{dx}{d\xi^k}\, d\xi^k \ . \end{aligned} \tag{5.78}$$

Now the mapping $\xi^k : \Omega^P \to \Omega_k$ is explicitly known and therefore also the derivatives $d\xi^k/dx$ and $dx/d\xi^k$. Furthermore, the calculation of the stiffness matrix is now decomposed in into K calculations of stiffness matrices on the parent element in which the two basis functions are explicitly known as a function of ξ.

As an example, let us calculate the element matrices, using the recipe described above. Note that since we use elements of the same size $h = 1/K$, where K is the number of elements, we have that

$$\frac{dx}{d\xi^k} = x_k - x_{k-1} = h \quad \text{therefore} \quad \frac{d\xi^k}{dx} = \frac{1}{x_k - x_{k-1}} = \frac{1}{h}\ . \tag{5.79}$$

Furthermore, for the derivatives of the basis functions ψ^P on the parent element we have

$$\frac{d\psi_I^P}{d\xi} = -1 \quad \text{and} \quad \frac{d\psi_{II}^P}{d\xi} = 1\ . \tag{5.80}$$

So K_{ij}^k, which is the *element stiffness matrix* is a 2×2 matrix with the following entries

$$K_{11}^k = \int_0^1 \left((-1)\frac{1}{h}(-1)\frac{1}{h} + (1-\xi)^2 \right) h \, d\xi = \frac{1}{h} + \frac{h}{3} \, . \tag{5.81}$$

$$K_{12}^k = \int_0^1 \left((-1)\frac{1}{h}\frac{1}{h} + (1-\xi)\xi \right) h \, d\xi = -\frac{1}{h} + \frac{h}{6} \, , \tag{5.82}$$

and

$$K_{22}^k = K_{11}^k = \frac{1}{h} + \frac{h}{3} \quad \text{and} \quad K_{21}^k = K_{12}^k = -\frac{1}{h} + \frac{h}{6} \, . \tag{5.83}$$

So in this specific case all the element stiffnes matrices are the same. This will not be the case on a non-uniform mesh or when one wants to approximate a differential equation with variable coefficients.

Similarly, we can transform the calculation of the right-hand-side vector to an operation on the parent element, using

$$\int_0^1 x\phi_i \, dx = \sum_{k=1}^{K} \int_{\Omega_k} x\phi_i \, dx = \sum_{k=1}^{K} \int_{\Omega^P} x(\xi)\psi_i^P \frac{dx}{d\xi} \, d\xi \, . \tag{5.84}$$

Since we have defined only two basis functions on our parent domain so-called *element stiffness matrix* is a 2×1 vector whose components are given by

$$K_1^k = \int_0^1 ((1-\xi) * x_{k-1} + \xi * x_k)(1-\xi) h \, d\xi = \frac{h}{6} \left(2x_{k-1} + x_k \right) \, , \tag{5.85}$$

and

$$K_2^k = \int_0^1 ((1-\xi) * x_{k-1} + \xi * x_k)\xi h \, d\xi = \frac{h}{6} \left(x_{k-1} + 2x_k \right) \, . \tag{5.86}$$

So at the element level we have now

$$\boldsymbol{K}^k = \begin{bmatrix} \frac{1}{h} + \frac{h}{3} & -\frac{1}{h} + \frac{h}{6} \\ -\frac{1}{h} + \frac{h}{6} & \frac{1}{h} + \frac{h}{3} \end{bmatrix} \, , \quad \vec{F}^k = \frac{h}{6} \begin{pmatrix} 2x_{k-1} + x_k \\ x_{k-1} + 2x_k \end{pmatrix} \, . \tag{5.87}$$

Note that these element matrices and and element load vectors can be set up independently of all the other elements. So setting up the discrete system can be efficiently established on a parallel machine.

The only question we now have, is how the small element matrices and load vectors fit into the global stiffness matrix and the global load vector. The operation of placing the small element matrices $\boldsymbol{K}^k$ into the matrix $\boldsymbol{K}$ is called *assembly*. This can be efficiently established with a map which relates the local nodes on the parent element to the global node numbers in the mesh. Each local node is determined by the element number k and the degree of freedom with the parent domain (in our example 2). If we set up a array or a look-up table which connects these local degrees of freedom to the global degrees of freedom the global matrices follow very quickly from the local matrices. Such a look-up table has the following form

$$GM(elementnr, nodenr) = globalnr \ . \tag{5.88}$$

In our example depicted in Fig. 5.2 we have 6 global node numbers (numbered 0 until 5), 5 elements containing 2 unknowns per element, so this look-up table has the following form

$$\begin{aligned} GM(1,1) &= 0 \\ GM(1,2) &= 1 \\ GM(2,1) &= 1 \\ GM(2,2) &= 2 \\ GN(3,1) &= 2 \\ GM(3,2) &= 3 \\ GM(4,1) &= 3 \\ GM(4,2) &= 4 \\ GM(5,1) &= 4 \\ GM(5,2) &= 5 \end{aligned} \tag{5.89}$$

Now the whole exercise of setting up the global stiffnes matrix and the global load vector can be written as

```
K=0.0
F=0.0
Do k=1,K
  Do i=1,2
    Do j=1,2
      K(GM(k,i),GM(k,j)) = K(GM(k,i),GM(k,j)) + K^k(i,j)
```

```
    Enddo
    F(GM(k,i)) = F(GM(k,i)) + F^k(i)
  Enddo
Enddo
```

It is quite easy to see that this procedure places the element matrices at the right position in the global matrix and vector. It also has the advantage, that if you want to change the global node numbering, you only need to modify the GM-matrix.

Using the fact that $K = 5$ we have $h = 0.2$ and therefore the element matrices and element load vectors become

$$\boldsymbol{K}^k = \frac{1}{30}\begin{bmatrix} 152 & -149 \\ -149 & 152 \end{bmatrix}, \quad \vec{F}^k = \frac{1}{30}\begin{pmatrix} 2x_{k-1} + x_k \\ x_{k-1} + 2x_k \end{pmatrix}. \tag{5.90}$$

So the global system becomes

$$\frac{1}{30}\begin{pmatrix} 152 & -149 & 0 & 0 & 0 & 0 \\ -149 & 304 & -149 & 0 & 0 & 0 \\ 0 & -149 & 304 & -149 & 0 & 0 \\ 0 & 0 & -149 & 304 & -149 & 0 \\ 0 & 0 & 0 & -149 & 304 & -149 \\ 0 & 0 & 0 & 0 & -149 & 152 \end{pmatrix}\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} = \frac{1}{150}\begin{pmatrix} 1 \\ 6 \\ 12 \\ 18 \\ 24 \\ 14 \end{pmatrix}. \tag{5.91}$$

Now the values α_0 and α_5 correspond to the prescribed values at the end points we have chosen to be zero. We can eliminate these values and the correspondig rows in the system (i.e. setting β_0 and β_5 also equal to zero) and we are left with

$$\frac{1}{30}\begin{pmatrix} 304 & -149 & 0 & 0 \\ -149 & 304 & -149 & 0 \\ 0 & -149 & 304 & -149 \\ 0 & 0 & -149 & 304 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \frac{1}{150}\begin{pmatrix} 6 \\ 12 \\ 18 \\ 24 \end{pmatrix}. \tag{5.92}$$

The solution of this linear system is given by

$$\vec{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 0.02877 \\ 0.05064 \\ 0.05844 \\ 0.04443 \end{pmatrix}. \tag{5.93}$$

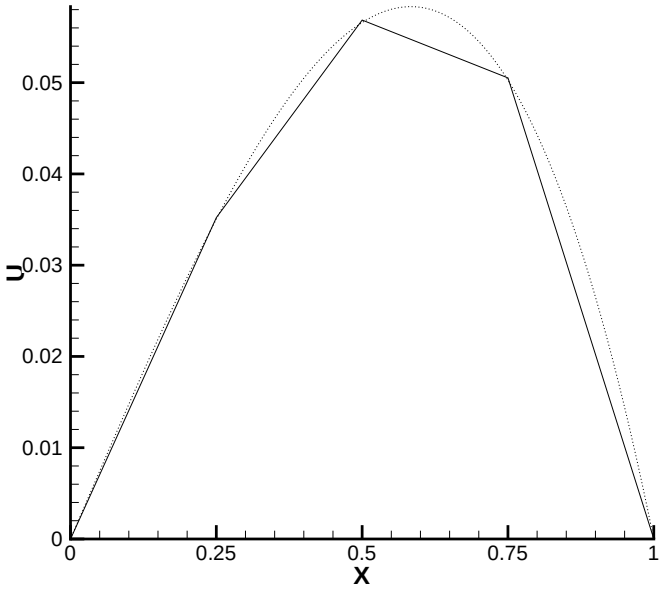

Figure 5.4: The finite element approximation (solid line) and the exact solution (dotted line) of the one-dimensional model problem

The numerical approximation versus the exact solution is depicted in Fig. 5.4. If we increase the number of elements, i.e. if we take a larger finite dimensional subspace of $H_0^1(0,1)$ we expect improved approximations. The results for various values of N ranging from $N = 6$ to $N = 25$ are shown in Fig. 5.5.

Also shown in Fig. 5.5 is the exact solution (dotted line). As is to be expected of a good numerical approximation, increasing the number of elements leads to an improved approximation to the exact solution. However, mere inspection of a figure is insufficient to judge a numerical scheme. Firstly, in the cases we are interested in we do not have the exact solution to compare our numerical scheme with (although in the development phase of numerical schemes, we always chose test cases where exact solution are available) and secondly there is the question how fast the approximate solution converges to the exact solution as a function of the number of elements. In order to analyze the speed of convergence we have to have a way of measuring the difference between the approximate solution and the exact solution. This measuring can be done in a lot of ways, for instance, we can chose the value at a point in our domain to measure the difference between the approximate value at that particular point and the exact solution at that point. We can also chose a functional of the solution, for instance the lift on a airfoil and see how the lift obtained from our approximate solution converges to the true value of the lift. Likewise, we can take the drag to measure the convergence. What you take to measure the quality and accuracy of your numerical scheme highly depends on your application. We will come back to this after we have looked at measures which are considerd 'natural' in the finite

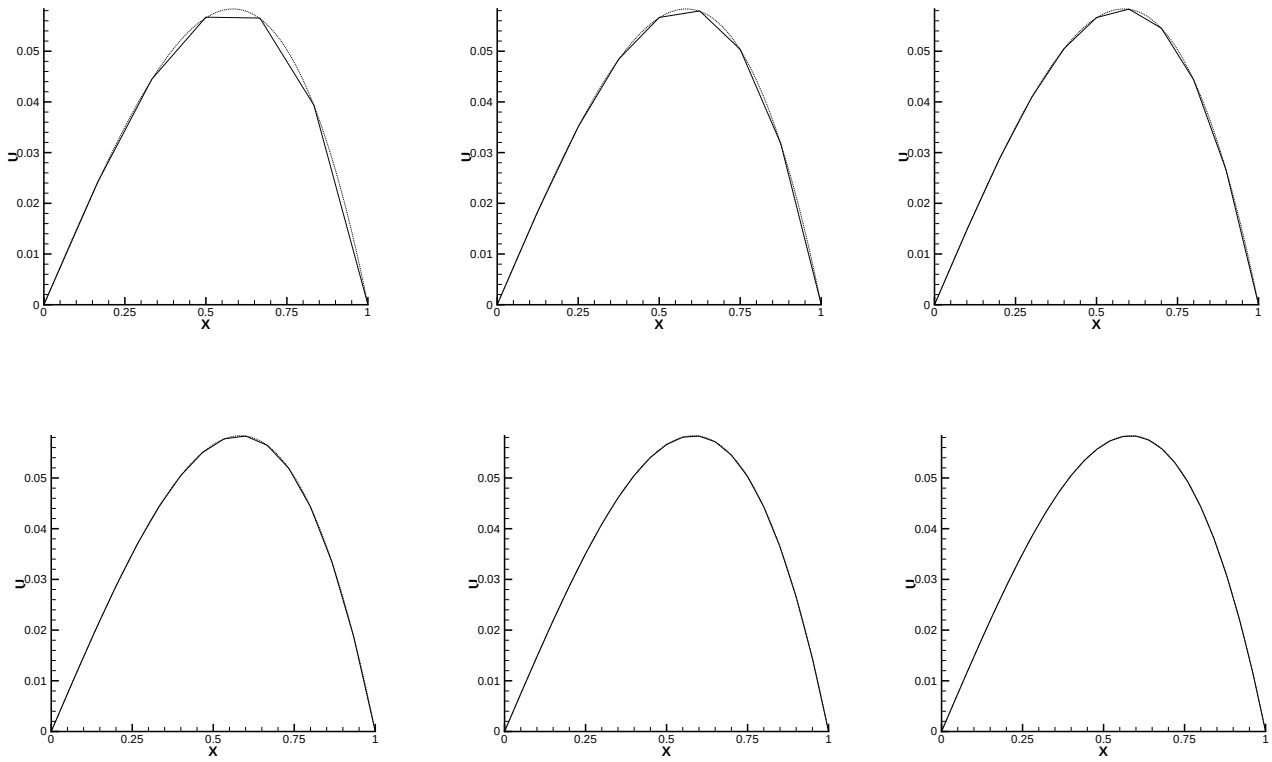

Figure 5.5: Finite element approximation for $K = 6, 8, 10, 15, 20$ and $K = 25$, respectively

element context. Such a 'natural' norm is for instance the L^2-norm of the error between the approximate solution and the exact solution. This error is defined by

$$\|e_N\|_0 := \left\{ \int_\Omega |u_{ex} - u_{appr}|^2 \, dx \right\}^{\frac{1}{2}} . \tag{5.94}$$

Since we know the exact solution for our one dimensional model problem we can explicitly calculate this error for various values of the number of elements. The results are tabulated in the follwing table.

N	h	$\|e\|_0$	$\ln h$	$\ln\|e\|_0$
4	$2.500 \cdot (-1)$	$2.930 \cdot (-3)$	-1.386	-5.833
6	$1.667 \cdot (-1)$	$1.307 \cdot (-3)$	-1.792	-6.640
8	$1.250 \cdot (-1)$	$7.363 \cdot (-4)$	-2.079	-7.214
10	$1.000 \cdot (-1)$	$4.716 \cdot (-4)$	-2.303	-7.659
15	$6.667 \cdot (-2)$	$2.097 \cdot (-4)$	-2.708	-8.470
20	$5.000 \cdot (-2)$	$1.180 \cdot (-4)$	-2.996	-9.045
25	$4.000 \cdot (-2)$	$7.552 \cdot (-5)$	-3.219	-9.491

As was expected the L^2-norm of the error decreases with increasing number of elements K. The question now is how does the error decrease with increasing number of elements. In order to answer this question we anticipate that the error is related to the length of the elements as

$$\|e_N\|_0 = Ch^p \ , \tag{5.95}$$

where C is constant independent of h. Consistency requires that $\|e_N\| \to 0$ if $h \to 0$ which is the case with this error formula. p is called the order of the scheme, so if $p = 2$ we call the scheme second order accurate. From the error formula we deduce that

$$\ln\|e_N\|_0 = p \ln h + \ln C \ . \tag{5.96}$$

So if p does not depend on h we expect a straight line when we plot $\ln\|e_N\|_0$ versus $\ln h$. This graph is depicted in Fig. 5.6.

Fig. 5.6 demonstrates that the use of linear elements in our one dimensional finite element program leads to a second order scheme. In many cases, however, we are not only interested what the difference is between the exact solution and the approximate solution, but we want to know how well the derivative of the approximate solution approximates the derivative of the exact solution. This is particularly the case if we are interested in derived quatities such as the stresses and the vorticity (which are derivatives of the velocity field). In Fig. 5.7 the derivative of the exact solution and the derivative of the approximate solution are plotted for 4 elements.

Note that since the approximate solution consists of piecewise linear functions, the derivative of the approximate solution will consist of piecewise constant solutions. This gives a rather crude approximation to the derivative, but if we increase the number of elements the approximate staircase functions will resemble the derivative of the exact solution better and better. Fig. 5.8 displays the approximation to the derivative of the exact solution for an increasing number of elements.

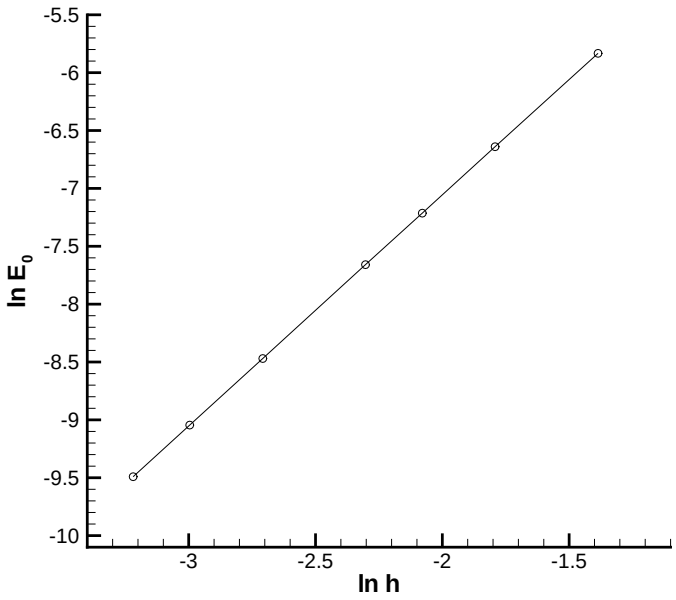

Figure 5.6: Log-log plot of the L^2 norm of the error as a function of h for the model problem

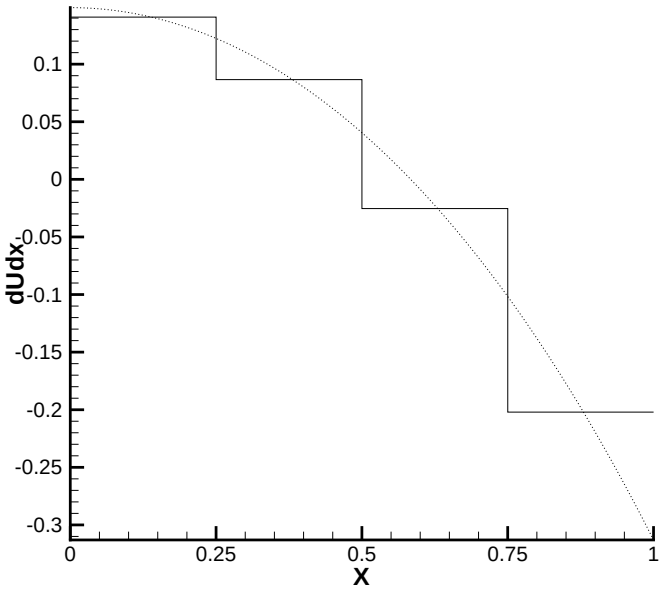

Figure 5.7: The derivative of the exact and the approximate solution for 4 elements

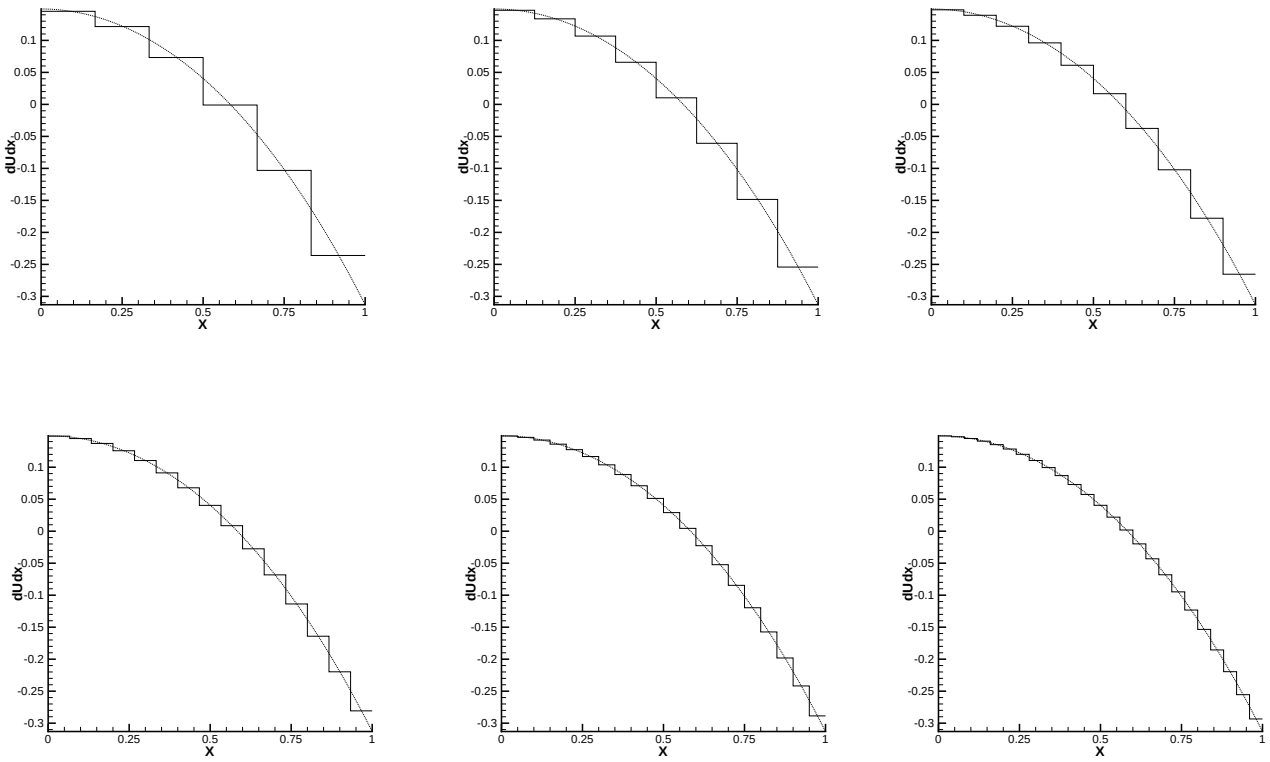

Figure 5.8: Finite element approximation of the derivative for $K = 6, 8, 10, 15, 20$ and $K = 25$, respectively

As in the case of the function itself, we can ask ourselves how the approximation of the derivative converges to the exact solution. The next table gives the L^2-norm of the error between du_{ex}/dx and du_N/dx for the values of K depicted in in Fig. 5.8.

K	h	$\|e'\|_0$	$\ln h$	$\ln\|e'\|_0$
4	$2.500 \cdot (-1)$	$3.885 \cdot (-2)$	-1.386	-3.248
6	$1.667 \cdot (-1)$	$2.601 \cdot (-2)$	-1.792	-3.649
8	$1.250 \cdot (-1)$	$1.954 \cdot (-2)$	-2.079	-3.935
10	$1.000 \cdot (-1)$	$1.565 \cdot (-2)$	-2.303	-4.158
15	$6.667 \cdot (-2)$	$1.044 \cdot (-2)$	-2.708	-4.562
20	$5.000 \cdot (-2)$	$7.830 \cdot (-3)$	-2.996	-4.850
25	$4.000 \cdot (-2)$	$6.265 \cdot (-3)$	-3.219	-5.073

Again we assume that there is some relation between the error in the derivative and

the element width h as given by (5.95), prompting us to plot the logarithm of the error versus the logarithm of h.

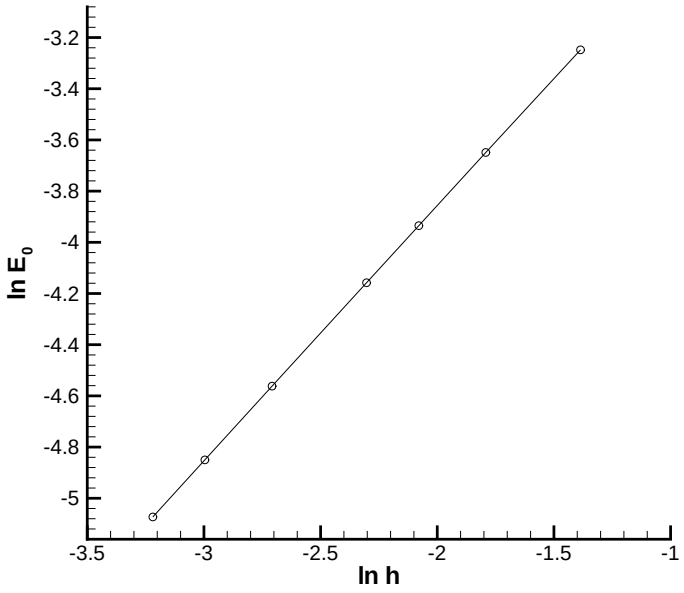

Figure 5.9: Log-log plot of the L^2 norm of the error in the derivative as a function of h for the model problem

This dependence is shown in Fig. 5.9. It turns out that, although the scheme was second order accurate with respect to the function itself, the derivative of the approximate solution only converges order one to the derivative of the exact solution. So the approximation of the derivative is one order less accurate than the approximation of the solution itself. Continuing in this way, we readily see that the approximation of the second derivative does not converge at all. In fact, the second derivative of our piecewise linear interpolation functions all reduce to delta distributions at the node points x_i, where the coefficient of the delta distributions denote the jump in the derivative of the approximate solution between adjacent elements. So if one is interested in derived quantities which contain the second order derivatives of the approximate solution, piecewise linear elements do not suffice.

The norm $\|e'\|_0$ is usually denoted by $|e|_1$. e represents the error in the quantity of interest and the subscript 1, means that we consider the first derivative. This quantity is not a norm for e, because $|e|_1 = 0$ does not necessarily imply that $e = 0$, therefore $|\cdot|_1$ is called a *semi-norm*.

5.3 Error analyis

In the previous section we have seen that piecewise linear elements are second order accurate with respect to the solution and first order accurate with respect to the derivative. Sofar, this could only be established by explicitly measuring the difference between the approximate solution and the exact solution, whereas in many practical problems we do not know what the exact solution is (this is the reason one wants to use CFD). So we have to have a different method in order to assess the accuracy of the scheme.

Before diving into the error analysis let us define a few function spaces. Let $L^p(\Omega)$ consist of all functions which satisfy

$$\|f\|_{L^p(\Omega)} := \left\{ \int_\Omega |f(x)|^p \, d\Omega \right\}^{\frac{1}{p}} < \infty \, . \tag{5.97}$$

When $p = 2$ we retrieve our familiar L^2 space, which we have encountered before. The main difference between the space $L^p(\Omega)$ and a general $L^p(\Omega)$, $p \neq 2$, is that only L^2 is a Hilbert space, i.e. a space in which an inner product can be defined. For general p, $p \neq 2$ this is not possible.

Just as we did with derivatives in the L^2-norm we can also define Sobolev spaces for L^p norms, denoted by $W^{m,p}(\Omega)$, by requiring that

$$f \in W^{m,p}(\Omega) \quad \Longleftrightarrow \quad \int_\Omega |D^k f|^p \, d\Omega < \infty \, , \;\; 0 \leq k \leq m \, , \tag{5.98}$$

in which $D^k f$ denotes all possible partial derivatives up to order k. Note that this definition is not restricted to one spatial dimension. Taking $m = 1$ and $p = 2$ gives the Sobolev space $W^{1,2}(\Omega) = H^1(\Omega)$ which we have seen before.

Now let u be a function such that $u \in W^{m,p}(\Omega)$ and construct a finite element space in the way we did in the previous section. Let us call this space $S^h \subset W^{m,p}(\Omega)$. Now let us define the interpolation operator

$$\Pi_h \, : \, W^{m,p}(\Omega) \rightarrow S^h(\Omega) \, , \tag{5.99}$$

such that

$$\Pi_h u = \sum_{i=1}^{M} l_i(u) \phi_i \, , \tag{5.100}$$

where $\{\phi_i\}_{i=1}^M$ are the global basis functions for the space $S^h(\Omega)$ generated by the finite element method and $\{l_i\}_{i=1}^M$ are the global degrees of freedom. What we are interested in is the quality of the approximation $\Pi_h u$ of u and its behavior as the mesh is refined. The key to this issue lies in the character of the local interpolation operators $\Pi_e = \Pi_h|_e$, i.e. the restriction of the interpolation operator to a finite element. To set the stage for this analysis we need some ingredients, the first one being a purely geometric result.

Theorem 5 *Let Ω and $\hat{\Omega}$ be two affine equivalent domains; that is*

$$\forall x \in \Omega \ , \quad x = \boldsymbol{T}\hat{x} + c \ , \quad \hat{x} \in \hat{\Omega} \ , \tag{5.101}$$

where $\boldsymbol{T}$ is an invertible matrix and c is a translation vector. Further, let

$$\left.\begin{array}{ll} h = dia(\Omega) & \hat{h} = dia(\hat{\Omega}) \\ \rho = \sup\{dia(S); & S \text{ is a sphere contained in } \Omega\} \\ \hat{\rho} = \sup\left\{dia(\hat{S}); & \hat{S} \text{ is a sphere contained in } \hat{\Omega}\right\} \end{array}\right\} \tag{5.102}$$

Then

$$\|\boldsymbol{T}\| \leq \frac{h}{\hat{\rho}} \quad \text{and} \quad \|\boldsymbol{T}^{-1}\| \leq \frac{\hat{h}}{\rho} \ . \tag{5.103}$$

Before proving this result it is worthwhile to look at a geometrical interpretation of the variables involved. Fig. 5.10 shows the two coordinate systems (x_1, x_2) and $(\hat{x}_1, \hat{x}_2)$ and the affine transformation between. ρ and $\hat{\rho}$ denote the radius of the largest circles contained in the domains whereas h and $\hat{h}$ denotes the radius of the smallest circles which enclose the domains.

Note that we have encountered such a mapping when we mapped our finite element to the parent element

$$(0,1) \to (x_{i-1}, x_i) \ , \quad x = (1-\xi)x_{i-1} + \xi x_i = x_{i-1} + \xi(x_i - x_{i-1}) \ . \tag{5.104}$$

where

$$\boldsymbol{T} : (0,1) \to (x_{i-1}, x_i) \ , \quad \boldsymbol{T}(\xi) = \xi\,(x_i - x_{i-1}) \ , \tag{5.105}$$

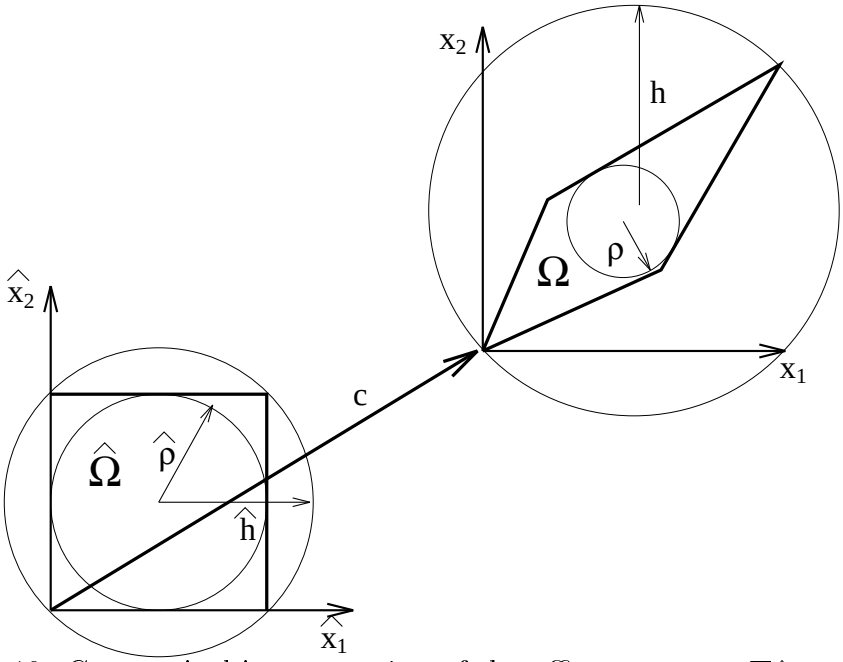

Figure 5.10: Geometrical interpretation of the affine map $x = \boldsymbol{T}\hat{x} + c$ and the definition of the various radii.

and

$$c = x_{i-1} \; . \tag{5.106}$$

For the one-dimensional case $\rho = h$ and $r\hat{h}o = \hat{h}$. Now that we have seen what the theorem geometrically means and that we have already encountered such a mapping it is useful to proof the theorem.

Proof For any constant $\lambda > 0$ we have

$$\|T\| = \sup_{\hat{x}\in\hat{\Omega}} \frac{\boldsymbol{T}\hat{x}}{\hat{x}} = \sup_{\hat{x}\in\hat{\Omega}} \frac{1}{\lambda} \|\boldsymbol{T}\left(\frac{\lambda\hat{x}}{\|\hat{x}\|}\right)\| = \frac{1}{\lambda} \sup_{\|z\|=\lambda} \|\boldsymbol{T}z\| \; . \tag{5.107}$$

The first equality is the definition of the norm of $\boldsymbol{T}$, the second equality is due to the fact that $\boldsymbol{T}$ is linear and the last equality is just a change of variables. Hence we have

$$\|\boldsymbol{T}\| = \frac{1}{\hat{\rho}} \sup_{\|z\|=\hat{\rho}} \|\boldsymbol{T}z\| \; . \tag{5.108}$$

Let $\hat{z}$ be any such vector for which $\|\hat{z}\| = \hat{\rho}$. Then for such a $\hat{z}$ we can find two vectors

$\hat{x}, \hat{y} \in \hat{\Omega}$ such that $\hat{z} = \hat{x} - \hat{y}$. This follows from the definition of $\hat{\rho}$. So

$$z = \boldsymbol{T}\hat{z} = \boldsymbol{T}\hat{x} + c - \boldsymbol{T}\hat{y} - c = x - y \in \Omega \ . \tag{5.109}$$

Since $z \in \Omega$ we have $\|z\| = \|\boldsymbol{T}\hat{z}\| \leq h$. Hence

$$\|T\| = \frac{1}{\hat{\rho}} \sup_{\|z\|=\hat{\rho}} \|Tz\| \leq \frac{h}{\hat{\rho}} \ . \tag{5.110}$$

The second inequality is proven analogously. **End Proof**

We next prepare two important lemmas, one involving a property of the (local) interpolation error $v - \Pi v$ and another on the behavior of seminorms of functions in Sobolev spaces under affine transformations.

Lemma 2 *Let $W^{k+1,p}(\Omega)$ be a Sobolev space continuous embedded in another Sobolev space $W^{m,q}(\Omega)$, and let $\Pi \in \mathcal{L}(W^{k+1,p}(\Omega), W^{m,q}(\Omega))$ be a continuous linear operator from $W^{k+1,p}(\Omega)$ onto $W^{m,q}(\Omega)$ which preserves polynomials of degree $\leq k$; that is*

$$\Pi w = w \ , \quad \forall w \in \mathcal{P}_k(\Omega) \ . \tag{5.111}$$

Then there exists a constant $C = C(\Omega)$ such that for every $v \in W^{k+1,p}(\Omega)$

$$|v - \Pi v|_{m,q,\Omega} \leq C(\Omega)\|I - \Pi\|_{\mathcal{L}(W^{k+1,p}(\Omega),W^{m,q}(\Omega)}|v|_{k+1,p,\Omega} \ . \tag{5.112}$$

Before proving this, some additional comments. A normed space U is embedded in a normed space V (with norms $\|\cdot\|_U$ and $\|\cdot\|_V$, respectively) if

- U is a linear subspace of V, and
- the injection (the identity operator $i : U \to V$ such that $iu = u \in V$, $\forall u \in U$) for U into V is continuous. Since i is linear, this condition is equivalent to the condition that there exists a constant $C > 0$ such that

$$\|u\|_V \leq C\|u\|_U \quad \forall u \in U \ . \tag{5.113}$$

The semi-norms appearing in this lemma are defined as

$$|f|_{m,p,\Omega} = \left\{ \sum_{|\alpha|=k} \int_\Omega |D^\alpha f|^p \, d\Omega \right\}^{\frac{1}{p}} . \tag{5.114}$$

Again, this is not a genuine norm, since $|f|_{m,p,\Omega}$ does not imply that $f = 0$, therefore it is called a semi-norm. (A semi-norm can be turned into a norm, when all normal derivatives of f along $\partial\Omega$ up to order $m-1$, including f itself are set to zero. This implies boundedness in at least one direction. These subtleties are beyond the scope of these lectures). Now that we know what the lemma means it is time to actually prove the lemma.

Proof Let $w \in \mathcal{P}_k(\Omega)$, i.e. the space of polynomials defined over Ω up to degree k. Then

$$v - \Pi v = v - \Pi v + w - \Pi w = (I - \Pi)(v + w) , \tag{5.115}$$

for every $v \in W^{k+1,p}(\Omega)$. We have

$$\begin{aligned} |v - \Pi v|_{m,q,\Omega} &\leq \|v - \Pi v\|_{m,q,\Omega} \\ &= \|(I - \Pi)(v + w)\|_{m,q,\Omega} \\ &\leq \|I - \Pi\|_{\mathcal{L}(W^{k+1,p}(\Omega), W^{m,q}(\Omega))} \inf_{w \in \mathcal{P}_k(\Omega)} \|v + w\|_{m,q,\Omega} \end{aligned} \tag{5.116}$$

Since $W^{k+1,p}(\Omega)$ is continuously embedded in $W^{m,q}(\Omega)$ it follows that

$$|v - \Pi v|_{m,q,\Omega} \leq C \|I - \Pi\|_{\mathcal{L}(W^{k+1,p}(\Omega), W^{m,q}(\Omega))} \inf_{w \in \mathcal{P}_k(\Omega)} \|v + w\|_{k+1,p,\Omega} . \tag{5.117}$$

The only thing we have to proof now is that $\inf_{w \in \mathcal{P}_k(\Omega)} \|v+w\|_{k+1,p,\Omega} = |v|_{k+1,p,\Omega}$. In order to do this we have to introduce the quotient space

$$Q^{k+1,p}(\Omega) = W^{k+1,p}(\Omega)/\mathcal{P}_k(\Omega) . \tag{5.118}$$

The elements of $Q^{k+1,p}(\Omega)$ are cosets $[v]$ of functions such that $\forall u, v \in W^{k+1,p}(\Omega)$

$$u \in [v] \Longrightarrow u - v \in \mathcal{P}_k(\Omega) . \tag{5.119}$$

The natural norm on $Q^{k+1,p}(\Omega)$ is given by

$$\|[v]\|_{Q^{k+1,p}(\Omega)} = \inf_{g \in \mathcal{P}_k(\Omega)} \|v + g\|_{k+1,p,\Omega} \ . \tag{5.120}$$

So with the introduction of the quotient space our interpolation estimate reduces to

$$|v - \Pi v|_{m,q,\Omega} \leq C \|I - \Pi\|_{\mathcal{L}(W^{k+1,p}(\Omega), W^{m,q}(\Omega))} \|[v]\|_{Q^{k+1,p}(\Omega)} \ . \tag{5.121}$$

So we now claim that there exists a positive constant $C = C(\Omega)$ such that

$$\|[v]\|_{Q^{k+1,p}(\Omega)} \leq C |v|_{k+1,p,\Omega} \ . \tag{5.122}$$

If we manage to prove this, the proof of the lemma is complete. Well, let the dimension of $\mathcal{P}_k(\Omega)$ be N (this value depends on the number of spatial dimensions, for instance, in one dimension $N = k + 1$) and let l_i, $1 \leq i \leq N$, denote a basis for the dual space. Then $l_i(g) = 0$, $i = 1, 2, \ldots, N$ and $g \in \mathcal{P}_k(\Omega)$, if and only if $g \equiv 0$. By the Hahn-Banach theorem, each l_i can be extended to continuous linear functionals on the whole space $W^{k+1,p}(\Omega)$, also denoted by l_i which coincides with those in $\mathcal{P}_k'(\Omega)$ and which therefore also have the property that if $l_i(g) = 0$ for all l_i, then $g = 0$. Let us first prove that there exists a constant C such that

$$\|v\|_{k+1,p,\Omega} \leq C \left\{ |v|_{k+1,p,\Omega} + \sum_{i=1}^{N} |l_i(v)| \right\} \ , \quad \forall v \in W^{k+1,p}(\Omega) \ . \tag{5.123}$$

Suppose this were not true, then there would be a sequence $\{v_k\}$ with $\|v_k\|_{k+1,p,\Omega} = 1$ in $W^{k+1,p}(\Omega)$ such that

$$\lim_{k \to \infty} \left\{ |v|_{k+1,p,\Omega} + \sum_{i=1}^{N} |l_i(v)| \right\} = 0 \ . \tag{5.124}$$

Now the space $W^{k+1,p}(\Omega)$ is compact in $W^{k,p}(\Omega)$, hence we extract a subsequence of $\{v_k\}$, also denoted by $\{v_k\}$ which converges strongly in $W^{k,p}(\Omega)$ to some function v. However, (5.124) implies that $|v|_{k+1,p,\Omega} \to 0$ and since $W^{k+1,p}(\Omega)$ is complete, the limit v must be such that $\|D^\alpha v\|_{0,p,\Omega} = \lim_{k \to \infty} \|D^\alpha v_k\|_{0,p,\Omega} = 0$ for $|\alpha| = k + 1$. In other words

$$\left. \begin{array}{c} v_k \to v \quad \text{in} \ \ W^{k,p}(\Omega) \\ \text{and} \\ |v_k|_{k+1,p,\Omega} \to 0 \end{array} \right\} \ . \tag{5.125}$$

$$\Longrightarrow \left\{ \begin{array}{ll} v_k \to v & \text{in } W^{k+1,p}(\Omega) \\ & \text{and} \\ D^\alpha v = 0 & \text{in } \mathcal{D}'(\Omega) \text{ for } |\alpha| = k+1 \end{array} \right. \tag{5.126}$$

The fact that $D^\alpha v = 0$ for $|\alpha| = k+1$ implies that v is a polynomial of degree k. Since $l_i(v) = \lim_{k\to\infty} l_i(v_k) = 0$, we conclude that $v = 0$. But this contradicts the fact that $\|v\|_{k+1,p,\Omega} = 1$, therefore (5.123) must be true.

Returning now to our intitial point: pick a $q \in \mathcal{P}_k(\Omega)$ such that $l_i(v+q) = 0$, $i = 1, 2, \ldots, N$, then

$$\begin{aligned} \|[v]\|_{Q^{k+1,p}(\Omega)} &= \inf_{g\in\mathcal{P}_k(\Omega)} \|v+g\|_{k+1,p,\Omega} && (5.127)\\ &\le \|v+q\|_{k+1,p,\Omega} && (5.128)\\ &\le C\left\{ |v+q|_{k+1,p,\Omega} + \sum_{i=1}^{N} |l_i(v+q)| \right\} && (5.129)\\ &= C|v+q|_{k+1,p,\Omega} && (5.130) \end{aligned}$$

Inserting this result in (5.121) gives

$$|v - \Pi v|_{m,q,\Omega} \le C(\Omega)\|I-\Pi\|_{\mathcal{L}(W^{k+1,p}(\Omega),W^{m,q}(\Omega))}|v|_{k+1,p,\Omega}\ . \tag{5.131}$$

End Proof

The next lemma tells us how the semi-norms are related under a affine transformation.

Lemma 3 *Let Ω and $\hat{\Omega}$ be two open affine equivalent subsets in $\mathcal{R}^n$; that is, let there be an invertible matrix $\boldsymbol{T}$ and a vector c such that*

$$\forall x \in \Omega \quad x = \boldsymbol{T}\hat{x} + c\ , \quad \hat{x} \in \hat{\Omega}\ . \tag{5.132}$$

Then there exists a constant $C = C(n,m)$ such that

$$\left.\begin{array}{ll} \forall v \in W^{m,q}(\Omega) & \\ & |\hat{v}|_{m,q,\hat{\Omega}} \le C\|\boldsymbol{T}\|^m |det(\boldsymbol{T})|^{-1/q}|v|_{m,q,\Omega} \\ \forall \hat{v} \in W^{m,q}(\hat{\Omega}) & \\ & |v|_{m,q,\Omega} \le C\|\boldsymbol{T}^{-1}\|^m |det(\boldsymbol{T})|^{1/q}|\hat{v}|_{m,q,\hat{\Omega}} \end{array}\right\}\ , \tag{5.133}$$

where $\hat{v}(\hat{x}) = v(x)$.

Proof We start with the defintion of the seminorm of $\hat{v}$

$$|\hat{v}|_{m,q,\hat{\Omega}} = \left\{ \int_{\hat{\Omega}} \sum_{|\alpha|=m} |D^{\alpha}\hat{v}(\hat{x})|^q \, d\hat{\Omega} \right\}^{\frac{1}{q}} . \tag{5.134}$$

Recall that if $\mathcal{D}^m\hat{v}$ denotes the Fréchet derivative of $\hat{v}$, then, for any vectors $y_i \in \mathcal{R}^n$, $1 \leq i \leq m$ we have

$$\mathcal{D}^m\hat{v}(\hat{x}) \cdot (y_1, y_2, \ldots, y_m) = \mathcal{D}^m v(x) \cdot (\boldsymbol{T}y_1, \boldsymbol{T}y_2, \ldots, \boldsymbol{T}y_m) , \tag{5.135}$$

so that

$$\|\mathcal{D}^m\hat{v}(\hat{x})\| \leq \|\boldsymbol{T}\|^m \|\mathcal{D}^m v(x)\| , \tag{5.136}$$

where

$$\|\mathcal{D}^m\hat{v}(\hat{x})\| = \sup_{\|y_i\| \leq 1} |\mathcal{D}^m\hat{v}(\hat{x}) \cdot (y_1, y_2, \ldots, y_m)| . \tag{5.137}$$

Since

$$D^{\alpha}\hat{v}(\hat{x}) = \mathcal{D}^m\hat{v}(\hat{x}) \cdot (i_1^{\alpha_1}, i_2^{\alpha_2}, \cdots, i_n^{\alpha_n}) \quad |\alpha| = m , \tag{5.138}$$

where the i_i are the orthonormal basisvectors for $\mathcal{R}^n$, we have

$$|D^{\alpha}\hat{v}(\hat{x})| \leq \|\mathcal{D}^m\hat{v}(\hat{x})\| , \quad |\alpha| = m . \tag{5.139}$$

Hence

$$|\hat{v}|_{m,q,\hat{\Omega}} \leq C(m,n) \left[\int_{\hat{\Omega}} \|\mathcal{D}^m\hat{v}(\hat{x})\|^q \, d\hat{\Omega} \right]^{1/q}$$

$$\leq C(m,n)\|T\|^m \left[\int_{\Omega} \|\mathcal{D}^m v(x)\|^q |\det(\boldsymbol{T}^{-1})| \, d\Omega \right]^{1/q} . \tag{5.140}$$

However, $\|\mathcal{D}^m v(x)\| \leq C_1(m,n) \max_{|\alpha|=m} D^\alpha v(x)|$, so that

$$\int_\Omega \|\mathcal{D}^m v(x)\|^q |\det(\boldsymbol{T}^{-1})|\, d\Omega \leq C_2(m,n)|\det(\boldsymbol{T}^{-1})| \cdot |v|^q_{m,q,\Omega} \,. \tag{5.141}$$

Combining this results with (5.140) gives the first inequality. The second inequality is proven in an analogous manner. **End Proof**

All these lemmas have prepared us for the final Theorem.

Theorem 6 *Let $(\widehat{\Omega}, \widehat{D}, \widehat{P})$ be a finite element for which the set $\widehat{D}$ of degrees of freedom involves the specification of partial derivatives of order $s \geq 0$. In addition, for positive integers m and k, let*

$$\left.\begin{array}{c} W^{k+1,p}\left(\widehat{\Omega}\right) \sqsubset \overline{C}^s\left(\widehat{\Omega}\right) \\ W^{k+1,p}\left(\widehat{\Omega}\right) \sqsubset W^{m,q}\left(\widehat{\Omega}\right) \\ \mathcal{P}_k\left(\widehat{\Omega}\right) \subset \widehat{P} \subset W^{m,q}\left(\widehat{\Omega}\right) \end{array}\right\} \tag{5.142}$$

Then there exists a positive constant $C = C(\widehat{\Omega}, \widehat{D}, \widehat{P})$, depending on the properties of $(\widehat{\Omega}, \widehat{D}, \widehat{P})$, such that for all elements $(\overline{\Omega}_e, D_e, P_e)$ affine equivalent to $(\widehat{\Omega}, \widehat{D}, \widehat{P})$ and all $v \in W^{k+1,p}\left(\widehat{\Omega}\right)$, we have

$$|v - \Pi_e v|_{m,q,\Omega_e} \leq C(\widehat{\Omega}, \widehat{D}, \widehat{P}) meas\,(\Omega_e)^{1/q-1/p} \frac{h_e^{k+1}}{\rho_e^m} |v|_{k+1,p,\Omega_e} \tag{5.143}$$

where $\Pi_e v$ denotes the P_e-interpolant of v and

$$h_e = dia\,(\Omega_e)$$

$$\rho_e = \sup\{dia\,(S)\,; S \text{ is a sphere contained in } \Omega_e\}$$

Proof Since $\mathcal{P}_k\left(\widehat{\Omega}\right) \subset \widehat{P}$ and $\widehat{D}$ is $\widehat{P}$-unisolvent, we observe that if $\widehat{\Pi}$ is a $\widehat{\Omega}$-interpolant operator, then

$$\widehat{\Pi}\widehat{w} = \widehat{w} \qquad \forall \widehat{w} \in \mathcal{P}_k\left(\widehat{\Omega}\right) \tag{5.144}$$

The inclusions (5.142) show that $\widehat{\Pi} \in \mathcal{L}\left(W^{k+1,p}\left(\widehat{\Omega}\right), W^{m,q}\left(\widehat{\Omega}\right)\right)$. In view of (5.144) and Lemma ??, we have

$$\left|\widehat{v} - \widehat{\Pi}\widehat{v}\right|_{m,q,\widehat{\Omega}} \leq C\left(\widehat{\Omega}\right) \left\|\widehat{I} - \widehat{\Pi}\right\|_{\mathcal{L}\left(W^{k+1,p}\left(\widehat{\Omega}\right), W^{m,q}\left(\widehat{\Omega}\right)\right)} \cdot |\widehat{v}|_{k+1,p,\widehat{\Omega}} = C(\widehat{\Omega}, \widehat{D}, \widehat{P})\, |\widehat{v}|_{k+1,p,\widehat{\Omega}} \quad (5.145)$$

for every $\widehat{v} \in W^{k+1,p}\left(\widehat{\Omega}\right)$. In light of Theorem ??, we have

$$\widehat{v} - \widehat{\Pi}\widehat{v} = v - \Pi_e v$$

Thus, turning to Lemma ?? and noting that the open sets $\text{int}\left(\widehat{\Omega}\right)$ and $\text{int}(\Omega_e)$ are necessarily equivalent we have

$$|v - \Pi_e v|_{m,q,\Omega_e} \leq C \left\|\mathbf{T}_e^{-1}\right\|^m |\det(\mathbf{T}_e)|^{1/q} \left|\widehat{v} - \widehat{\Pi}\widehat{v}\right|_{m,q,\widehat{\Omega}} \quad (5.146)$$

$$|\widehat{v}|_{k+1,p,\widehat{\Omega}} \leq C \|\mathbf{T}_e\|^{k+1} |\det(\mathbf{T}_e)|^{-1/p} |v|_{k+1,p,\Omega_e} \quad (5.147)$$

where $\mathbf{T}_e$ is an invertible matrix defining the affine mapping $F : \widehat{\Omega} \longrightarrow \Omega_e$. Combining (5.144) and (5.146), and (5.147) we have

$$\begin{aligned} |v - \Pi_e v|_{m,q,\Omega_e} \quad \leq \quad & C(\widehat{\Omega}, \widehat{D}, \widehat{P}) \|\mathbf{T}_e^{-1}\|^m \|\mathbf{T}_e\|^{k+1} \\ & \cdot |\det(\mathbf{T}_e)|^{1/q} |\det(\mathbf{T}_e)|^{-1/p} \\ & \cdot |v|_{k+1,p,\Omega_e} \end{aligned}$$

Noting that

$$|\det(\mathbf{T}_e)| = \frac{\text{meas}(\Omega_e)}{\text{meas}\left(\widehat{\Omega}\right)}$$

and recalling that

$$\left\|\mathbf{T}_e^{-1}\right\| \leq \frac{\widehat{h}}{\rho_e} \quad \text{and} \quad \|\mathbf{T}_e\| \leq \frac{h_e}{\widehat{\rho}}$$

we obtain the desired inequality. **End Proof.**

This error estimate can be embellished by setting $p = q$ and to define $h = \max_{1 \leq e \leq K} h_e$, so h is the largest mesh width in the grid. Then we have

$$\begin{aligned} |v - \Pi_h v|^p_{m,p,\Omega} &= \sum_{e=1}^{K} |v|_{\bar{\Omega}_e} - \Pi_e v|^p_{\bar{\Omega}_e} \\ &\leq \sum_{e=1}^{K} C \left(h_e^{k+1-m} |v|_{k+1,p,\Omega_e} \right)^p \\ &\leq C \left(h^{k+1-m} \sum_{e=1}^{K} |v|_{k+1,p,\Omega_e} \right)^p \\ &= C \left(h^{k+1-m} |v|_{k+1,p,\Omega} \right)^p \end{aligned}$$

and therefore we have

$$|v - \Pi_h v|_{m,p,\Omega} \leq C h^{k+1-m} |v|_{k+1,p,\Omega} \, . \tag{5.148}$$

Since the interpolant is any function in the finite dimensional function space, combining the above Theorem with Cea's theorem gives an error approximation of the finite element method. Note that for our one dimensional example $\rho_e = h_e = h$ and $\widehat{\rho} = \widehat{h} = 1$. So assuming that the exact solution is an element of $H^2(0,1)$ we have with $m = 0$, $k = 1$ and $p = q = 2$

$$\|u - u_N\|_{0,\Omega} \leq C h^2 \, |u|_{2,\Omega}$$

So we expect second order convergence as we have established in the previous section. If we want to know how the solution converges with respect to its first derivative in the L^2-norm, we take $m = 1$ and we find that convergence is only first order, precisely as was established in the previous section. So it is possible to theoretically find the rate of convergence without knowing the exact solution in advance.

There are even more reasons why robust error estimates are necessary in numerical analysis of schemes and the application of schemes. First of all the error estimates tells you on what numerical parameters the error depends. For instance, it is a well-known fact that the aspect ratio of elements should not be too large. The above theorem tells you why. Effectively, this means that h_e/ρ_e becomes large and then it depends on the m and k

whether this effects the error estimate or not. Furthermore the fact that $W^{k+1,p}(\Omega)$ should be closely embedded in $W^{m,q}(\Omega)$ tells you which error estimates are possible or which degree of polynomials are possible. The Sobolev embedding theorem gives conditions under which $W^{k+1,p}(\Omega)$ is continuous embedded in $W^{m,q}(\Omega)$. The following relations must hold

$$\left.\begin{array}{lll} (k+1-m)p < n & \text{and} & k+1-m > 0 \\ (k+1-m)p = 0 & \text{and} & p \leq q < \infty \\ (k+1-m)p > n & \text{and} & q \in \mathcal{R},\ \ 1 \leq q \end{array}\right\} \tag{5.149}$$

For our linear approximation $k = 1$ in one dimension $n = 1$ with the use of L^2 norms $p = q = 2$ this means that the error estimate is only usefull if

$$2(2-m) > 1\ . \tag{5.150}$$

This is only the case for $m = 0$ and $m = 1$ and this is precisely what we have found. The one dimensional finite element method leads to a convergent solution for the approximation itself and the approximation of the derivative, but the second derivative, i.e. $m = 2$ does not converge.

And another reason to have these error estimates is to decide whether it is useful to increase the polynomial order p-refinement, or to refine the mesh, h-refinement. If the exact solution is only contained in $H^s(\Omega)$, $s \geq 1$ it will not be useful to take polynomials of degree higher than $s-1$. (Why?). So in order to get a more accurate solution, the only way is to refine the mesh. If on the other hand s is sufficiently high it might be better to increase the polynomial degree to obtain a more accurate solution. These considerations play an essential role in so-called hp-adaptive methods, where in certain parts of the domain it is preferable to decrease h, i.e. refine the mesh, and in other parts to increase p, i.e. the polynomial degree. It is expected that these kinds of methods will increase the accuracy of the computations considerably, or, if one pursues a certain accuracy, to decrease the number of degrees of freedom. The next section will give a first step in higher order approximation methods.

5.4 Spectral Methods

In the previous section the basic idea behind the finite element method was explained. We started with a weak formulation and applied the Galerkin approximation. The next step was to chose local basis functions which led naturally to the finite element method. It is straightforward to represent the solution in a finite element by higher order polynomials.

For instance, instead of the linear functions on the parent element we could have used quadratic functions, defined on the parent element as

$$\psi^I(\xi) = 2(\xi - \frac{1}{2})(\xi - 1) \ , \quad \psi^{II}(\xi) = 4\xi(1 - \xi) \ , \quad \psi^{III}(\xi) = 2\xi(\xi - \frac{1}{2}) \ . \tag{5.151}$$

We may of course continue in this way, by increasing the polynomial degree. Note however, that we have chosen as a additional node in our finite element the point $\xi = 1/2$ in our quadratic element. We could also have chosen this additional node at $\xi = 1/4$ rendering the following basis functions

$$\psi^I(\xi) = 4(\xi - \frac{1}{4})(\xi - 1) \ , \quad \psi^{II}(\xi) = \frac{16}{3}\xi(1 - \xi) \ , \quad \psi^{III}(\xi) = \frac{4}{3}\xi(\xi - \frac{1}{4}) \ . \tag{5.152}$$

In order to extend the low order method systematically to higher order schemes the concept of spectral methods will be introduced. This can be done per finite element, thus leading to so-called *spectral element methods*, but we will take the Galerkin formulation as discussed before as our starting point. The extension of a spectrale method to a spectral element method will from then on be straightforward.

The new concept we will have to learn will be *orthogonal polynomials*. The study of orthogonal polynomials is one of the basic building blocks of physical analysis. Sines and Cosines are not polynomials, but constitute a system of orthogonal functions. The expansion of a given function in terms of sines and cosines, i.e. an expansion in terms of orthogonal functions, will be familiar to all of you. This expansion is known as the Fourier series. In signal analysis this expansion is also interpreted as a mapping from physical space to frequency space and the weights associated with a frequency constitute the spectrum of the signal. This idea of calculating the spectrum of the solution led to the name *spectral methods*. However, there are much more systems of orthogonal polynomials for which the term 'spectral analysis' seems to be ill-chosen. For instance, spherical Legendre polynomials describe the spin state of an elementary particle in quantum mechanics, but you have probably also have heard of Hermite polynomials, Laguerre polynomials, Gegenbauer polynomials, Chebyshev polynomials or Jacobi polynomials. And this list is not exhaustive; without too much difficulty you may come up with a set of functions which may bear your name. We will focuss mainly on the Legendre polynomials, despite the fact that other types might be preferable in some problems. However, the judgement which polynomials to use under which circumstances requires in-depth knowledge of the functional properties of orthogonal polynomials and the Legendre polynomials are closely related to the linear basis functions we have discussed previously.

5.4.1 Legendre Polynomials

At the beginning of this chapter we have introduced the concept of a inner product defined on a function space. For instance, the L^2 inner product was defined as

$$(f, g) := \int_\Omega f g \, d\Omega \ . \tag{5.153}$$

It can be readily shown that this definition satisfies all the requirements of an inner product. From this inner product we can deduce a norm, defined as

$$\|f\|^2 = (f, f) = \int_\Omega f^2 \, d\Omega \ . \tag{5.154}$$

We also know that the following relation holds

$$(f, g) = \|f\|\|g\| \cos\theta \ . \tag{5.155}$$

For vectors in Euclidian space, we know that θ is the angle between the two vectors f and g, but for functions it is harder to visualize what this angle means. We do know that when $\theta = \pi/2$ the two functions are 'perpendicular' to eachother (whatever that means for functions), or the two functions are *orthogonal*. Suppose now that we can construct a set of functions $\{L_i\}_{i=0}^{\infty}$ such that

$$(L_i, L_j) = c_i \delta_{ij} \ , \tag{5.156}$$

in which δ_{ij} denotes the Kronecker delta which equals 1 when $i = j$ and 0 otherwise. Then we can formally express any function u in Ω as

$$u(x) = \sum_{i=0}^{\infty} \alpha_i L_i(x) \ . \tag{5.157}$$

Note that we already used this idea in the development of the Galerkin formulation. The new thing is now the fact that the functions L_i are orthogonal, so if we want to know the coefficient α_j we simply multiply the above expansion with L_j, integrate over the domain Ω and divide by the constant c_j

$$\frac{1}{c_j} \int_\Omega u(x) L_j(x) \, d\Omega = \frac{1}{c_j} \sum_{i=0}^{\infty} \alpha_i \int_\Omega L_i(x) L_j(x) \, d\Omega = \alpha_j \ , \tag{5.158}$$

where the last equality follows from the fact that all integrals yield zero except when $i = j$. Note that if the functions L_i would not be orthogonal, the previous operation would yield one equation for infinitely many unknowns α_i. Essentially the whole procedure is pretty much the same as the one used in Fourier series. In order to get the Legendre polynomials we take $\Omega = [-1, 1]$ and use polynomials for the functions L_i.

We are free to chose our first basis function and we will take $L_0 = 1$ as our first basis function. Now we are going to determine $L_1 = ax + b$ such that this function is orthogonal to L_0. This means that

$$\int_{-1}^{1} (ax + b)\, dx = \left[\frac{1}{2}ax^2 + bx\right]_{-1}^{1} = 2b = 0 \; . \tag{5.159}$$

So setting $b = 0$ makes L_1 perpendicular to L_0. If we impose the condition that all $L_i(1) = 1$ we can determine a unique $L_1(x) = x$. The next polynomial will be a polynomial of degree two, $L_2(x) = ax^2 + bx + c$ and we have to chose the coefficients such that $(L_0, L_2) = 0$, $(L_1, L_2) = 0$ and $L_2(1) = 1$. These three conditions lead to

$$\begin{array}{ccccc} \frac{2}{3}a & & 2c & = & 0 \\ & \frac{2}{3}b & & = & 0 \\ a & b & c & = & 1 \end{array} \tag{5.160}$$

So we find

$$L_2(x) = \frac{1}{2}\left(3x^2 - 1\right) \; . \tag{5.161}$$

The next polynomial will be of degree three, must be orthogonal to all the previous polynomials and should equal 1 at the point $x = 1$. Solving for the four coefficients produces the desired polynomial. One can readily see that the calculation of the higher order orthogonal polynomials becomes more and more tedious. Fortunately, orthogonal polynomials obey a nice recurrence relation which for the Legendre polynomials is given by

$$L_{k+1}(x) = \frac{2k+1}{k+1} x L_k(x) - \frac{k}{k+1} L_{k-1}(x) \; . \tag{5.162}$$

So instead of solving 4 equations for 4 unknowns to obtain L_3 we simply take

$$L_3(x) = \frac{1}{3}\left(5xL_2(x) - 2L_1(x)\right) = \frac{1}{3}\left(\frac{5}{2}\left[3x^3 - x\right] - 2x\right) = \frac{1}{2}\left(5x^3 - 3x\right) \; . \tag{5.163}$$

Another important feature is that the Legendre polynomials are generated by a differential equation. This differential equation is given by

$$\big((1-x^2)L_k'(x)\big)' + k(k+1)L_k(x) = 0\ , \quad L_k(1) = 1\ . \tag{5.164}$$

Now that we know how the Legendre polynomials are defined and how we can compute them we can expand 'any' function in terms of the Legendre polynomals

$$u(x) = \sum_{i=0}^{\infty} \alpha_i L_i(x)\ , \quad \forall x \in [-1,1]\ . \tag{5.165}$$

The derivative of the function u can also be expressed in terms of the Legendre polynomials.

$$u'(x) = \sum_{i=0}^{\infty} \alpha_i L_i'(x) = \sum_{i=0}^{\infty} \tilde{\alpha}_i L_i(x)\ . \tag{5.166}$$

Now using the fact that for Legendre polynomials we have

$$(2k+1)L_k(x) = L_{k+1}'(x) - L_{k-1}'(x)\ , \tag{5.167}$$

we can write the last expansion as

$$\begin{aligned}
u'(x) &= \sum_{i=0}^{\infty} \tilde{\alpha}_i L_i(x) \\
&= \sum_{i=0}^{\infty} \left[\frac{\tilde{\alpha}_i}{2i+1} L_{i+1}'(x) - \frac{\tilde{\alpha}_i}{2i+1} L_{i-1}'(x) \right] \\
&= \sum_{i=1}^{\infty} \frac{\tilde{\alpha}_{i-1}}{2i-1} L_i'(x) - \sum_{i=-1}^{\infty} \frac{\tilde{\alpha}_{i+1}}{2i+3} L_i'(x) \\
&= \sum_{i=1}^{\infty} \left[\frac{\tilde{\alpha}_{i-1}}{2i-1} - \frac{\tilde{\alpha}_{i+1}}{2i+3} \right] L_i'(x) + \tilde{\alpha}_0 L_{-1}(x) + \frac{\tilde{\alpha}_1}{3} L_0'(x) \\
&= \sum_{i=1}^{\infty} \left[\frac{\tilde{\alpha}_{i-1}}{2i-1} - \frac{\tilde{\alpha}_{i+1}}{2i+3} \right] L_i'(x)
\end{aligned}$$

So the coefficients of the expansion in terms of $L_i(x)$ and $L_i'(x)$ are related by

$$\alpha_i = \frac{\tilde{\alpha}_{i-1}}{2i-1} - \frac{\tilde{\alpha}_{i+1}}{2i+3} , \quad \forall i \geq 1 . \tag{5.168}$$

So if we know the expansion of a function in terms of the Legendre polynomials we can recursively compute the coefficients for the derivative.

Example 30 *Suppose the function $u(x)$ in terms of a Legendre series exapnsion is given by*

$$u(x) = 1 \cdot L_0(x) + 3 \cdot L_1(x) + 2 \cdot L_2(x) . \tag{5.169}$$

Let us write the Legendre expansion for the derivative, $u'(x)$.

From (5.168) we know that

$$\begin{array}{l} \alpha_1 = \tilde{\alpha}_0 - \frac{1}{5}\tilde{\alpha}_2 = 3 \\ \alpha_2 = \frac{1}{3}\tilde{\alpha}_1 - \frac{1}{7}\tilde{\alpha}_3 = 2 \\ \alpha_k = \frac{\tilde{\alpha}_{k-1}}{2k-1} - \frac{\tilde{\alpha}_{k+1}}{2k+3} = 0 \quad k \geq 3 \end{array} \tag{5.170}$$

The last equations can be satisfied by setting $\tilde{\alpha}_k = 0$ for $k \geq 2$, then the first two equations yield $\tilde{\alpha}_1 = 3\alpha_2 = 6$ and $\tilde{\alpha}_0 = \alpha_1 = 3$ and therefore we obtain

$$u'(x) = 3 \cdot L_0(x) + 6 \cdot L_1(x) . \tag{5.171}$$

Let us check whether this is correct. Writing out $u(x)$ as a polynomial gives

$$u(x) = 1 + 3x + 3x^2 - 1 = 3 \cdot (x^2 + x) \tag{5.172}$$

The derivative of this function is then given by

$$u'(x) = 3 + 6x = 3 \cdot L_0(x) + 6L_1(x) . \tag{5.173}$$

Exactly as expected!!

5.5 Definitions and theorems

5.5.1 Cauchy-Schwartz inequality

Theorem 7 *Let X be an inner-product space, then for all u, $v \in X$ we have*

$$|(u,v)_X| \leq \|u\|_X \|v\|_X \ . \tag{5.174}$$

The equality holds if and only if u and v are linearly dependent.

Proof For every scalar t we have

$$0 \leq (u - tv, u - tv)_X = (u,u) - 2t(u,v) + t^2(v,v) \ . \tag{5.175}$$

If $(v,v)_X = 0$, then $(u,u) - 2t(u,v) \geq 0$ for all scalars t, which forces $(u,v)_X = 0$, in which case the inequality holds trivially. Now suppose $9v,v) \neq 0$. Substituting $t = (u,v)_X/(v,v)_X$ into this inequality, we obtain

$$0 \leq (u,u)_X - |(u,v)_X|^2/(v,v)_X \ , \tag{5.176}$$

which proves the Cauchy-Schwartz inequality.

If u and v are linearly dependent, i.e. if there exists a λ for which $u = \lambda v$ then the Cauchy-Schwartz inequality becomes an equality. Conversely, if for a certain t we have an equality in (5.175), then $u = tv$ which implies linear dependence between u and v. **End Proof**

5.5.2 Continuity and boundedness

Theorem 8 *Let $T : \mathcal{D}(T) \to Y$ be a linear operator, where $\mathcal{D}(T) \subset X$ and X, Y are normed spaces. Then*

- *T is continuous if and only if T is bounded.*
- *If T is continuous at a single point, T is continuous everywhere.*

Begin Proof (a) For $T = 0$ the statement is trivial. Let $T \neq 0$, then $\|T\| \neq 0$. We assume that T is bounded and consider any $x_0 \in \mathcal{D}(T)$. Let $\epsilon > 0$ be given. Then, since T is linear, for every $x \in \mathcal{D}(T)$ such that

$$\|x - x_0\| \leq \delta \qquad \text{where} \quad \delta = \frac{\epsilon}{\|T\|} , \tag{5.177}$$

we obtain

$$\|Tx - Tx_0\| = \|T(x - x_0)\| \leq \|T\|\|x - x_0\| < \|T\|\delta = \epsilon . \tag{5.178}$$

Since $x_0 \in \mathcal{D}(T)$ was arbitrary, this shows that T is continuous.

Conversely, assume that T is continuous at an arbitrary point $x_0 \in \mathcal{D}(T)$. Then, given any *epsilon* > 0, there exists a $\delta > 0$ such that

$$\|Tx - Tx_0\| \leq \epsilon \quad \text{for all } x \in \mathcal{D}(T) \text{ satisfying} \quad \|x - x_0\| \leq \delta . \tag{5.179}$$

We now take any $y \neq 0$ in $\mathcal{D}(T)$ and set

$$x = x_0 + \frac{\delta}{\|y\|} y \quad \text{Then} \quad x - x_0 = \frac{\delta}{\|y\|} y . \tag{5.180}$$

Hence $\|x - x_0\| = \delta$, so that we can use (5.179). Since T is linear, we have

$$\|Tx - Tx_0\| = \|T(x - x_0)\| = \|T\left(\frac{\delta}{\|y\|} y\right)\| = \frac{\delta}{\|y\|}\|Ty\| , \tag{5.181}$$

and (5.179) implies that

$$\frac{\delta}{\|y\|}\|Ty\| \leq \epsilon . \tag{5.182}$$

Thus

$$\|Ty\| \leq \frac{\epsilon}{\delta}\|y\| . \tag{5.183}$$

This can be written as $\|Ty\| \leq c\|y\|$, where $c = \epsilon/\delta$, and shows that T is bounded.

(b) Continuity of T at a point implies boundedness of T by the second part of (a), which in turn implies continuity of T by (a). **End proof**

5.5.3 Riesz's Representation Theorem

It is of practical importance to know the general form of bounded linear functionals opn various spaces. For general Banach spaces such formulae and their derivation can sometimes be complicated. However, for a Hilbert space the situation is suprisingly simple:

Theorem 9 *Every bounded linear functional f on a complete inner product space (Hilbert space) H can be represented in terms of an inner product, namely*

$$f(x) = (x, z)_H \ , \tag{5.184}$$

where z depends on f, is uniquely determined from f and has norm

$$\|z\| = \|f\| \tag{5.185}$$

Begin proof We will follow the following steps in the proof

a f has a representation of the form (5.184),

b z in (5.184) is unique,

c formula (5.185) holds.

The details are as follows:

a If $f = 0$, then (5.184) and (5.185) hold if we take $z = 0$. Let therefore $f \neq 0$. To motivate the idea of the proof, let us ask what properties z must have if a represenatation (5.184) exists. First of all $z \neq 0$ since otherwise $f = 0$. Second, $(x, z) = 0$ for all x for which $f(x) = 0$, that is, for all x in the null space of f, $\mathcal{N}(f)$. Hence $z \perp \mathcal{N}(f)$. This suggests that we consider $\mathcal{N}(f)$ and its orthogonal complement $\mathcal{N}(f)^\perp$.

Now we know that $\mathcal{N}(f)$ is a vector space and it is closed. Furthermore, $f \neq 0$ implies that $\mathcal{N}(f) \neq H$, so $\mathcal{N}(f)^\perp \neq \{0\}$. Hence $\mathcal{N}(f)^\perp$ contains an element $z_0 \neq 0$. We set

$$v = f(x)z_0 - f(z_0)x \ , \tag{5.186}$$

where $x \in H$ is arbitrary. Applying f, we obtain

$$f(v) = f(x)f(z_0) - f(z_0)f(x) = 0 \ . \tag{5.187}$$

This shows that $v \in \mathcal{N}(f)$. Since $z_0 \perp \mathcal{N}(f)$, we have

$$0 = (v, z_0) = (f(x)z_0 - f(z_0)x, z_0) = f(x)(z_0, z_0) - f(z_0)(x, z_0) \ . \tag{5.188}$$

Noting that $(z_0, z_0) = \|z_0\|^2 \neq 0$, we can solve for $f(x)$. The result is

$$f(x) = \frac{f(z_0)}{\|z_0\|^2}(x, z_0) \ . \tag{5.189}$$

This can be written in the form (5.184), where

$$z = \frac{f(\bar{z}_0)}{\|z_0\|^2} z_0 \ . \tag{5.190}$$

Since $x \in H$ was arbitrary, (5.184) has been proven.

b We prove that z in (5.184) is unique. Suppose that for all $x \in H$ we have

$$f(x) = (x, z_1) = (x, z_2) \ . \tag{5.191}$$

Then $(x, z_1 - z_2) = 0$ for all x. Choosing in particular $x = z_1 - z_2$, we have

$$(x, z_1 - z_2) = (z_1 - z_2, z_1 - z_2) = \|z_1 - z_2\|^2 = 0 \ . \tag{5.192}$$

Hence, $z_1 - z_2 = 0$, so that $z_1 = z_2$, the uniqueness.

c We finally prove that (5.185) holds. If $f = 0$, then $z = 0$ and (5.185) holds. Let $f \neq 0$. Then $z \neq 0$. From (5.184) with $x = z$ we obtain

$$\|z\|^2 = (z, z) = f(z) \leq \|f\|\|z\| \ . \tag{5.193}$$

Division by $\|z\| \neq 0$ yields $\|z\| \leq \|f\|$. It remains to show that $\|f\| \leq \|z\|$. From (5.184) and the Schwartz inequlatity we see that

$$|f(x)| = |(x, z)| \leq \|x\|\|z\| \ . \tag{5.194}$$

This implies that

$$\|f\| = \sup_{\|x\|=1} |(x,z)| \leq \|z\| \ . \tag{5.195}$$

End proof

5.6 Exercises

Exercise 24 *Determine the exact solution of equation (5.1).*

Exercise 25 *Show that* $x \in H^{-1}(0,1)$. *Determine* $\|x\|_{H^{-1}(0,1)}$.

Exercise 26 *Apply the Galerkin method to the one dimensional sample problem, using* $N = 3$ *and choose as basis functions* $\phi_i = \sin i\pi x$, $i = 1,2,3$. *Calculate the stiffness matrix* $\boldsymbol{K}$ *and the right-hand-side vector* $\boldsymbol{F}$ *and solve for the coefficients* α_i *and construct then the approximate solution* u_N. *Plot the exact and approximate solutions and comment on the accuracy of your approximation.*

Exercise 27 *Construct basis functions* ϕ_i, $i = 1,2,3$ *that are polynomials of degree* $(i+1)$. *Make sure that the basis functions satisfy the homogeneous boundary conditions. Use these basis functions to construct a Galerkin approximation. Compare the approximation with results obtained in the previous exercise.*

Exercise 28 *Show that for a conforming approximation of out one-dimensional model problem the basis functions need to be continuous. Kinks in the basis functions are allowed.*

Exercise 29 *Consider a continuous, coercive and bilinear form* $a(\cdot,\cdot)$ *and show that the solution to the problem*

$$a(u,v) =< F, v > \quad \forall v \in X$$

is bounded by

$$\|u\|_X \leq \frac{\|F\|_{X'}}{\alpha} \ ,$$

where α is the coercivity constant. Show that for small values of α small perturbation in F are amplified in the solution u.

Exercise 30 *Use the finite element method with piecewise linear basis functions to solve the equation*

$$-u" = 4\delta(x - \frac{1}{2}) \ , \quad 0 \leq x \leq 1 \ . \tag{5.196}$$

with $u(0) = u(1) = 0$. Show that the exact solution is given by

$$u(x) = \begin{cases} 2x & 0 \leq x \leq \frac{1}{2} \\ 2 - 2x & \frac{1}{2} \leq x \leq 1 \end{cases} \tag{5.197}$$

Calculate explicitly the error between the numerical solution and the approximate solution. What is the usefulness of the error estimates given in this chapter?

Exercise 31 *Apply quadratic elements to our one dimensional model problem and determine the rate of convergence for decreasing mesh size. How does this compare with the error estimates. Plot the error between the exact solution and the approximate solution versus h for both the piecewise linear approximation and the quadratic approximation.*

Exercise 32 *Show that the Legendre polynomial $L_3(x)$ obtained by the recurrence relation is orthogonal to all lower order polynomials and satsifies $L_3(1) = 1$. Demonstrate that the Legendre polynomials $L_2(x)$ which was constructed to be orthogonal to all lower order polynomials actually satisfies the recurrence relation. Write a small program to calculate $L_100(\frac{1}{2})$ without actually calculating all previous polynomials.*

Exercise 33 *Proof that $L_k(-1) = (-1)^k$.*

Exercise 34 *Proof that $|L_k(x)| \leq 1$ for all $k \geq 0$ and all $x \in [-1, 1]$.*

Exercise 35 *Check whether the first few Legendre polynomials which have been computed actually satisfy the differential equation (5.164).*