Finite Element Methods

Max D. Gunzburger Janet S. Peterson

August 26, 2014

Chapter 1

Introduction

{c

Many mathematical models of phenomena occurring in the universe involve differential equations for which analytical solutions are not available. For this reason, we must consider numerical methods for approximating the solution of differential equations. The finite element method is one such technique which has gained widespread use in a diverse range of areas such as fluid mechanics, structural mechanics, biological science, chemistry, electromagnetism, financial modeling, and superconductivity, to name a few. One can find articles where finite element methods have been employed to study everything from stress analysis of a human tooth to design of an airplane wing.

Although the foundations for the finite element method were laid in the first half of the twentieth century, it did not become widely used until much later. Structural engineers were the first to use the technique in the 1940's and 1950's; mathematicians became interested in analyzing and implementing the method in the late 1960's. The first symposium on the mathematical foundations of the finite element method was held in June of 1972 with over 250 participants and resulted in a now famous book by I. Babuska and A. Aziz. Prior to this symposium there had already been numerous national and international conferences held on the finite element method but mainly with an emphasis on engineering applications. In the following decades the finite element method has grown in popularity as a useful tool in design and application as well as a fertile area for mathematical analysis.

This first chapter is motivational in intent. We define, in the simplest possible setting, a finite element method. We then make an attempt to analyze the method; this attempt fails to be rigorous because we do not have in hand the necessary mathematical tools. However, in making the attempt, we learn something about the nature of the tools that we need to acquire. We then compare and contrast finite element methods to the finite difference approach and discuss some of the attractive features of finite element methods.

1.1 What are finite element methods?

{i

Finite element methods are a class of methods for obtaining approximate solutions of differential equations, especially partial differential equations.[1] As such, they can be compared to other methods that are used for this purpose, e.g., finite difference methods, finite volume methods or spectral methods. There are seemingly countless finite element methods in use, so that one cannot refer to any method as *the* finite element method any more that one can refer to any particular method as being *the* finite difference method. In fact, there are numerous subclasses of finite element methods, each saddled with a modifier, e.g., *Galerkin*, *mixed*, or *collocation* finite element methods. We draw distinctions between these different subclasses of finite element methods in later chapters.

The finite element method is distinguished from other approaches to approximating differential equations by the combination of variational methods and piecewise polynomial approximation. Piecewise polynomial approximation is very attractive due to the ease of use, its approximation properties, and the availability of bases which are locally supported; that is, bases that are nonzero over a small portion of the domain. Variational methods have their roots in the combination of partial differential equations and the calculus of variations. The Rayleigh-Ritz Method, conceived individually by Lord Rayleigh and Walther Ritz, is a variational technique to find the minimum of a functional defined on an appropriate space of functions as a linear combination of elements of that space. The variational aspect of the finite element method usually takes the form of a weak or variational problem. In this and later chapters we see that some of the problems we consider are equivalent to an unconstrained minimization problem such as Rayleigh-Ritz. On the other hand, the variational principles that the finite element method encompasses can handle problems which are related to constrained minimization and even those not related to optimization problems.

1.2 A Simple Example

{intro_sec_example}

In order to begin to understand the basic idea of the finite element method and the steps involved, we define a finite element method for the very simple two-point boundary value problem {intro_simpex}

{intro_simpde}
$$-u''(x) = f(x) \quad 0 < x < 1\,, \tag{1.1a}$$

{intro_simpbc0}
$$u(0) = 0\,, \tag{1.1b}$$

and

{intro_simpbc1}
$$u'(1) = 0\,. \tag{1.1c}$$

Here, $f(x)$ is a given function defined for $x \in (0,1)$ and $u(x)$ is the unknown function to be determined by solving (1.1). This boundary value problem can represent

[1] Finite element methods were not always thought of in this manner, at least in the structural mechanics community. In an alternate definition, structural systems are directly discretized into approximate submembers such as beams, plates, shells, etc., without any recourse to differential equations. These submembers are then called "finite elements."

a number of different physical situations; e.g., the temperature distribution in a uniform rod. It is important to note that this differential equation arises from a steady-state problem, that is, one that does not result from the time evolution of some initial condition.

The finite element approximation $u^h(x)$ to the solution $u(x)$ of (1.1) is defined to be the solution of the following problem:

$$\textit{find } u^h(x) \in V^h \textit{ such that } \quad \int_0^1 \frac{du^h}{dx}\frac{dv^h}{dx}\,dx = \int_0^1 f v^h\,dx \quad \forall\, v^h \in V^h\,, \tag{1.2}$$
{i

where V^h is a finite dimensional set (more precisely, a linear space[2]) of functions that vanish at $x = 0$ and are sufficiently smooth. Actually, Problem 1.2 defines a finite element method only if the approximating set V^h is chosen to consist of piecewise polynomial functions. This choice of approximating functions, along with a judicious choice of basis for V^h, is primarily responsible for the success of the finite element method as a computational method.

We now ask ourselves what (1.2) has to do with the original problem (1.1). An obvious connection is that since functions belonging to V^h vanish at $x = 0$ by definition, we have that $u^h(x)$ satisfies the boundary condition (1.1b). To see further connections, consider the following problem which is analogous to (1.2) except it is posed over an infinite dimensional vector space V instead of the finite dimensional space V^h:

$$\begin{gathered}\textit{find } u(x) \textit{ such that } u(0) = 0 \textit{ and}\\ \int_0^1 u'v'\,dx = \int_0^1 f v\,dx \quad \forall\, v \in V\,,\end{gathered} \tag{1.3}$$
{i

where for each $v \in V$, $v(0) = 0$ and v is "sufficiently smooth". One can view Problem 1.2 as an approximation of Problem 1.3. Integrating the left-hand side of (1.3) by parts and using the fact that $v(0) = 0$ allows us to write

$$v(1)u'(1) - \int_0^1 \big(u''(x) + f(x)\big)v(x)\,dx = 0\,. \tag{1.4}$$
{i

Now the arbitrariness of $v(x)$ implies that $u(x)$ also satisfies (1.1a) and (1.1c). To see this, we first choose an arbitrary $v(x)$ that vanishes at $x = 1$ as well as at $x = 0$. For all such $v(x)$, we have that

$$\int_0^1 \big(u''(x) + f(x)\big)v(x)\,dx = 0\,,$$

so that $u(x)$ satisfies (1.1a). However, if $u(x)$ satisfies (1.1a), then (1.4) simplifies to

$$v(1)u'(1) = 0\,,$$

where now again $v(1)$ is arbitrary. Thus, we obtain (1.1c) as well. Hence we have demonstrated that if $u(x)$ is a sufficiently smooth solution of (1.3) then it also satisfies (1.1).

[2] Linear or vector spaces will be discussed in Chapter 2.

Now, let us reverse the above steps that took us from (1.3) to (1.1). Specifically, we require that $u(0) = 0$ and we multiply (1.1a) by a sufficiently smooth function $v(x)$ that vanishes at $x = 0$ but is otherwise arbitrary. Then, we integrate the term involving the second derivative of u by parts and use the boundary condition (1.1c) to obtain (1.3). In this manner, one can show[3] that any solution $u(x)$ of (1.1) is also a solution of the problem (1.3). Is the converse true? We have seen that the answer is yes only if the solution of (1.3) is sufficiently differentiable so that substitution into (1.1a) makes sense. For this substitution to make sense, $u(x)$ should be (at least) twice continuously differentiable which, of course, requires that the given function $f(x)$ be continuous on $(0, 1)$. On the other hand, (1.3) may have solutions that cannot be substituted into (1.1a) because they are not sufficiently differentiable. For example, we will see in later chapters that (1.3) has a solution for some functions f that are not continuous; these solutions cannot be solutions of (1.1).

1.2.1 Some Terminology

{intro_sec_terminology}

Let us now introduce some terminology that will be used throughout this book. We call $u(x)$ a *classical solution* of (1.1) if, upon substitution into these relations, equality holds at every point $x \in (0, 1)$. We call solutions of (1.3) that are not classical solutions of (1.1) *weak solutions* of the latter problem and (1.3) itself is referred to as a *weak formulation* of (1.1).[4] Analogously, problem (1.2) is termed a *discrete weak problem.*

The functions v^h and u^h in (1.2) are called *test* and *trial* functions, respectively. The same terminology is used for the corresponding functions v and u appearing in (1.3). Where do these names come from? Suppose someone gave us a function $u^h(x)$ and claimed that it was a solution of the discrete weak problem (1.2). To verify the claim, we would put the function $u^h(x)$ on "trial," i.e., we would determine if substituting it into (1.2) results in the left-hand side equal to the right-hand side for all possible test functions $v^h(x) \in V^h$.

The Dirichlet boundary condition (1.1b) and the Neumann boundary condition (1.1c) are treated differently within the framework of the weak formulation (1.3) or its approximation (1.2). First, we note that the Neumann boundary condition (1.1c) is not imposed on the test or trial functions; however, we saw that if $u(x)$ satisfies the weak problem (1.3), then this Neumann boundary condition is indeed satisfied. Such boundary conditions, i.e., boundary conditions that are not required of the trial functions but are satisfied "naturally" by the weak formulation, are called *natural boundary conditions.* On the other hand, nothing in the process we used to go from the weak problem (1.3) to the classical problem (1.1) implied that the Dirichlet boundary condition (1.1b) was satisfied. For this reason, we imposed the boundary condition as a constraint on the possible trial functions. Such bound-

[3] All the necessary steps can be made rigorous.

[4] The terminology about solutions is actually richer than we have indicated. There are also solutions called *strong solutions* intermediate between weak and classical solutions. We postpone further discussions of the different types of solutions until we have developed some additional mathematical background.

ary conditions are called *essential boundary conditions.* Note that for the discrete problem, the approximate solution $u^h(x)$ satisfies (by construction) the essential boundary condition (1.1b) exactly, but that the natural boundary condition (1.1c) is only satisfied in a weak sense.

1.2.2 Polynomial Approximation

The two main components of the finite element method are its variational principles which take the form of weak problems and the use of piecewise polynomial approximation. In our example we use the discrete weak or variational formulation (1.2) to define a finite element method but we have not used piecewise polynomials yet. In this example we choose the simple case of approximating with piecewise linear polynomials; that is, a polynomial which is linear when restricted to each subdivision of the domain. To define these piecewise polynomials, we first discretize the domain $[0,1]$ by letting N be a positive integer and setting the *grid points* or *nodes* $\{x_j\}_{j=0}^N$ so that $0 = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = 1$. Consequently we have $N+1$ nodes and N elements. These nodes serve to define a partition of the interval $[0,1]$ into the subintervals $T_i = [x_{i-1}, x_i]$, $i = 1, \ldots, N$; note that we do not require the partition to be uniform. The subintervals T_i are the simplest examples of *finite elements.* We choose the finite dimensional space V^h in (1.2) to be the space of continuous piecewise linear polynomials over this partition of the given interval so that each v^h is a continuous piecewise linear polynomial. In particular, we denote $v_i^h(x) = v^h(x)|_{T_i}$ for $i = 1, \ldots, N$; i.e., $v_i^h(x)$ is the restriction of $v^h(x)$ to element T_i. For continuous piecewise linear polynomials, we formally define the set of functions V^h as follows: $v^h(x) \in V^h$ if

$$\begin{aligned} &\text{(i)} \quad v_i^h(x) \text{ is a linear polynomial for } i = 1, \ldots, N; \\ &\text{(i)} \quad v_i^h(x_i) = v_{i+1}^h(x_i) \text{ for } i = 1, \ldots, N-1, \text{ and} \\ &\text{(iii)} \quad v^h(x_0) = 0\,. \end{aligned} \tag{1.5}$$

Condition (i) of (1.5) guarantees that the function $v^h(x)$ is a piecewise linear polynomial, Condition (ii) guarantees continuity and Condition (iii) guarantees that v^h vanishes at $x = 0$. With this choice for V^h, (1.2) is called a *piecewise linear finite element method* for (1.1).

1.2.3 Connection with Optimization Problem

We note that the weak problems (1.2) and (1.3) can be associated with an optimization problem. For example, for a given $f(x)$ consider the functional

$$\mathcal{J}(v; f) = \frac{1}{2} \int_0^1 (v')^2 \, dx - \int_0^1 f v \, dx \tag{1.6}$$

and the unconstrained minimization problem:

$$\textit{find } u(x) \in V \textit{ such that} \quad \mathcal{J}(u; f) \le \mathcal{J}(v; f) \quad \forall\, v \in V\,,$$

where the space V is defined as before. Using standard techniques of the calculus of variations, one can show that a necessary requirement for any minimizer of (1.6) is satisfying the weak problem (1.3). The converse is also true so that the two problems (1.6) and (1.3) are equivalent. In fact, in engineering applications this minimization approach is often used since it has the interpretation of minimizing an energy. However, not all weak problems have an equivalent minimization problem. We discuss this and its implications in later chapters.

{intro_sec_implement}

1.3 How do you implement finite element methods?

We now translate the finite element method defined by (1.2) into something closer to what a computer can understand. To do this, we first show that (1.2) is equivalent to a linear algebraic system once a basis for V^h is chosen. Next we indicate how the entries in the matrix equation can be evaluated.

Let $\{\phi_i(x)\}_{i=1}^N$ be a basis for V^h, i.e., a set of linearly independent functions such that any function belonging to V^h can be expressed as a linear combination of these basis functions. Note that we have assumed that the dimension of V^h is N which is the case if we define V^h by (1.5). Thus, the set $\{\phi_i(x)\}|_{i=1}^N$ has the property that it is linearly independent, i.e.,

$$\sum_{i=1}^N \alpha_i \phi_i(x) = 0 \quad \text{implies} \quad \alpha_i = 0 \quad \text{for } i = 1, \ldots, N$$

and it spans the space. That is, for each $w^h \in V^h$ there exists real numbers w_i, $i = 1, \ldots, N$, such that

$$w^h(x) = \sum_{i=1}^N \omega_i \phi_i(x)\,.$$

In the weak problem (1.2), the solution $u^h(x)$ belongs to V^h and the test function $v^h(x)$ is arbitrary in V^h. Since the set spans V^h we can set $u^h = \sum_{j=1}^N \mu_j \phi_j$ and then express (1.2) in the following equivalent form: find $\mu_j \in \mathbb{R}^1$, $j = 1, \ldots, N$, such that

$$\int_0^1 \frac{d}{dx}\left(\sum_{j=1}^N \mu_j \phi_j(x)\right) \frac{d}{dx}\left(v^h\right) dx = \int_0^1 f(x) v^h \, dx \quad \forall\, v^h \in V^h\,.$$

Since this equation must hold for each function $v^h \in V^h$ then it is enough to test the equation for each element in the basis; that is, for each ϕ_i, $i = 1, \ldots, N$. Using this fact, the discrete problem is rewritten as

{intro_simpdweak2}

find μ_j, $j = 1, \ldots, N$, *such that*

$$\sum_{j=1}^N \left(\int_0^1 \phi_i'(x)\phi_j'(x)\,dx\right)\mu_j = \int_0^1 f\phi_i(x)\,dx \qquad \text{for } i = 1, \ldots, N. \tag{1.7}$$

Clearly (1.7) is a linear algebraic system of N equations in N unknowns. Indeed, if the entries of the matrix K and the vectors $\vec{U}$ and $\vec{b}$ are defined by

$$K_{ij} = \int_0^1 \phi_i'(x)\phi_j'(x)\,dx\,, \quad U_j = \mu_j\,, \quad \text{and} \quad b_j = \int_0^1 f(x)\phi_i\,dx \quad \text{for } i,j = 1,\ldots,N\,,$$

then, in matrix notation, (1.7) is given by

$$K\vec{U} = \vec{b}\,. \tag{1.8}$$

{i

However, we have not yet completely formulated our problem so that it can be implemented on a computer. We first need to choose a particular basis set and then the integrals appearing in the definition of K and $\vec{b}$ must be evaluated or approximated. Clearly there are many choices for a basis for the space of continuous piecewise linear functions defined by (1.5). We will see in Section 1.6 that a judicious choice of the basis set will result in (1.8) being a tridiagonal system of equations and thus one which can be solved efficiently in $O(N)$ operations.

For now, let's assume that we have chosen a specific basis and turn to the problem of evaluating or approximating the integrals appearing in K and $\vec{b}$. For a simple problem like ours we can often determine the integrals exactly; however, for a problem with variable coefficients or one defined on a general polygonal domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ this would not be practical. Even if we have software available that can perform the integrations, this would not lead to an efficient implementation of the finite element method. Thus to obtain a general procedure which would be viable for a wide range of problems, we approximate the integrals by a quadrature rule. For example, for the particular implementation we are developing here, we use the midpoint rule in each element to define the composite rule

$$\int_0^1 g(x)\,dx = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} g(x)\,dx \approx \sum_{k=1}^{N} g\left(\frac{x_{k-1}+x_k}{2}\right)(x_k - x_{k-1})\,.$$

Using this rule for the integrals that appear in (1.8), we are led to the problem

$$K^h\vec{U^h} = \vec{b^h}\,, \tag{1.9}$$

{i

where the superscript h on the matrix K and the vector $\vec{b}$ denotes the fact that we have approximated the entries of K and $\vec{b}$ by using a quadrature rule to evaluate the integrals. Using the midpoint rule, the entries of K^h and $\vec{b^h}$ are given explicitly by

$$K_{ij}^h = \sum_{k=1}^{N}(x_k - x_{k-1})\phi_i'\left(\frac{x_{k-1}+x_k}{2}\right)\phi_j'\left(\frac{x_{k-1}+x_k}{2}\right)\,, \quad \text{for } i,j = 1,\ldots,N$$

and

$$b_i^h = \sum_{k=1}^{N}(x_k - x_{k-1})f\left(\frac{x_{k-1}+x_k}{2}\right)\phi_i\left(\frac{x_{k-1}+x_k}{2}\right)\,, \quad \text{for } i = 1,\ldots,N\,.$$

In our example, $K^h = K$. To see this, recall that we have chosen V^h as the space of continuous piecewise linear functions on our partition of $[0,1]$ and thus the integrands in K are constant on each element T_i. The midpoint rule integrates constant functions exactly so even though we are implementing a quadrature rule, we have performed the integrations exactly. However, in general, $\vec{b}^h \neq \vec{b}$ so that $\vec{U}^h \neq \vec{U}$.

Once the specific choice of a basis set for V^h is made, the matrix problem (1.9) can be directly implemented on a computer. A standard linear systems solver can be used to obtain $\vec{U}^h$. To efficiently solve (1.9) the structure and properties of K^h should be taken into consideration.

There are an infinite number of possible basis sets for a finite element space. If the basis functions have global support, e.g., if they are nonzero over the whole interval $(0,1)$, then, in general, the resulting discrete systems such as (1.8) or (1.9) will involve full matrices, i.e., matrices having possibly all nonzero entries.

In order to achieve maximum sparsity in the discrete systems such as (1.8) or (1.9), the basis functions should be chosen to have local support, i.e., to be nonzero on as small a portion of the domain as possible. Typically the basis functions are required to have *compact support*, that is, they are zero outside of a compact set; in finite elements the compact set consists of adjacent elements. In the one dimensional case we have considered here, the basis functions should be nonzero over as small a number of subintervals as possible. Such a basis set is provided by the "hat" functions defined by

{intro_simphat}

$$\text{for } i = 1, \ldots, N-1, \qquad \phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{for } x_{i-1} \le x \le x_i \\[2ex] \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & \text{for } x_i \le x \le x_{i+1} \\[2ex] 0 & \text{otherwise} \end{cases} \tag{1.10}$$

and

{intro_simphatn}

$$\phi_N(x) = \begin{cases} \dfrac{x - x_{N-1}}{x_N - x_{N-1}} & \text{for } x_{N-1} \le x \le x_N \\[2ex] 0 & \text{otherwise.} \end{cases} \tag{1.11}$$

A sketch of these functions for the case $N = 4$ on a nonuniform partition of $[0,1]$ is given in Figure 1.3. Note that for all $i = 1, \ldots, N$, $\phi_i(0) = 0$, $\phi_i(x)$ is continuous on $[0,1]$, is a linear polynomial on each subinterval $[x_{j-1}, x_j]$, $j = 1, \ldots, N$, and $\phi_i(x)$ is nonzero only in $[x_{i-1}, x_{i+1}]$. It can be shown that the set $\{\phi_i(x)\}_{i=1}^N$ given by (1.10) and (1.11) is linearly independent and forms a basis for the space V^h defined by (1.5).

Now let's examine the entries of the matrices K and K^h appearing in the linear systems (1.8) or (1.9), respectively, for the basis functions defined in (1.10) and (1.11). It is easy to see that both $K_{ij} = 0$ and $K^h_{ij} = 0$ unless $|i - j| \le 1$. Thus, for any number of elements N, these matrices have nonzero entries only along the main diagonal and the first upper and lower subdiagonals, i.e., they are

{intro_fig_hat}

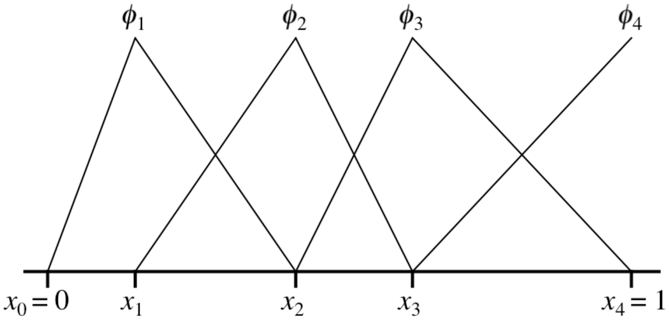


Figure 1.1. *Example of the hat basis functions for four intervals.*

tridiagonal matrices. This is the optimal sparsity achievable with piecewise linear finite elements. As a result, one can apply very inexpensive methods to solve the linear systems (1.8) or (1.9).[5]

1.4 What is needed to analyze finite element methods?

{i

In the previous section we saw how to implement the finite element method for a simple two point boundary value problem. In this section we turn to the question of determining how accurate the approximate solution is in our example. Specifically, we want to derive an error estimate, i.e., a bound for the difference between the finite element approximation u^h satisfying (1.2) and the weak solution u satisfying (1.3). In deriving this estimate we ignore the fact that in the implementation stage of solving our problem we introduced another error via the use of a quadrature rule to evaluate the integrals. This is reasonable because, in general, one chooses a quadrature rule whose error is small enough that it will be dominated by the finite element error. We discuss this more in later chapters.

To derive the estimate we first note that since the test function v in (1.3) satisifes $v(0) = 0$ and is sufficiently smooth but otherwise“arbitrary”, we may choose $v = v^h \in V^h \subset V$ in that equation since the same smoothness is required of v^h and $v^h(0) = 0$. We have

$$\int_0^1 \frac{du}{dx}\frac{dv^h}{dx}\,dx = \int_0^1 f(x)v^h(x)\,dx \quad \forall\, v^h \in V^h\,.$$

Then, we subtract (1.2) from this equation to obtain

$$\int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right)\frac{dv^h}{dx}\,dx = 0 \quad \forall\, v^h \in V^h\,. \tag{1.12}$$

{i

[5] For example, if one uses a direct, elimination algorithm, tridiagonal systems can be solved using $O(N)$ multiplications; this should be constrasted with the $O(N^3)$ work needed to solve a full linear system by Gaussian elimination.

This equation is called an *orthogonality condition.* We now use this orthogonality condition with $u^h \in V^h$ to write

$$\begin{aligned}\int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right)^2 dx &= \int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right) \frac{du}{dx}\, dx - \int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right) \frac{du^h}{dx}\, dx \\ &= \int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right) \frac{du}{dx}\, dx \\ &= \int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right) \left(\frac{du}{dx} - \frac{dw^h}{dx}\right) dx\,,\end{aligned}$$

where w^h is an arbitrary element of V^h. Using a standard inequality, we are lead to the expression

$$\begin{aligned}&\int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right) \left(\frac{du}{dx} - \frac{dw^h}{dx}\right) dx \\ &\qquad \le \sqrt{\int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right)^2 dx}\ \sqrt{\int_0^1 \left(\frac{du}{dx} - \frac{dw^h}{dx}\right)^2 dx}\end{aligned}$$

so that

{intro_simperror}
$$\sqrt{\int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right)^2 dx} \ \le\ \sqrt{\int_0^1 \left(\frac{du}{dx} - \frac{dw^h}{dx}\right)^2 dx} \qquad (1.13)$$

for arbitrary $w^h \in V^h$. The relationship (1.13) says the following: the root mean square error in the derivative of the finite element solution is less than or equal to the root mean square difference between the derivatives of the exact solution and any function in the approximating space V^h. In this sense, finite element approximations are " best approximations."

It would be nice if we could estimate the right-hand side of (1.13) in terms of parameters of the problem such as the grid spacing. In fact, this is possible. If we let $h = \max_{i=1}^N |x_i - x_{i-1}|$, it can be shown that the right-hand side of (1.13) is bounded by the product of a constant times h by using standard results from approximation theory when we approximate using continuous piecewise linear polynomials. Thus, we have the *error estimate*

{intro_simperrest}
$$\sqrt{\int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right)^2 dx} \ \le\ Ch \qquad (1.14)$$

in the case where V^h is defined by (1.5). Among other things, (1.14) implies that as $h \to 0$, i.e., as we increase the number of intervals N while reducing the maximum length, h, of the intervals, the error in the finite element solution as measured by the left-hand side of (1.14) tends to zero as well. Thus, we say that the finite element method is *convergent.*

The structure of the derivation of many finite element error estimates is similar to that outlined above. One first obtains an *orthogonality result* as typified by (1.12). Then, one derives a *best approximation result* such as (1.13). Finally, one uses approximation theoretic results about the best approximation to obtain an *error estimate* such as (1.14).

In our derivation of (1.14) we have omitted several details and have not been precise in the definition of the function space where we seek the solution to the weak problem. What is needed to make it precise? First of all, we have to go back to the beginning and make precise what we mean by "sufficiently smooth" in (1.3) and what are the functions f for which the weak problem possesses a unique solution. Next, we have to make precise all the steps that led to (1.13) and to the estimate of the right-hand side of (1.13) to arrive at (1.14). To obtain this estimate, we need to investigate approximation theory in finite element spaces, i.e., how well can piecewise polynomials approximate given functions. This theory, as well as estimates of the constant C in (1.14), which actually depends on the exact solution u, will need regularity results for solutions of (1.3), i.e., results relating the differentiability of the solution u to, among other things, the differentiability of the data function f. All of this will require some background knowledge of linear functional analysis and partial differential equations. In addition, when proving existence and uniqueness of a weak problem or deriving an error estimate, we don't want to consider each individual problem but rather obtain the results for a general weak problem. To do this, we must formulate a general weak problem which requires the introduction of appropriate function spaces and bilinear forms. In Chapter 2 we give a brief overview of selected topics from functional analysis and in Chapter 3 we introduce the appropriate function spaces, provide the machinery for defining an abstract weak problem, proving its existence and uniqueness and providing error estimates. Regularity results concerning differential equations will be quoted and referenced as needed.

1.5 A comparison with finite difference methods

{i

Like finite difference methods, particular finite element methods are ultimately defined based on a grid, i.e., a partition of a given domain in Euclidian space into subdomains. More often than not, the grid itself is defined by selecting a finite number of points in the given domain. Thus, both classes of methods may be thought of as grid-based methods so that, with some justice, one may view any finite element method as merely being a realization of a particular finite difference method. Conversely, with a little bit of work and ingenuity, finite difference methods may be given a finite element derivation.

What separates the two classes of methods? There are many good and valid answers that can and have been given to this question. A fundamental difference between the two methods is this: unlike finite difference methods, a finite element method can easily be "lifted" from the grid into a function space consisting of functions defined, for all practical purposes, everywhere in the given domain. This is exactly what we did in the example from Sections 1.1-1.4 by working with the set V^h so that details about the grid, although needed for the implementation of the finite element method, were incidental to its definition and analysis. On the other hand, finite difference methods remain intimately tied to the grid and their analysis involves functions defined over a discrete set of points, i.e., the grid. In general, this renders finite difference methods much more difficult to analyze.

To compare the two methods more fully, let us define a simple finite difference method for (1.1). As before, we begin with a uniform partition of the interval $[0,1]$ into N subintervals. Let $h = 1/N$ and $x_i = ih$ for $i = 0, \dots, N$. We let U_i denote an approximation to $u(x_i)$ which is the exact solution evaluated at x_i. We apply the boundary condition (1.1b) by setting $U_0 = 0$. In order to determine U_i for $i = 1, \dots, N$, we first approximate the differential equation (1.1a) by replacing, at every interior grid point x_i, the second derivative by a second central difference quotient to obtain

{intro_simpfd}

$$-\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} = f(x_i) \qquad \text{for } i = 1, \dots, N-1\,. \tag{1.15}$$

To approximate the remaining boundary condition (1.1c), we can use a one-sided difference quotient to obtain

{intro_simpfdbc}

$$\frac{U_N - U_{N-1}}{h} = 0\,. \tag{1.16}$$

Clearly, (1.15)-(1.16) form the linear algebraic tridiagonal system of N equations for the N unknowns U_i, $i = 1, \dots, N$ given by

{intro_simpfdsystem}

$$\frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots & 0 \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ 0 & \cdots & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-1} \\ U_N \end{pmatrix} = h \begin{pmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_N) \end{pmatrix}\,. \tag{1.17}$$

We can immediately point out another difference, albeit a somewhat philosophical one, between finite difference and finite element methods. The finite difference method (1.15)-(1.16) was derived primarily by *approximating operators*, i.e., derivatives. On the other hand, the primary approximation step in deriving the finite element method (1.2) was to replace the solution u in (1.3) by an approximation u^h, i.e., by *approximating the solution.*

To explore the relationship between the two types of methods, let's return to the finite element method of (1.9). If we assume that the partition T_i, $i = 1, \dots, N$, is uniform with $h = x_i - x_{i-1}$ for $i = 1, \dots, N$, that V^h is defined by (1.5), and that the midpoint rule is used, then the entries of K^h and $\vec{b}^h$ in (1.9) can be evaluated to obtain (see exercises for details)

{intro_simpfe}

$$
\begin{aligned}
K^h_{ii} &= \frac{2}{h} \quad \text{for } i = 1, \ldots, N-1, \quad K^h_N = \frac{1}{h} \\
K^h_{ij} &= -\frac{1}{h} \quad \text{for } i = 1, \ldots, N-1, |j-i| = 1, \quad K^h_{N,N-1} = -\frac{1}{h} \\
b^h_j &= \frac{h}{2}\left[f\left(\frac{x_{j-1}+x_j}{2}\right) + f\left(\frac{x_j + x_{j+1}}{2}\right)\right] \quad \text{for } j = 1, \ldots, N-1 \\
b^h_N &= \frac{h}{2} f\left(\frac{x_{N-1}+x_N}{2}\right).
\end{aligned}
\tag{1.18}
$$

Note the similarities and differences between the finite difference method (1.15)-(1.16) and our finite element method. In particular, notice that the coefficient matrix is identical to (1.17) but the right-hand side has an averaging performed on the right-hand side function $f(x)$ in the finite element approach. This is a typical feature of finite element methods that partially accounts for some of its advantageous properties.

1.6 What are the advantages of the finite element methods?

{i

Finite element methods possess many desirable properties that account for their popularity in a variety of settings. Some of these we have already encountered. For example, within the finite element method framework, natural boundary conditions such as (1.1c) are very easily enforced. We also saw that there is no difficulty in treating problems with nonuniform grids. A third advantage that we have alluded to is that, due to being able to introduce sophisticated function theoretic machinery, finite element methods can be "easily" analyzed with complete rigor. All three of these are thought of as posing difficulties within the finite difference framework.

There are other good features inherent in finite element methodologies. Perhaps the most important one is the ability of finite element methods to "easily" treat problems in complicated, e.g., non-rectangular, domains.[6] Another good feature is that finite element methods preserve certain symmetry and positivity properties possessed by problems such as (1.1). In particular, in this case, the matrices K and K^h appearing in (1.8) and (1.9), respectively, are symmetric and positive definite.

A final desirable feature of finite element methods is that, when they are properly implemented, they lead to sparse discrete problems. This spasity property is crucial to the efficiency of finite element methods and results from a judicious choice for the basis set $\{\phi_i(x)\}_{i=1}^N$ for the finite element space V^h.

1.7 Which method is best?

{i

Unfortunately, there is no best method for all problems. Which method is best, be it of the finite difference, finite element, finite volume, spectral, etc., type, depends

[6] Of course, since we have only looked at problems in one dimension, we have not yet been exposed to such domains.

on the class of problems under consideration or, often, on the specific problem itself. It is not even possible to say that one finite element method is better than another one in all settings. In order to determine which method is best (or even good) for a given problem, one must understand its definition, implementation, and analysis. The purpose of this book is to provide the tools, knowledge, and experience so that such judgments can be made of finite element methods and, if a similar familiarity with other classes of methods is obtained, one can then make rational comparisons and decisions.

There are some areas of applications wherein finite element methods are preponderant and therefore, at least judging by the number of users, are best. Chief among these is structural mechanics. In other areas, e.g., incompressible flows, heat transfer, electromagnetism, etc., finite element methods, although not quite so ubiquitous as in structural mechanics, have gained, if not dominance, at least widespread popularity. There are areas of applications wherein finite element methods, although in use, have not achieved anything near dominance. One example is inviscid, compressible flows containing shock waves and other discontinuities. Two interesting observations are in order. The first is that finite element methods have attained something close to dominance in those areas for which they can be fully and rigorously analyzed. The second is that the lack of such analyses in other areas is usually due to the lack of results about the partial differential equations themselves. These relationships between popularity and analyses may or may not be purely coincidental.

Exercises

1.1. Consider the two-point boundary value problem (BVP)

{intro_ex1.1}

$$\begin{aligned} -u'' + u &= x \quad 0 < x < 1\,, \\ u'(0) &= 2\,, \\ u(1) &= 0\,. \end{aligned} \tag{1.19}$$

Write down a weak formulation for the BVP given in (1.19); at this point you do not have to be specific about the underlying spaces. Show that a solution of your classical problem satisfies your weak formulation and that the converse is also true provided your solution of the weak problem is sufficiently smooth.

1.2. In the BVP given in (1.19), which boundary condition is essential and which is natural? Why?

1.3. Consider the piecewise linear function $\psi(x)$ given by

$$\psi(x) = \begin{cases} 8x & 0 \le x \le 0.25 \\ -2x + 2.5 & 0.25 \le x \le 0.5 \\ -4x + 3.5 & 0.5 \le x \le 0.75 \\ 6x - 4 & 0.75 \le x \le 1.0 \end{cases} \tag{1.20}$$

Write $\psi(x)$ as a linear combination of the standard "hat" basis functions on the given partition of $[0, 1]$.

1.4. Show that the entries of K^h and $\vec{b}^h$ are given by (1.18) provided the basis functions defined by (1.10) are used and the midpoint rule is employed to evaluate the integrals.

Chapter 2

Results from Linear Analysis

{c

In the last chapter we began to see the need for certain mathematical tools in order to rigorously analyze the finite element method. In an effort to have this book as self-contained as possible, we provide here a short summary of many of the commonly used results from functional analysis.

The main goal of this chapter is to introduce the mathematical tools necessary to precisely formulate and analyze a *general* weak problem and its discrete analogue. The advantage to this abstraction is that we are able to treat a wide class of problems within the same general framework. In later chapters when we investigate particular differential equations, we see that many of the weak formulations fit into this general framework. Thus, if we determine conditions which guarantee existence, uniqueness, and continuous dependence on the data of our general weak problem and derive an error estimate, then we can easily analyze a particular weak formulation by showing that it satisfies the hypotheses for the general problem.

In formulating weak problems we need to determine appropriate classes of function spaces to use for our test and trial spaces and to examine some of their basic properties. The particular types of spaces that are needed are certain Hilbert spaces which are named after the German mathematician David Hilbert (1862-1943). These spaces offer a natural setting for weak problems and can be considered as generalizations of Euclidean space. In general, these Hilbert spaces are infinite dimensional. When attempting to understand various concepts and results on infinite dimensional spaces, it is always helpful to ask oneself what this corresponds to in a finite dimensional setting such as $\mathbb{R}^n$. In many situations we attempt to point out the analogous results in $\mathbb{R}^n$.

We remark that this chapter is by no means a complete exposition of the topic; rather, it is merely intended to prepare the reader for subsequent chapters. For a more detailed exposition of the topics in Section 2.1-2.3, one may consult any functional analysis text; e.g., see [Schechter], [Kreysig], [Yoshida].

2.1 Linear spaces

{1

The goal of this section is to recall some basic definitions for *linear* or *vector spaces*,[7] inner products, and norms and specify some of the notation we use throughout the book. For simplicity of exposition, we only consider real linear spaces.

{la_defn_vectorspace} **Definition 2.1.** *A* **linear space** *(or* **vector space***) V is a set of objects on which two operations are defined;*[8] *the first determines the sum of two elements belonging to V and the second determines the product of any scalar (a real number) α and any element of V. These sum and product operations must satisfy the following properties:*

(i) $u + v \in V$ *for all* $u, v \in V$*;*

(ii) $u + v = v + u$ *for all* $u, v \in V$*;*

(iii) $u + (v + w) = (u + v) + w$ *for all* $u, v, w \in V$*;*

(iv) *there is an element* $0 \in V$ *such that* $u + 0 = u$ *for all* $u \in V$*;*

(v) *for each* $u \in V$ *there exists an element* $(-u) \in V$ *such that* $u + (-u) = 0$*;*

(vi) $\alpha u \in V$ *for each scalar* α *and all* $u \in V$*;*

(vii) $1u = u$ *for all* $u \in V$*;*

(viii) $\alpha(u + v) = \alpha u + \alpha v$ *for all scalars* α*, and for all* $u, v \in V$*;*

(ix) $(\alpha + \beta)u = \alpha u + \beta u$ *for all scalars* α, β *and for all* $u \in V$*;*

(x) $\alpha(\beta u) = (\alpha\beta)u$ *for all scalars* α, β *and for all* $u \in V$*.*

These axioms are simply the well-known properties satisfied by the set of all vectors in $\mathbb{R}^n$ with the usual definitions for the sum and scalar product operations. However, more general collections of objects such as the set of all continuous functions defined on the interval $[a, b]$ with the usual definitions of sum and product are also linear spaces.

The elements of a linear space V are called *vectors.* An expression of the form

$$\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n \,,$$

where $\alpha_i \in \mathbb{R}$ and $u_i \in V$, $i = 1, \ldots, n$, is called a *linear combination* of the vectors u_i. In the simple example of the previous chapter, we saw that our approximate solution was chosen to be a linear combination of functions which formed a basis for the approximating space. The two underlying properties of a basis are linear independence and spanning. Clearly we can always take a linear combination of m vectors and get the zero vector by choosing all the coefficients to be zero. The

[7] The terms are used interchangeably.

[8] There is also an associated field which we always choose to be the real numbers.

concept of linear independence/dependence characterizes whether this is the only way to get the zero vector. Recall that for m vectors in $\mathbb{R}^n$, this reduces to the question of whether the linear system $A\vec{x} = \vec{0}$ has only the trivial solution; here the m columns of A are the vectors. The question of whether a set of m vectors in $\mathbb{R}^n$ span $\mathbb{R}^n$ reduces to the question of whether $A\vec{x} = \vec{b}$ has a unique solution for any $\vec{b} \in \mathbb{R}^n$.

Definition 2.2. *The set of vectors $\{u_i\}_{i=1}^n$ is called* **linearly dependent** *if there exist real numbers α_i, $i = 1, \ldots, n$, not all of which are zero, such that*

$$\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n = 0\,. \tag{2.1}$$

Otherwise, the set is called **linearly independent***; i.e., the set is linearly independent if the only solution to (2.1) is $\alpha_i = 0$, $i = 1, \ldots, n$.*

Definition 2.3. *A subset of vectors of a finite dimensional vector space V is called a spanning set if every vector belonging to V can be written as a linear combination of the elements of the subset.*

To define a basis for a linear space, we need enough vectors to span the space but not too many so that they are linearly dependent.

Definition 2.4. *If V is a linear space and $S = \{v_1, v_2, \ldots, v_r\}$ is a finite set of vectors in V, then S is called a* **basis** *for V if it is a linearly independent spanning set of V.*

To clarify the difference between a finite dimensional and an infinite dimensional linear space, we make the following definition.

Definition 2.5. *A linear space V is called finite dimensional of dimension n if V contains n linearly independent elements and if any set of $(n+1)$ vectors belonging to V is linearly dependent.*

When posing a discrete weak problem, we use a finite dimensional space so we can generate a basis and hence write our approximating solution as a linear combination of the basis elements. In fact, we usually choose our approximating spaces as finite dimensional *subspaces* of the underlying infinite dimensional space on which the weak problem is posed.

Definition 2.6. *A subset S of a vector space V is called a* **subspace** *of V if $u \in S$ and $v \in S$ implies that $\alpha u + \beta v \in S$ for every $\alpha, \beta \in \mathbb{R}$.*

Example 2.7 Consider the infinite dimensional linear space of all continuous functions defined on $\Omega = [0,1]$ with the usual definition of addition and scalar multiplication; we denote this space as $C^0(\Omega)$. Define the following two subsets of $C^0(\Omega)$

$$S_1 = \{v \in C^0(\Omega)\ :\ v(0) = 0\}$$

and

$$S_2 = \{v \in C^0(\Omega)\ :\ v(0) = 1\}\,.$$

The set S_1 is a subspace of $C^0(\Omega)$ since if we take a linear combination of any two continuous functions that are zero at $x = 0$ then the result is a continuous function that is zero at $x = 0$. However, the set S_2 is not a subspace because if we add two functions which are one at $x = 0$ then the resulting function has the value two at $x = 0$. This will be important to us when we are satisfying inhomogeneous boundary conditions. ■

Mappings or operators on linear spaces play an important role, especially linear mappings.

{la_defn_linearmap} **Definition 2.8.** *A mapping f of a linear space V onto a linear space W , denoted $f : V \to W$, is called a linear mapping or equivalently a linear operator provided*

{la_linear}

$$f(\alpha u + \beta v) = \alpha f(u) + \beta f(v) \quad \forall\, u, v \in V,\ \alpha, \beta \in \mathbb{R}\,. \tag{2.2}$$

The kernel of a mapping $f : V \to W$ is defined to be the set $\{v \in V\ :\ f(v) = 0\}$ and the range is defined to be the set of all $w \in W$ such that there exists a $u \in V$ where $f(u) = w$.

For example, matrix multiplication using an $m \times n$ matrix is a linear map from $\mathbb{R}^n \to \mathbb{R}^m$. The kernel of the mapping is just the null space of the matrix and the range is just the span of the columns of the matrix.

The structure of a general linear space is not rich enough to be of use in analyzing the finite element method. In this section the goal is to build a particular class of linear spaces which have the properties that we need to state and analyze our problems. In particular, we need a distance function or metric to measure the "size" of a vector, such as an error vector. However, to be useful the metric must be defined in such a way that there is a relationship between the algebraic structure of the vector space and the metric. To guarantee this relationship we first introduce the concept of a *norm* which uses the algebraic properties of the space and then we use the norm to define a metric.

{la_sec_norms}

2.1.1 Norms

One familiar distance or metric function is the Euclidean distance formula for measuring the length of a given vector in $\mathbb{R}^n$ or equivalently the distance between two points in $\mathbb{R}^n$. This concept of length of a vector in $\mathbb{R}^n$ can be generalized to include other measures such as the maximum component of a vector in $\mathbb{R}^n$. This generalization is accomplished by introducing the notion of a *norm* on $\mathbb{R}^n$ which is a real-valued function from $\mathbb{R}^n$ to $\mathbb{R}$ satisfying important properties that the Euclidean distance possesses. This concept of a norm can be extended to general linear spaces. A norm on a linear space V can be used to measure the "size" of an element of V, such as the size of an error.

{la_defn_norm} **Definition 2.9.** *A norm on a linear space V is a real-valued function from V to $\mathbb{R}$, denoted by $\|\cdot\|$, such that*

(i) $\|u\| \geq 0$ *for all* $u \in V$ *and* $\|u\| = 0$ *if and only if* $u = 0$*;*

(ii) $\|\alpha u\| = |\alpha| \; \|u\|$ *for all* $u \in V$ *and all* $\alpha \in \mathbb{R}$*;*

(iii) $\|u + v\| \leq \|u\| + \|v\|$ *for all* $u, v \in V$*.*

The last property is known as the *triangle inequality* due to its interpretation in $\mathbb{R}^n$ using the standard Euclidean norm. In the exercises we consider the three most common norms on $\mathbb{R}^n$.

If we relax the first property of a norm to allow $\|u\| = 0$ for $u \neq 0$, but still require properties (ii) and (iii), then we call the resulting function a *semi-norm.* A linear space V equipped with a norm as defined above is called a *normed linear space* so that we think of a normed linear space as a pair $(V, \|\cdot\|)$.

Example 2.10 If we return to our linear space $C^0(\Omega)$ where $\Omega = [0, 1]$ we can define a norm as

$$\|f\| \equiv \max_{x \in [0,1]} |f(x)| \, .$$

Clearly, all three properties of the norm are satisfied. To measure the difference between two vectors, $f, g \in C^0(\Omega)$ we determine

$$\|f - g\| = \max_{x \in [0,1]} |f(x) - g(x)| \, .$$

∎

Since there is always a choice of norms to use on a given vector space, we would like to know if these different measures are somehow comparable.

Definition 2.11. *Two norms, $\|\cdot\|_a$, $\|\cdot\|_b$ defined on a linear space V are said to be equivalent if there are constants C_1, C_2 such that* {l

$$C_1 \|u\|_a \leq \|u\|_b \leq C_2 \|u\|_a \, . \tag{2.3}$$ {l

Of course, if (2.3) holds then we also have

$$\frac{1}{C_2} \|u\|_b \leq \|u\|_a \leq \frac{1}{C_1} \|u\|_b \, .$$

In a course in linear algebra, it is usually proved that all norms on $\mathbb{R}^n$ are equivalent. In the exercises, the actual constants in the equivalence relations for the three standard norms on $\mathbb{R}^n$ are investigated; of course, these constants can depend upon n. In functional analysis, one can show a more general, i.e., that in a finite dimensional vector space all norms are equivalent. For the proof of the result, see [Schechter].

Lemma 2.12. *If V is a finite dimensional normed linear space, then all norms are equivalent.* {l

2.1.2 Inner products

{l

Recall that in dealing with vectors in $\mathbb{R}^n$, one defines a scalar product of two vectors $\vec{a} = (a_1, a_2, \cdots, a_n)$ and $\vec{b} = (b_1, b_2, \cdots, b_n) \in \mathbb{R}^n$ as

$$(\vec{a}, \vec{b}) = \vec{a}^T \vec{b} = \sum_{i=1}^{n} a_i b_i \,.$$

The result of the scalar product is just a number so it can be viewed as a function from $\mathbb{R}^n \to \mathbb{R}$. The scalar product is useful in many applications such as determining if two vectors are perpendicular or equivalently, orthogonal. This concept can be generalized to elements of a linear space in the following manner.

{la_defn_innerproduct} **Definition 2.13.** *An* **inner product** *or* **scalar product** *on a (real) linear space V is a real-valued function from V to $\mathbb{R}$, denoted by $(\cdot,\cdot)$, satisfying*

(i) $(u,u) \geq 0$ *for all* $u \in V$ *and* $(u,u) = 0$ *if and only if* $u = 0$*;*

(ii) $(u,v) = (v,u)$ *for all* $u, v \in V$*;*

(iii) $(\alpha u + \beta v, w) = \alpha\,(u,w) + \beta\,(v,w)$ *for all* $u, v, w \in V$ *and all* $\alpha, \beta \in \mathbb{R}$*.*

A vector space V equipped with an inner product is aptly called an *inner product space.*

Example 2.14 Returning to our example $C^0(\Omega)$ of a linear space we can define an inner product as

$$(f,g) = \int_0^1 f(x) g(x)\, dx \,.$$

The three properties of the inner product are easily shown to be satisfied by using the properties of integrals. See exercises. ■

Analogous to the case of $\mathbb{R}^n$, we say that two vectors in an inner product space are orthogonal if their scalar product is zero.

Definition 2.15. *Let V be an inner product space. Then $u, v \in V$ are orthogonal if and only if*

{la_orthogonality}

$$(u,v) = 0 \,. \tag{2.4}$$

One can use the inner product to define a norm for a vector space. Indeed, if we let $\|v\| = (v,v)^{1/2}$ for all $v \in V$, one can readily show that this defines a norm on V; see the exercises for details. We refer to a norm defined in this manner on an inner product space as the *induced norm.*

To complete this section we present an inequality for inner product spaces which is extremely useful. Recall that the scalar product of two vectors in $\mathbb{R}^n$ can also be written as $(\vec{a}, \vec{b}) = \|\vec{a}\|\,\|\vec{b}\| \cos\theta$ where $\|\cdot\|$ denotes the standard Euclidean norm and θ is the angle between the two vectors. Obviously this says that $(\vec{a},\vec{b}) \leq \|\vec{a}\|\,\|\vec{b}\|$. The Cauchy-Schwarz inequality generalizes this result to an inner product space.

{la_cs} **Lemma 2.16.** *Let V be an inner product space. The Cauchy-Schwarz inequality is given by*

$$(u,v) \leq (u,u)^{\frac{1}{2}}\,(v,v)^{\frac{1}{2}} \quad \forall\, u,v,\in V\,. \tag{2.5}$$ {1

If $\|\cdot\|$ denotes the induced norm on V then this inequality can also be written as

$$(u,v) \leq \|u\|\,\|v\| \quad \forall\, u,v,\in V\,. \tag{2.6}$$ {1

Proof. To verify (2.5), we first note that it is trivially satisfied if $u=0$ or $v=0$ so we consider the case where $u,v \neq 0$. By the first property of inner products, we know that $(u-\alpha v, u-\alpha v) \geq 0$ for any $\alpha \in \mathbb{R}$. Using the linearity property of the inner product we rewrite this as

$$\begin{aligned} 0 \leq (u-\alpha v, u-\alpha v) &= (u,u) - 2\alpha(u,v) + \alpha^2(v,v) \\ &= (u,u) - \alpha(u,v) - \alpha\Big[(u,v) - \alpha(v,v)\Big]\,. \end{aligned}$$

Now the term in brackets is zero if we choose $\alpha = (u,v)/(v,v)$. Note that this is possible since we are considering the case $v \neq 0$. Thus

$$(u,u) - \frac{(u,v)^2}{(v,v)} \geq 0$$

and simplification yields the Cauchy-Schwarz inequality (2.5). The second form of the inequality given in (2.6) follows directly from the definition of the norm on V induced by the scalar product.

2.1.3 Topological concepts

{1

One of our goals in analyzing the finite element method is to determine the error between the solution of the discrete weak problem and the solution of the continuous weak problem. We can use the concept of norm introduced in the last section to measure the distance between these two solutions. For a normed vector space V, we define the *distance* ρ between two vectors u and v as $\rho(u,v) = \|u-v\|$.

In discretizing a problem, we expect to have a sequence of solutions which are generated by using successively finer meshes. We expect that these solutions converge, in some sense, to the solution of the continuous problem. We now make precise what this means.

Definition 2.17. *A sequence of vectors $u_1, u_2, u_3, \ldots$ belonging to a normed linear space V is called* **convergent** *if there exists a vector $u \in V$ such that given any $\epsilon > 0$, there exists a postive integer $N = N(\epsilon)$ such that* {1

$$\|u_n - u\| < \epsilon \quad \forall\, n \geq N\,.$$

We call u the limit of the sequence $\{u_i\}_{i\geq 1}$ and write

$$\lim_{n\to\infty} u_n = u \quad or \quad u_n \to u \text{ in } V \text{ as } n\to\infty\,.$$

It can be shown that a convergent sequence has only one limit and $u_n \to u$ in V if and only if $\|u_n - u\| \to 0$ as $n \to \infty$.

An important tool in analysis is the Cauchy sequence. If we use Definition 2.17 to show that a sequence is convergent, then we need to know its limit. However, sometimes we don't know the actual limit or it may not even be in our linear space. Oftentimes the important issue is that a sequence converges rather than what its limit is. A Cauchy sequence is one in which its terms ultimately become arbitrarily close. In fact, we can discount a finite number of terms at the beginning of the sequence and then guarantee that any two of the remaining terms are closer than some prescribed value.

{la_defn_cauchyseq} **Definition 2.18.** *A sequence of vectors $\{u_i\}_{i\geq 1}$, $\{u_i\}\in V$, is called a* **Cauchy sequence** *if, given any $\epsilon > 0$, there exists an integer $N = N(\epsilon)$ such that*

$$\|u_n - u_m\| \leq \epsilon \quad \forall\, m, n \geq N\,.$$

Here $\|\cdot\|$ defines a norm on a normed linear space V.

Every convergent sequence is clearly a Cauchy sequence since

$$\|u_m - u_n\| = \|(u_m - u) + (u - u_n)\| \leq \|u - u_m\| + \|u - u_n\|$$

and we can make the right-hand side arbitrarily small as $m, n \to \infty$. However, the converse is not always true as the following example illustrates.

Example 2.19 The Weierstrass Approximation Theorem states that a continuous function defined on $[a,b]$ can be uniformly approximated as closely as desired by a polynomial defined on $[a,b]$. More precisely, suppose f is an arbitrary continuous function defined on $[a,b]$. For every $\epsilon > 0$, there exists a polynomial function $p(x)$ such that $\max_{x\in[a,b]} |f(x) - p(x)| < \epsilon$. Thus we can construct a sequence of polynomials in the linear space of polynomials defined on $[a,b]$ which form a Cauchy sequence using the max-norm but its limit is not a polynomial. ■

We would like to avoid this situation by imposing on our space of functions the property that every Cauchy sequence in V has a limit in V. In addition to properties (i)–(x) which characterize linear spaces, we would like to add the property of *completeness*, i.e.,

(xi) if $\{v_n\}$ is a sequence of elements in V such that $\|v_n - v_m\| \to 0$ as $m, n \to \infty$, then there exists an element $v \in V$ such that $\|v - v_n\| \to 0$ as $n \to \infty$.

Another way to state property (xi) is to require that every Cauchy sequence in V is convergent.

A *complete normed vector space*, i.e., a collection of objects satisfying properties (i)–(xi) with a norm defined on the space, is of such importance that it is given a special name: a *Banach space*. Euclidean n-dimensional space is the most familiar example of a Banach space. Given any (noncomplete) normed space S it can be proved that by adding new elements, S can be extended to a complete normed space (a Banach space), V. This process is referred to as the *completion* of S or the *closure* of S in V.

Since we work with finite dimensional subspaces when we discretize, we often have sequences on these subspaces and need to know if their limit is in the subspace.

Definition 2.20. *A subset S of a Banach space V is said to be a* **closed subspace** *of V if it is a subspace of V with the property that whenever $\{u_i\}_{i\geq 1}$ is a convergent sequence in V such that $u_i \in S$, $i = 1, 2, \ldots,$ then $u = \lim_{n\to\infty} u_n$ belongs to S also.* {1

It can be shown that every finite dimensional subspace is closed; this is important for us since our approximating spaces are finite dimensional.

Our search for the appropriate function spaces to use in analyzing the finite element method is almost at an end. In the next section we add a final property to our complete, normed linear space, that of an inner product.

2.1.4 Hilbert spaces

{1

A complete inner product space is called a *Hilbert space*; these spaces extend the ideas of the Euclidean space $\mathbb{R}^n$ to infinite dimensional spaces. For example, the parallelogram law in $\mathbb{R}^2$ states that the the sum of the squares of the lengths of the two diagonals in a parallelogram equals the sum of the squares of the lengths of the four sides. This law can be shown to hold in all Hilbert spaces and is written as

$$\|f+g\|^2 + \|f-g\|^2 = 2\big(\,\|f\|^2 + \|g\|^2\,\big)\,, \tag{2.7}$$ {1

where $\|\cdot\|$ denotes the induced norm and f, g are any elements of the Hilbert space. See the exercises for a proof of this result.

Clearly, every Hilbert space is a Banach space; one simply uses the norm induced by the inner product, i.e., $\|v\| = (v,v)^{1/2}$. However, the converse is not true. A standard counterexample is to consider the Banach space of all bounded linear functions with the uniform or max norm. In this example, one can demonstrate that the parallelogram law fails to hold so it can not also be a Hilbert space; see the exercises for details. The most commonly used spaces of admissible test and trial functions for weak formulations of boundary value problems for partial differential equations are Hilbert spaces.

Example 2.21 An example of a Hilbert space that is central to our discussions is $L^2(\Omega)$ where Ω denotes an open, connected subset of $\mathbb{R}^n$. To construct this space, we consider the set of real-valued, continuous functions $u(\mathbf{x}) = u(x_1, x_2, \ldots, x_n)$ defined on Ω where $(x_1, x_2, \ldots, x_n)$ denotes a point in $\mathbb{R}^n$. Addition and scalar multiplication are defined in the usual manner. We define an inner product as

$$(u, v) = \int_\Omega u(\mathbf{x})v(\mathbf{x})\, d\Omega\,, \tag{2.8}$$

{1

where $d\Omega$ is the volume element in Ω. Clearly this satisifes all the properties of an inner product given in Definition 2.13. In order to guarantee that the integral defining this inner product exists, we restrict our attention to functions $u(\mathbf{x})$ on Ω such that

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

We now define S to be the space described above; i.e.,

$$S = \{u \mid u = u(\mathbf{x}), u(\mathbf{x}) \text{ is continuous for all } \mathbf{x} \in \Omega \text{ and } \int_\Omega |u(\mathbf{x})|^2\, d\Omega < \infty\}\,.$$

Then S is an inner product space with the inner product defined by (2.8). The norm on S is given by

$$\|u\| = (u, u)^{1/2} = \left(\int_\Omega |u(\mathbf{x})|^2\, d\Omega\right)^{1/2}\,.$$

In general, the space S is not complete. For example in $\mathbb{R}$, let $\Omega = (-1, 1)$ and let S be defined as above. Consider the sequence $u_1(x), u_2(x), \cdots$ where

$$u_j(x) = \begin{cases} -1 & \text{for} \quad -1 < x \le -1/j \\ jx & \text{for} \quad -1/j \le x \le 1/j \\ 1 & \text{for} \quad 1/j < x < 1\,. \end{cases}$$

It is straightforward to show that $\{u_j(x)\}$, $j = 1, 2, \cdots$, is a Cauchy sequence in S. Moreover, the sequence converges to the discontinuous function $f(x)$ where

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

However, $f(x) \notin S$ and so there is no continuous function $u(x)$ on $(-1, 1)$ for which $\|u_n - u\| \to 0$ as $n \to \infty$. By adding new elements to S we can complete the space to form a Hilbert space V. These additional functions may be piecewise continuous, but, in general, are highly discontinuous. This extended *complete* space V is called $L^2(\Omega)$ which is a complete inner product space, i.e., a Hilbert space. ■

Remark The space $L^2(\Omega)$ is really a special case of the Banach space of functions on Ω which are p-integrable denoted $L^p(\Omega)$, $p \ge 1$, The norm is given by

$$\|u\|_{L^p} = \left(\int_\Omega |u|^p\, d\Omega\right)^{1/p}$$

for $1 \leq p < \infty$ and for $p = \infty$

$$\|u\|_{L^\infty} = \sup_\Omega |u| \,.$$

Remark In a manner analogous to the construction of $L^2(\Omega)$, we can construct a weighted L^2-space. Given a weight $w(\mathbf{x})$, integrable on Ω, we define the inner product as

$$(u,v) = \int_\Omega u(\mathbf{x})v(\mathbf{x})w(\mathbf{x})\, d\Omega \,.$$

We denote this space $L^2(\Omega; w)$.

2.2 Best approximations

{1

In this section we want to investigate some geometric properties of Hilbert spaces. A central idea in approximation theory is to determine an element of a subspace of a given vector space which is closest (with respect to the given metric) to a particular element of the vector space; that is, to find the *best approximation* of the given vector in the subspace. (In fact, this is the basis for least squares methods.) We would like to know when it is possible to assert in advance that a best approximating element exists. Moreover, we want to know whether this best approximating element is unique.

{1

Example 2.22 Consider the situation illustrated in Figure 2.1. Here we assume that we have a given plane S in $\mathbb{R}^3$, a vector $\mathbf{u} \notin S$, and we want to find a vector $\mathbf{s}$ in S that is nearest $\mathbf{u}$; i.e., $\|\mathbf{u} - \mathbf{s}\| \leq \|\mathbf{u} - \phi\|$ for all $\phi \in S$ where we are using the standard Euclidean norm. Clearly in this case, there is a a unique $\mathbf{s}$ and it is found by drawing a perpendicular from $\mathbf{u}$ to S; that is, *projecting* the vector $\mathbf{u}$ onto S. Also, we can uniquely write the vector $\mathbf{u}$ as the sum of the vector $\mathbf{s} \in S$ and a vector not in S. ■

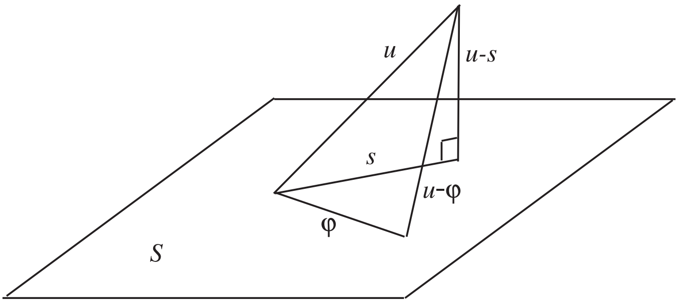


Figure 2.1. {1

There are analogous results for general Hilbert spaces. However, they do not hold for a general subspace but only a *closed subspace* (see Definition 2.20). In the finite element setting, we are guaranteed that the subspace is closed since it is finite dimensional.

The following result is known as the *Projection Theorem* and it states that given an element in a Hilbert space, its orthogonal projection onto a closed subspace is the element of the subspace that is "nearest" the given vector where the distance is measured using the induced norm. Of course one must keep in mind that this depends upon the choice of the inner product (and thus the norm) on the given Hilbert space.

{la_thm_projection} **Theorem 2.23. (The Projection Theorem)** *Let S be a closed subspace of a Hilbert space V which is not the whole of V. Then given $u \in V$ there exists a unique element $Pu \in S$ such that*

{la_proj1}
$$\|u - Pu\| = \inf_{\phi \in S} \|u - \phi\| \tag{2.9}$$

where Pu satisfies

{la_proj2}
$$(u - Pu, \phi) = 0 \qquad \text{for every } \phi \in S\,. \tag{2.10}$$

Proof. Let $u \in V$ such that $u \notin S$. We know that for all $\phi \in S$, $\|u - \phi\| > 0$ so that if we define the distance from u to S as the lower bound

$$\delta = \inf_{\phi \in S} \|u - \phi\| \ ,$$

then there exists a sequence $\phi_n \in S$ such that $\|u - \phi_n\| \to \delta$ as $n \to \infty$. Our goal is so show that $\{\phi_n\}$ is a Cauchy sequence in S and thus conclude that the limit of the sequence is also in S because S is a closed subspace of the Hilbert space V; we call Pu this limit.

To demonstrate that $\{\phi_n\}$ is a Cauchy sequence we apply the parallelogram law (2.7)

with $f = u - \phi_m$ and $g = u - \phi_n$. We have

$$\|(u - \phi_m) + (u - \phi_n)\|^2 + \|(u - \phi_m) - (u - \phi_n)\|^2 = 2\,\|u - \phi_m\|^2 + 2\,\|u - \phi_n\|^2$$

and simplifying the left-hand side gives

{la_proj3}
$$4\left\|\left(u - \frac{\phi_m + \phi_n}{2}\right)\right\|^2 + \|\phi_n - \phi_m\|^2 = 2\,\|u - \phi_m\|^2 + 2\,\|u - \phi_n\|^2 \tag{2.11}$$

because

$$\|(u - \phi_m) + (u - \phi_n)\|^2 = \|2u - \phi_m - \phi_n\|^2 = 4\left\|u - \frac{1}{2}\left(\phi_m + \phi_n\right)\right\|^2 .$$

Now $\phi_n, \phi_m \in S$ and S is a subspace of V, so we have that $(\phi_m + \phi_n)/2 \in S$ and thus by the definition of δ, the first term on the left of (2.11) is nonnegative and at least as large as $4\delta^2$. Thus

$$\|\phi_n - \phi_m\|^2 \leq 2\,\|u - \phi_m\|^2 + 2\,\|u - \phi_n\|^2 - 4\delta^2\,.$$

We conclude that because $\|\phi_m - u\| \to \delta$, $\|\phi_n - u\| \to \delta$ as $m, n \to \infty$, the right-hand side goes to zero as $m, n \to \infty$, and hence $\{\phi_n\}$ is a Cauchy sequence in a closed subspace and therefore convergent.

Uniqueness of the limit $s \in S$ is proved in the standard way by assuming there are two elements in S which satisfy (2.9). Let s_1 and s_2 have the property that

$$\delta = \inf_{\phi \in S} \|u - \phi\| = \|u - s_1\| = \|u - s_2\| \, .$$

Then because $(s_1 + s_2)/2 \in S$, we have that

$$\delta \le \left\| u - \frac{1}{2}(s_1 + s_2) \right\| \le \frac{1}{2} \|u - s_1\| + \frac{1}{2} \|u - s_2\| = \delta$$

where we have used the triangle inequality for the last inequality. This implies that for $u - s_1$ and $u - s_2$ the triangle inequality must hold as an equality. However, this can only be true if $u - s_1 = \alpha(u - s_2)$ for some α. If we choose $\alpha = 1$ then this leads to a contradiction because it would imply $s_1 = s_2$; if $\alpha \neq 1$ then we have $(\alpha - 1)u = \alpha s_2 - s_1$ and this contradicts the fact that $u \notin S$. Consequently we have uniqueness.

To prove (2.10) we assume that there is some $\hat{\phi} \in S$ such that $(u - Pu, \hat{\phi}) \neq 0$ and show that this assumption leads to the existence of an $s \in S$ such that $\|u - s\| < \inf_{\phi \in S} \|u - \phi\|$; thus we obtain a contradiction. Let $s \in S$ be given by

$$s = Pu + \frac{\left(u - Pu, \hat{\phi}\right)}{\left(\hat{\phi}, \hat{\phi}\right)} \hat{\phi} \, .$$

Then

$$\begin{aligned} \|u - s\|^2 &= \left(u - Pu - \frac{\left(u - Pu, \hat{\phi}\right)}{\|\hat{\phi}\|^2} \hat{\phi}, u - Pu - \frac{\left(u - Pu, \hat{\phi}\right)}{\|\hat{\phi}\|^2} \hat{\phi} \right) \\ &= \|u - Pu\|^2 - \frac{2}{\|\hat{\phi}\|^2} \left(u - Pu, \hat{\phi}\right) \left(u - Pu, \hat{\phi}\right) + \frac{1}{\|\hat{\phi}\|^4} \left(u - Pu, \hat{\phi}\right)^2 \|\hat{\phi}\|^2 \\ &= \|u - Pu\|^2 - \frac{1}{\|\hat{\phi}\|^2} \left(u - Pu, \hat{\phi}\right)^2 \, . \end{aligned}$$

Because our assumption was that $(u - Pu, \hat{\phi}) \neq 0$, we have

$$\|u - s\| < \|u - Pu\| = \inf_{\phi \in S} \|u - \phi\|$$

which is the contradiction we sought. ■

In Example 2.22, we wrote the vector which we projected into the subspace as the sum of a vector in the subspace and one orthogonal to the subspace. Theorem 2.23 guarantees that we can do this in a Hilbert space when we are projecting onto a closed subspace. Given u in a Hilbert space V then the unique Pu in a closed subspace S guaranteed by Theorem 2.23 is called the *orthogonal projection* of u onto the closed subspace S. In this case, if we let $r = u - Pu$, then we can write

$$u = Pu + r \quad \text{where } Pu \in S \text{ and } (r, \phi) = 0 \text{ for all } \phi \in S\,. \tag{2.12}$$

Hence the vector $r = u - Pu$ is *orthogonal* to all vectors in S; we call the set of all vectors orthogonal to S the *orthogonal complement* of S and denote it by $S^{\perp}$. Thus we have written u as the sum of an element in S and one orthogonal to S, i.e., in $S^{\perp}$. In the exercises, we explore the fact that the analogous result holds for the entire Hilbert space; that is, the Hilbert space V can be written as the direct sum of S and $S^{\perp}$.

Another interpretation of the vector s guaranteed by Theorem 2.23 is given by (2.9). From this equation we call Pu the *best approximation* of u in S. In the case when S is a finite dimensional subspace of a Hilbert space V (and thus automatically closed) then we can explicitly construct the best approximation to a given vector in V by using (2.10). To see this, we let ϕ_i, $i = 1, \ldots, m$ be a basis for S and from (2.10), $(u - Pu, \phi_i) = 0$ for $i = 1, \ldots, m$. Also Pu can be written as a linear combination of the basis elements; i.e., $Pu = \sum_{j=1}^m c_j \phi_j$. Thus

{la_bestapprox}

$$\sum_{j=1}^{m} c_j\,(\phi_j, \phi_i) = (u, \phi_i)\,, \quad i = 1, \ldots, m\,, \tag{2.13}$$

which is just a linear system for the unknowns c_j, $j = 1, \ldots, m$. The matrix $\mathcal{G}$ whose entries are given by $\mathcal{G}_{ij} = (\phi_j, \phi_i)$ is known as the *Gram matrix* associated with the basis functions $\{\phi_i\}$ of S and is guaranteed to be nonsingular. This can be easily seen by assuming that if $\mathcal{G}$ is singular, then we could find a vector $d = (d_1, d_2, \ldots, d_m)$ such that $\mathcal{G}d = 0$. This would imply $\left(\sum_{j=1}^m d_j \phi_j, \phi_i\right) = 0$ for all $i = 1, \ldots, m$ and thus the vector $u = \sum_{j=1}^m d_j \phi_j$ would be orthogonal to S which is a contraction.

In the following example we investigate the effect that the choice of the approximating subspace for the best approximation has on the properties of the Gram matrix.

Example 2.24 Consider the problem of determining the best approximation to a function $f(x)$ in two different subspaces of $L^2(0,1)$. We first consider the subspace consisting of all polynomials of degree three or less. In this case an obvious choice of a basis is $\{1, x, x^2, x^3\}$. The specific system we must solve is

$$\begin{pmatrix} \int_0^1 \,dx & \int_0^1 x\,dx & \int_0^1 x^2\,dx & \int_0^1 x^3\,dx \\ \int_0^1 x\,dx & \int_0^1 x^2\,dx & \int_0^1 x^3\,dx & \int_0^1 x^4\,dx \\ \int_0^1 x^2\,dx & \int_0^1 x^3\,dx & \int_0^1 x^4\,dx & \int_0^1 x^5\,dx \\ \int_0^1 x^3\,dx & \int_0^1 x^4\,dx & \int_0^1 x^5\,dx & \int_0^1 x^6\,dx \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} \int_0^1 f(x)\,dx \\ \int_0^1 x f(x)\,dx \\ \int_0^1 x^2 f(x)\,dx \\ \int_0^1 x^3 f(x)\,dx \end{pmatrix}.$$

Upon performing the integration the system becomes

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} \int_0^1 f(x)\, dx \\ \int_0^1 xf(x)\, dx \\ \int_0^1 x^2 f(x)\, dx \\ \int_0^1 x^3 f(x)\, dx \end{pmatrix} .$$

It is important to note that in this case the Gram matrix is not sparse and is the well-known Hilbert matrix which is notoriously ill-conditioned. Consequently, if we look for the best approximation to a function $f(x)$ out of the space of polynomials of degree n or less, then for even modest values of n our solution is not reliable using standard matrix solvers.

On the other hand, if we choose as a subspace of $L^2(0,1)$ the space of continuous piecewise linear functions over the uniform partition of $[0,1]$ into N subintervals and choose as a basis the piecewise linear "hat" functions described in the previous chapter, then the resulting matrix is well-conditioned and is tridiagonal. To see the structure of the matrix consider the uniform partition of $[0,1]$ into 3 subintervals with four grid points $x_1 = 0$, $x_2 = \frac{1}{3}$, $x_3 = \frac{2}{3}$, $x_4 = 1$. Let $\phi_i(x)$, $i = 1, \ldots, 4$ denote the basis functions

$$\phi_1 = \begin{pmatrix} 1-3x & 0 \le x \le \frac{1}{3} \\ 0 & \text{elsewhere} \end{pmatrix}, \qquad \phi_2 = \begin{pmatrix} 3x & 0 \le x \le \frac{1}{3} \\ 2-3x & \frac{1}{3} \le x \le \frac{2}{3} \\ 0 & \text{elsewhere} \end{pmatrix},$$

$$\phi_3 = \begin{pmatrix} 3x-1 & \frac{1}{3} \le x \le \frac{2}{3} \\ 3-3x & \frac{2}{3} \le x \le 1 \\ 0 & \text{elsewhere} \end{pmatrix}, \qquad \phi_4 = \begin{pmatrix} 3x-2 & 2/3 \le x \le 1 \\ 0 & \text{elsewhere} \end{pmatrix}.$$

The Gram matrix has the (i,j) entry given by $\int_0^1 \phi_i(x)\phi_j(x)\, dx$. Due to the fact that the basis function $\phi_i(x)$ has local support, we deduce that the Gram matrix is

$$\begin{pmatrix} \int_0^1 \phi_1^2\, dx & \int_0^1 \phi_2\phi_1\, dx & 0 & 0 \\ \int_0^1 \phi_1\phi_2\, dx & \int_0^1 \phi_2\phi_2\, dx & \int_0^1 \phi_3\phi_2\, dx & 0 \\ 0 & \int_0^1 \phi_2\phi_3\, dx & \int_0^1 \phi_3\phi_3\, dx & \int_0^1 \phi_4\phi_3\, dx \\ 0 & 0 & \int_0^1 \phi_3\phi_4\, dx & \int_0^1 \phi_4\phi_4\, dx \end{pmatrix},$$

which is a symmetric tridiagonal matrix. ■

Being able to find the best approximation to a given function using piecewise polynomials does not directly help us to find our finite element approximation. This is because in order to use (2.13) to find the best approximation to u, we need to know u, which in our case is the unknown solution to the weak problem. However, what we see in the next chapter is that when we measure the error in our finite element approximation, it will be bounded by a constant times the error in the best approximation in the approximating space.

{la_sec_blf}

2.3 Bounded linear functionals

Functional is just the name given to a special type of function which assigns a number to each element of a linear space. For example, for functions in $L^2((0,1))$ the integral over the domain is a functional. If V is a given Hilbert space with inner product $(\cdot,\cdot)$ and induced norm $\|\cdot\|$ then $\|v\|$ assigns a number to each element v in V and is thus a functional. If we fix an element u in V, then (v,u) assigns a value (i.e., a scalar) to each element. Such mappings are called *functionals.* Functionals which are linear and bounded are of particular interest.

{la_def_blf} **Definition 2.25.** *F is a functional on a Hilbert space V if it assigns to every $v \in V$ a unique number $F(v)$ and we write $F : V \to \mathbb{R}$. A functional is called linear if for every $u, v \in V$ and scalars α, β we have*

$$F(\alpha u + \beta v) = \alpha F(u) + \beta F(v)\,. \tag{2.14}$$

In addition, we say that a functional is bounded if

$$\sup_{v \in V} \frac{|F(v)|}{\|v\|} < \infty,\ v \neq 0\,, \tag{2.15}$$

where $\|\cdot\|$ is the induced norm on V. We call this finite number $\|F\|$.

We note that if F is a bounded linear functional on V then this is equivalent to saying that F is a linear functional which is a continuous function of its arguments.

Example 2.26 Let V be a Hilbert space; for a fixed $u \in V$ the inner product $F(v) = (v,u)$ denotes a bounded linear functional on V. Clearly it defines a functional and is linear because of the linearity of the inner product. Specifically, we have

$$F(\alpha v+\beta w) = (\alpha v+\beta w, u) = (\alpha v, u)+(\beta w, u) = \alpha(v,u)+\beta(w,u) = \alpha F(v)+\beta F(w)\,.$$

Boundedness follows from using the Cauchy-Schwarz inequality

$$F(v) = (v,u) \leq \|v\|\,\|u\|\ .$$

to obtain

$$\frac{|F(v)|}{\|v\|} \leq \|u\| < \infty \quad \forall\, v \in V\,,$$

for $v \neq 0$. ∎

In a similar manner, the linearity of the norm on a Hilbert space can be used to demonstrate that the norm is a bounded linear functional. Clearly, we can think of many other examples of bounded linear functionals, but what is surprising is that the inner product is really the only one on a Hilbert space; i.e., every bounded linear functional can be written as an inner product. This result is known as the Riesz Representation Theorem and is named after the Hungarian mathematician Frigyes Riesz (1880-1956).

Theorem 2.27. (Riesz Representation Theorem) *For every bounded linear functional F on a Hilbert space V there is a unique element $f \in V$ such that* {1

$$F(v) = (v, f) \quad \textit{for all } v \in V\,. \tag{2.16}$$

Moreover, $\|F\| = \|f\|$.

Proof. We first note that if F assigns to each $v \in V$ the value zero, then the proof is immediate by taking $f = 0$. In the sequel we assume that this is not the case. However, we do know that for each $v \in V$ which F assigns to zero we must have that f is orthogonal to it; i.e., $(v, f) = 0$. We call the set of all vectors v such that $F(v) = 0$ the *kernel* of F and denote it by $\mathcal{K}(F)$. Hence we must construct an f that is orthogonal to the kernel of F.

We first demonstrate that $\mathcal{K}(F)$ is a closed subspace of V. To show that it is a subspace we use the linearity of F; i.e., if $u, v \in \mathcal{K}(F)$ then

$$F(\alpha u + \beta v) = \alpha F(u) + \beta F(v) = 0\,.$$

To show that it is a closed subspace of V, we let $\{u_n\}$ be a sequence in $\mathcal{K}(F)$ such that $u_n \to u \in V$ as $n \to \infty$ and show that $u \in \mathcal{K}(F)$. We have

$$|F(u)| = |F(u) - F(u_n)| = |F(u - u_n)| \leq \|F\|\,\|u - u_n\|\;,$$

where we have used the fact that $u_n \in \mathcal{K}(F)$, the linearity of F, and the definition of the norm of a bounded linear functional. The right-hand side of this inequality goes to zero as $n \to \infty$ so that $F(u) = 0$.

We now proceed to construct an f that is orthogonal to $\mathcal{K}(F)$. From the comments following the projection theorem we know that we can write V as the direct sum of $\mathcal{K}(F)$ and $\mathcal{K}(F)^\perp$ since $\mathcal{K}(F)$ is a closed subspace of V. Our strategy is to take an arbitrary $\hat{f} \in \mathcal{K}(F)^\perp$, $\hat{f} \neq 0$ and construct an $f \in \mathcal{K}(F)$ using $\hat{f}$. Consider the vector $F(v)\hat{f} - F(\hat{f})v$. This vector is in $\mathcal{K}(F)$ since

$$F\Big(F(v)\hat{f} - F(\hat{f})v\Big) = F(v)F(\hat{f}) - F(\hat{f})F(v) = 0$$

and thus $\Big(F(v)\hat{f} - F(\hat{f})v, \hat{f}\Big) = 0$ for all $\hat{f} \in \mathcal{K}(F)^\perp$, $v \in V$. Therefore we have $F(v)\|\hat{f}\|^2 = F(\hat{f})(v, \hat{f})$ so that

$$F(v) = \left(v, \frac{F(\hat{f})}{\|\hat{f}\|^2}\hat{f}\right)\,.$$

Hence if we set $f = (F(\hat{f})/\|\hat{f}\|^2)\hat{f}$, we see that for each $v \in \mathcal{K}(F)$, this choice of f is orthogonal to v and we have the desired result.

To show uniqueness of f we assume that there are two vectors f_1 and f_2 such that

$$F(v) = (v, f_1) = (v, f_2) \qquad \forall\, v \in V\,.$$

But this implies that $(v, f_1 - f_2) = 0$ for all $v \in V$; specifically set $v = f_1 - f_2$ from which it follows that $f_1 - f_2 = 0$.

Lastly, we must demonstrate that $\|F\| = \|f\|$. This follows immediately from the definition of F and the Cauchy-Schwarz inequality. We have

$$|F(v)| = |\,(v, f)\,| \leq \|v\|\,\|f\|$$

so that if $v \neq 0$,

$$\frac{|F(v)|}{\|v\|} \leq \|f\|$$

and thus $\|F\| \leq \|f\|$. On the other hand, since $f \in V$, $F(f) = (f, f) = \|f\|^2$. Thus the supremum is attained at $v = f$ and we have equality. ■

The main goal of this chapter was to introduce the mathematical tools necessary to formulate and analyze a general weak problem. The last tool we need is a bilinear form. We define a *bilinear form* on a Hilbert space V to be a map from $V \times V$ into $\mathbb{R}^1$, denoted by $B(\cdot,\cdot)$, such that

$$\begin{aligned} B(\alpha_1 u_1 + \alpha_2 u_2, v) &= \alpha_1 B(u_1, v) + \alpha_2 B(u_2, v) \\ B(u, \beta_1 v_1 + \beta_2 v_2) &= \beta_1 B(u, v_1) + \beta_2 B(u, v_2) \end{aligned}$$

for all $u_i, v_i \in V$ and $\alpha_i, \beta_i \in \mathbb{R}^1$, $i = 1, 2$. That is, $B(\cdot,\cdot)$ is linear in each of its components. An example of a bilinear form on $L^2(0,1)$ is $\int_0^1 u(x)v(x)\,dx$. In fact, any inner product on a Hilbert space defines a bilinear form; this can easily be seen from the linearity of the inner product (see Definition 2.13). Another example of a bilinear form on a Hilbert space V is $(\mathcal{B}u, v)_V$ where $\mathcal{B}$ is a linear operator from V to V.

We say that a bilinear form $B(\cdot,\cdot)$ on V is bounded if there exists a positive constant C such that

$$|B(u, v)| \leq C\,\|u\|_V\,\|v\|_V\,.$$

If we fix an element $u \in V$ then the bilinear form $B(u, v)$ represents a linear functional on V; if $B(\cdot,\cdot)$ is bounded, then for a fixed $u \in V$, $B(u, v)$ represents a bounded linear functional $F(v)$ on V. The Riesz Representation Theorem 2.27 then guarantees that there exists a unique element $\hat{u} \in V$ such that $B(u, v)$ can be written as the inner product $(v, \hat{u})$. The ability to associate to each $u \in V$ a unique element $\hat{u}$ is central to our analysis of an abstract weak problem.

Exercises

{1

2.1. The most common examples of norms on $\mathbb{R}^n$ are the Euclidean norm defined by

$$\|\mathbf{x}\|_{\ell_2} = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} ,$$

the sum norm defined by

$$\|\mathbf{x}\|_{\ell_1} = \left(\sum_{i=1}^{n} |x_i| \right) ,$$

and the maximum norm defined by

$$\|\mathbf{x}\|_{\ell_\infty} = \max_{1 \le i \le n} |x_i| .$$

Here $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. For any norm, the set $\{\mathbf{x} : \|\mathbf{x}\| \le 1\}$ is called the unit ball.

a. Sketch the unit balls for each of the norms defined above.

b. Show that the norms are equivalent by explicitly determining the comparability constants.

2.2. Let V be a complete inner product space; define $\|u\|$ to be the non-negative number $(u, u)^{1/2}$. Show that this defines a norm on V.

2.3. Let V be a Hilbert space and let $f, g \in V$. Verify the parallelogram law

$$\|f + g\|^2 + \|f - g\|^2 = 2\,\|f\|^2 + 2\,\|g\|^2 .$$

Note that the name comes from the special case of $\mathbb{R}^2$ where we know that the sum of the squares of the sides of a parallelogram is equal to the sum of the squares of the diagonals.

2.4. In the previous exercise we saw that any two elements of a Hilbert space V satisfies the parallelogram law; in fact, one can show that if B is a Banach space which satisfies the parallelogram law then it is also a Hilbert space (see, e.g., [Schechter]). Consider the Banach space of all bounded real functions on the interval $[0, 1]$ with the norm

$$\|u\| = \sup_{0 \le x \le 1} |u(x)| .$$

Find functions $f, g \in B$ which violate the parallelogram law and thus conclude that B is a Banach space but not a Hilbert space. (Hint: for example, find functions f, g such that $\|f\| = \|g\| = \|f - g\| = \|f + g\|$.)

2.5. Let S be a closed subspace of a Hilbert space V. Let $S^\perp$ be defined by

$$S^\perp = \{u \in V \,:\, (u, \phi) = 0 \text{ for all } \phi \in S\} .$$

Show that $S^\perp$ is a closed subspace of V. Also show that $S \cap S^\perp = \{0\}$ and thus V can be written as the direct sum of S and $S^\perp$ so that every element $u \in V$ can be written as the sum of an element in S and in $S^\perp$.

2.6. Let P be the projection operator from a Hilbert space V to a *closed* subspace $S \subset V$; i.e., P is an operator $P : V \to S$ such that

$$Pu = \begin{cases} u & \text{if } u \in S \\ u_0 & \text{otherwise,} \end{cases}$$

where $u = u_0 + u_1$ uniquely with $u_0 \in S$ and $u_1 \in S^\perp$.

a. Show that P is linear.

b. Clearly, the range of P is S. What is the kernel of P? Why?

c. Show that $P^2 = P$.

d. Show that $\|P\| = 1$ where

$$\|P\| = \sup_{\phi \in V} \frac{\|P\phi\|}{\|\phi\|} \qquad \text{for } \phi \neq 0\,.$$

e. Show that $I - P$ is the projection operator onto the orthogonal complement of S.

2.7. Prove that if $\{u_i\}$ is a convergence sequence in a normed linear space then the limit is unique.

2.8. (Computational) Consider the function $u(x) = x^3 \sin \pi x$ on $[0, 1]$. We want to determine the *best approximation* in the L^2-norm, $\tilde{u}(x)$, to $u(x)$ out of the space of continuous piecewise linear functions which are zero at $x = 0$ and $x = 1$.

a. Choose a uniform partition of $[0, 1]$ with $h = 0.25$. Write a code to determine the best approximation $\tilde{u}$ to $u(x)$ using the standard “hat” basis functions for continuous piecewise linears. For the integration, use a two-point Gauss quadrature rule. Write your code so that you have a separate function or subroutine which evaluates a basis function at any given point.

b. Repeat (a) with $h = 0.125$ and $h = .0625$. For each value of h determine the $L^2(0, 1)$ error in $u(x)$ and $\tilde{u}(x)$; calculate a numerical rate of convergence (i.e., determine k such that the error is $\mathcal{O}(h^k)$) based upon your two calculations. To calculate the error, apply the two-point Gauss quadrature rule over each subinterval.

Chapter 3

Abstract Formulation

{c

The first step in a finite element approach is to write an appropriate weak or variational problem. In lieu of deriving existence and uniqueness results for each weak problem we encounter, our strategy is to formulate a general weak problem and prove existence and uniqueness for it. Then, as we encounter specific weak problems, we only need to show that each problem fits into the framework of the general problem and satisfies any conditions required by our analysis of the general problem. We repeat the procedure with the discrete weak problem, but, in addition, derive a general error estimate. The tools introduced in the last chapter easily allow us to formulate a general weak problem; the existence and uniqueness of its solution is established through the Lax-Milgram theorem which is proved with the aid of the Projection and the Riesz Representation theorems from the previous chapter.

The abstract weak problem which we study is posed on a general Hilbert space, but when we look at specific examples we need to completely specify the particular space. It turns out that the class of Hilbert spaces that are appropriate is Sobolev spaces. Before studying the general problem, we introduce these spaces and the concept of weak derivatives.

Not all weak problems we encounter fit into the framework of the general problem introduced in this chapter. In later chapters we consider an obvious generalization to this weak problem, introduce a so-called *mixed weak problem.* Consequently, by the completion of this book, we plan to analyze several general weak problems which can handle a wide variety of linear problems.

When we derived the weak formulation to our prototype example in Chapter 1, we saw that it was equivalent to solving a corresponding minimization problem. Not all variational problems have this corresponding Rayleigh-Ritz formulation. In Section 3.4 we prove a result which gives conditions when the two formulations are equivalent.

3.1 Weak L^2 derivatives and Sobolev spaces

{a

In this section we define the particular class of Hilbert spaces which we use as our spaces of admissible functions; these spaces are called *Sobolev spaces.* We want to generalize the concept of derivative to define what we refer to as a *weak* or *generalized derivative* and do it in such a way that if everything is "smooth enough" then the classical and weak derivatives coincide. The concept of a weak derivative is an extension of the classical derivative which maintains the validity of the integration by parts formula or its analogue in higher dimensions. Our generalization allows functions such as $u(\mathbf{x}) = |\mathbf{x}|$ on $[-1, 1]$ to have a derivative in the weak sense.

We use this weak derivative in our definition of Sobolev spaces, the particular Hilbert spaces we need. We begin this section with some notation which simplifies the exposition, follow with the definition of a weak derivative, and then introduce Sobolev spaces with their associated norms and inner products.

As usual, let Ω be an open, connected subset of $\mathbb{R}^n$ and let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote a general point in Ω. The set of all real-valued functions $u(\mathbf{x}) = u(x_1, \cdots, x_n)$ which are defined and continuous on Ω is denoted $C(\Omega)$ and the set of all continuous functions having derivatives of order less than or equal to k continuous in Ω is denoted $C^k(\Omega)$, $k < \infty$. We also need the space C_0^∞ which is the space of infinitely differentiable functions which have compact support. A function ϕ has compact support if $\phi = 0$ outside a closed and bounded subset of Ω; the support of a function $\phi(x)$ generally refers to the closure of the set of all x for which $\phi(x) \neq 0$.

To simplify the derivative notation we introduce the notation of a *multi-index* α which is defined as an n-tuple of non-negative integers, i.e., $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ where α_i, $i = 1, \ldots, n$ is a non-negative integer. We use the notation

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n \,.$$

In this way we can rewrite the partial differential operator as

$$D^\alpha \equiv \frac{\partial^{\alpha_1 + \alpha_2 + \cdots + \alpha_n}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}} \,.$$

For example, in $\mathbb{R}^1$, $\alpha = \alpha_1$ so that D^α denotes the ordinary differential operator; for example, $D^2 = d^2/dx^2$. For $\mathbb{R}^2$, $\alpha = (\alpha_1, \alpha_2)$ and so for $|\alpha| = 1$ we have the first order partial differential operators $D^{(1,0)} = \partial/\partial x_1$ and $D^{(0,1)} = \partial/\partial x_2$. For $|\alpha| = 2$ we have

$$D^{(2,0)} = \frac{\partial^2}{\partial x_1^2}\,, \qquad D^{(0,2)} = \frac{\partial^2}{\partial x_2^2}\,, \quad \text{and} \quad D^{(1,1)} = \frac{\partial^2}{\partial x_1 x_2}\,.$$

Using this notation we can define $C^k(\Omega)$ as

$$C^k(\Omega) = \{u \,:\, D^\alpha u \in C(\Omega),\ |\alpha| \leq k\}\,.$$

3.1.1 Weak derivatives

We now define the concept of the *weak* (or *generalized* or *distributional*) $L^2(\Omega)$ derivative of a function. Let $u \in L^2(\Omega)$; we say that u has a derivative of order α

in the weak L^2-sense if there exists a function $v \in L^2(\Omega)$ such that

{abs_weakder}

$$\int_\Omega u D^\alpha \phi \, d\Omega = (-1)^{|\alpha|} \int_\Omega v \phi \, d\Omega \tag{3.1}$$

holds for all $\phi \in C_0^\infty(\Omega)$.

To help us understand this definition we consider a specific case in $\mathbb{R}^1$ where $\Omega = (0,1)$. Suppose $\phi(x)$ is a continuously differentiable function on Ω which vanishes on the boundary of Ω, i.e., $\phi(0) = \phi(1) = 0$. Let $u \in C^1([0,1])$. Then

$$\int_0^1 u \frac{\partial \phi}{\partial x} \, dx = \phi u \big|_0^1 - \int_0^1 \phi \frac{\partial u}{\partial x} \, dx$$

and thus

$$\int_0^1 u \frac{\partial \phi}{\partial x} \, dx = - \int_0^1 \phi \frac{\partial u}{\partial x} \, dx \, .$$

So the classical derivative $\partial u / \partial x$ can be viewed as a function v satisfying

$$\int_0^1 u \frac{\partial \phi}{\partial x} \, dx = - \int_0^1 \phi v \, dx \, . \tag{3.2}$$

{a

Conversely, if we find a function v satisfying (3.2) then it behaves like the derivative when integrated against functions in $C_0^\infty(\Omega)$. Note that (3.2) is just (3.1) where $|\alpha| = 1$ since Ω is a subset of $\mathbb{R}^1$.

We conclude that the classical derivatives, if they exist and are continuous in the usual sense, coincide with the weak derivatives. However, there are functions which possess a weak L^2-derivative but have no classical derivatives.

{a

Example 3.1 We know that the function $u(x) = |x|$ on $\Omega = (-1,1)$ does not have a classical derivative at $x = 0$; however it does have a generalized L^2-derivative. To see this, let

$$v(x) = \begin{cases} -1 & \text{for} \quad -1 < x \le 0 \\ 1 & \text{for} \quad 0 < x < 1 \, . \end{cases}$$

Clearly, $v \in L^2(\Omega)$ and we claim that $v(x)$ is the weak L^2-derivative of $u(x) = |x|$. To show this, we note that if $\phi \in C_0^\infty(-1,1)$ we have

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

and thus (3.1) is satisfied with $|\alpha| = 1$. ■

There are functions in $L^2(\Omega)$ which do not possess weak or classical derivatives. The reader is referred to [Adams] for a complete exposition of generalized derivatives.

We note that it can be proved that weak L^2-derivatives are unique almost everywhere; that is, unique except on a set of measure zero. For example, in Example 3.1 we could have chosen the generalized L^2-derivative to be

$$w(x) = \begin{cases} -1 & \text{for} \quad -1 < x < 0 \\ 1 & \text{for} \quad 0 \leq x < 1 \,. \end{cases}$$

Note that $w(x)$ and $v(x)$ defined in Example 3.1 differ only at the point $x = 0$, i.e., on a set of measure zero.

3.1.2 Sobolev spaces

We are now ready to define the class of Hilbert spaces that we use to pose our weak problems. The *Sobolev space* $H^m(\Omega)$ is the set of functions $u \in L^2(\Omega)$ which possess generalized (weak) L^2-derivatives $D^\alpha u$ which are also in $L^2(\Omega)$ for $0 \leq |\alpha| \leq m$; i.e.,

{abs_definesobolev}

$$H^m(\Omega) = \{ u \in L^2(\Omega) \; : \; D^\alpha u \in L^2(\Omega) \text{ for } 0 \leq |\alpha| \leq m \; \} \,. \tag{3.3}$$

Clearly, $H^m(\Omega)$ is a subspace of $L^2(\Omega)$ and $H^0(\Omega) = L^2(\Omega)$. On $H^m(\Omega)$ we define the inner product

{abs_sobolevinner}

$$\begin{aligned} (u,v)_m &= \sum_{|\alpha| \leq m} \int_\Omega D^\alpha u D^\alpha v \, d\Omega \\ &= \sum_{|\alpha| \leq m} (D^\alpha u, D^\alpha v) \quad \forall \, u,v \in H^m(\Omega) \,, \end{aligned} \tag{3.4}$$

where $(\cdot,\cdot)$ denotes the standard inner product on $L^2(\Omega)$. Using this definition of inner product, we define the norm on $H^m(\Omega)$ as

{abs_sobolevnorm}

$$\|u\|_m = (u,u)_m^{1/2} = \left(\sum_{|\alpha| \leq m} \|D^\alpha u\|^2 \right)^{1/2} \qquad \forall \, u \in H^m(\Omega) \,, \tag{3.5}$$

where $\|\cdot\|$ denotes the standard norm on $L^2(\Omega)$. Clearly, $\|\cdot\|_0$ is the standard $L^2(\Omega)$ norm so in the sequel we denote the L^2-norm by $\|\cdot\|_0$.

The following result guarantees that $H^m(\Omega)$ is a complete inner product space; for the proof, see [Adams].

{abs_thm_sobolev}

Theorem 3.2. $H^m(\Omega)$, *equipped with the inner product and norm defined in (3.4) and (3.5), respectively, is a Hilbert space and thus a Banach space.*

We make extensive use of the space $H^1(\Omega)$; if $\Omega \subset \mathbb{R}^1$ then the norm on $H^1(\Omega)$ is explicitly given by

{abs_honenorm1}

$$\|u\|_1 = \Big(\|u\|_0^2 + \|u'\|^2 \Big)^{1/2} \tag{3.6}$$

and if $\Omega \subset \mathbb{R}^2$ then the norm is explicitly given by

{abs_honenorm2}
$$\|u\|_1 = \left(\|u\|_0^2 + \left\| \frac{\partial u}{\partial x_1} \right\|_0^2 + \left\| \frac{\partial u}{\partial x_2} \right\|_0^2 \right)^{1/2} . \tag{3.7}$$

Note that by construction, for a function $u \in H^m(\Omega)$ we have

$$\|u\|_0 \leq \|u\|_1 \leq \|u\|_2 \cdots \leq \|u\|_m .$$

We also make use of the Sobolev semi-norm on $H^m(\Omega)$ which is denoted by $|\cdot|$ and defined by

$$|u|_m = \left(\sum_{|\alpha|=m} \|D^\alpha u\|_0^2 \right)^{1/2} \qquad \forall\, u \in H^m(\Omega) . \tag{3.8}$$ {a

Thus for $\Omega \subset \mathbb{R}^n$ the H^1 semi-norm is explicitly given by

$$|u|_1 = \left(\sum_{i=1}^{n} \left\| \frac{\partial u}{\partial x_i} \right\|_0^2 \right)^{1/2} \qquad \forall\, u \in H^1(\Omega) . \tag{3.9}$$ {a

Again by definition of the norms, we have that for $u \in H^m(\Omega)$

$$|u|_m \leq \|u\|_m . \tag{3.10}$$ {a

Note that we are using the standard notation for partial derivative and D^α interchangeably; the context should make it clear if we are referring to the classical or weak derivative.

We also make use of the constrained space $H_0^1(\Omega)$ which denotes all functions in $H^1(\Omega)$ which are zero on the boundary; i.e.,

$$H_0^1(\Omega) = \{ u \in H^1(\Omega) \ : \ u|_{\partial\Omega} = 0 \} . \tag{3.11}$$ {a

Formally, $H_0^1(\Omega)$ is defined as the completion of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_1$ and it can be shown that it is a closed subspace of $H^1(\Omega)$ consisting precisely of those functions $u \in H^1(\Omega)$ which almost everywhere satisfy $u = 0$ on the boundary of Ω.

We comment that if $\Omega \subset \mathbb{R}^n$ for $n > 1$, then $H^m(\Omega)$ can contain functions which are not continuous. As an example, if $n = 2$ and Ω is the open unit disk with center at the origin, consider the function $u = (\ln(1/r))^k$ for $k < 1/2$ and $r = (x_1^2 + x_2^2)^{1/2}$. It can be shown that $u \in H^1(\Omega)$ but u is not continuous at the origin. A result known as Sobolev's Theorem (see [Adams]) gives the connection between $H^m(\Omega)$ and $C^m(\Omega)$ for arbitrary m.

We conclude this section with the following result, known as the Poincaré inequality, which is extremely useful in relating the L^2-norm of certain functions in $H^1(\Omega)$ with their corresponding semi-norm. Recall that by the definition of the Sobolev norm, it is always true that $\|u\|_0 \leq \|u\|_1$. However, it is not obvious if the result holds when we replace the one-norm with the one semi-norm. It turns out that it is true for functions in $H_0^1(\Omega)$ and even for functions which are zero on some portion of their boundary. It is important to realize that this result does not, in general, hold for all functions in $H^1(\Omega)$.

Lemma 3.3. (Poincaré Inequality) *Let $u \in H^1(\Omega)$ such that $u = 0$ on some portion of the boundary of Ω. Then there exists a constant C depending on Ω such that* {a

{abs_poincare}

$$\|u\|_0 \le C \left(\sum_{i=1}^{n} \left\| \frac{\partial u}{\partial x_i} \right\|_0^2 \right)^{1/2} = C|u|_1 \,. \tag{3.12}$$

Note that the Poincaré inequality, along with (3.10) gives that on $H_0^1(\Omega)$ the H^1-norm and H^1-seminorm are equivalent norms.

3.2 Formulation and analysis of a general weak problem

{abs_sec_weakproblem}

In this section we use the tools developed in the last chapter to formulate a general weak problem. We state and prove the Lax-Milgram theorem which is central to the theory of the finite element method since it provides us with conditions which guarantee the existence and uniqueness of the solution of our general weak problem.

Let V denote a Hilbert space, let $A(\cdot,\cdot)$ denote a bilinear form on $V \times V$ and let F denote a linear functional on V. The general weak problem we consider is to

{abs_wp}

$$\left\{ \begin{array}{l} \text{seek } u \in V \text{ satisfying} \\ \qquad A(u,v) = F(v) \quad \forall\, v \in V \,. \end{array} \right. \tag{3.13}$$

Many weak formulations that we encounter can easily be put into the general form of (3.13) with appropriate choices for the Hilbert space, the bilinear form, and the linear functional.

{abs_ex1} **Example 3.4** Consider the simple two-point boundary value problem

{abs_ex1_de}

$$-u''(x) = \sin \pi x \quad 0 < x < 1 \tag{3.14a}$$

and the boundary conditions

{abs_ex1_bcl}

$$u(0) = 0 \tag{3.14b}$$

and

{abs_ex1_bcr}

$$u(1) = 0 \,. \tag{3.14c}$$

In choosing the underlying Hilbert space for our weak formulation of (3.14), we must require our solution to be in $L^2(0,1)$ and to possess at least one weak L^2-derivative. In addition, we want to constrain our space so that we only consider functions which satisfy the homogeneous Dirichlet boundary conditions. Thus we choose $H_0^1(0,1)$ to be the underlying Hilbert space in which we seek a solution $u(x)$. In particular, we seek a $u \in H_0^1(0,1)$ satisfying

{abs_exwp}

$$\int_0^1 u'v'\,dx = \int_0^1 \sin\pi x v\,dx \quad \forall\, v \in H_0^1(0,1)\,. \tag{3.15}$$

Clearly any solution of this two-point boundary value problem is also a solution of (3.15). Now we can easily cast (3.15) into the general form of (3.13) if we let $V = H_0^1(0,1)$,

$$A(u,v) = \int_0^1 u'v'\,dx$$

and

$$F(v) = \int_0^1 \sin\pi x\; v(x)\,dx = (\sin\pi x, v)\,,$$

where $(\cdot,\cdot)$ denotes the $L^2(0,1)$ inner product. Clearly $A(u,v)$ defined in this way is a bilinear form on $H^1(0,1)$ and $F(v)$ is a linear functional on $H^1(0,1)$ and thus on $H_0^1(0,1)$. ■

If F is a bounded linear functional on the given Hilbert space V and the bilinear form $A(\cdot,\cdot)$ is bounded, or equivalently, continuous on the space V and, in addition, satisfies a property referred to as *coercivity* or equivalently as *V-ellipticity*, then the Lax-Milgram theorem guarantees existence and uniqueness of the solution of (3.13). Moreover, the theorem also provides a bound of the solution of the weak problem in terms of the data. This is analogous to bounds obtained in PDE theory.

Theorem 3.5. (Lax-Milgram Theorem) *Let V be a Hilbert space and let $A(\cdot,\cdot) : V \times V \to \mathbb{R}^1$ be a bilinear form on V which satisfies* {a

$$|A(u,v)| \le M\,\|u\|\,\|v\| \quad \forall\, u,v \in V \tag{3.16}$$

{a

and

$$A(u,u) \ge m\,\|u\|^2 \quad \forall\, u \in V\,, \tag{3.17}$$

{a

where M and m are positive constants independent of $u, v \in V$. Let $F : V \to \mathbb{R}^1$ be a bounded linear functional on V. Then there exists a unique $u \in V$ satisfying (3.13). Moreover

$$\|u\| \le \frac{1}{m}\,\|F\|\;. \tag{3.18}$$

{a

Proof. In order to prove this result we begin by fixing a $u \in V$ and demonstrating that $Q(v) = A(u,v)$ defines a bounded linear functional on V. We then apply the Riesz representation theorem to obtain a unique element $\hat{u} \in V$ such that

$$Q(v) = A(u,v) = (v,\hat{u}) \quad \forall\, v \in V\,.$$

This allows us to associate to each $u \in V$ a unique $\hat{u} \in V$. If we denote this correspondence by $\hat{u} = \mathcal{A}u$ we have

$$A(u,v) = (v,\mathcal{A}u) \quad \forall\, u,v \in V\,. \tag{3.19}$$

{a

We then demonstrate that $\mathcal{A}$ is a linear operator and that the range of $\mathcal{A}$, denoted $\mathcal{R}(\mathcal{A})$, is a closed subspace of V and finally that $\mathcal{R}(\mathcal{A}) = V$.

Once we have established these facts then we can establish the existence and uniqueness by the following argument. Since F is a bounded linear functional on V then the Riesz representation theorem guarantees the existence of a unique element $\phi \in V$ such that $F(v) = (\phi, v)$ for all $v \in V$. If the $\mathcal{R}(\mathcal{A}) = V$, then there exists a $u \in V$ such that $\mathcal{A}u = \phi$. Hence there exists a $u \in V$ such that

$$F(v) = (\mathcal{A}u, v) = A(u, v) \quad \forall\, v \in V\,.$$

Uniqueness is shown in the standard way of choosing $u_1 \neq u_2$ such that

$$A(u_1, v) = A(u_2, v) = F(v) \quad \forall\, v \in V\,.$$

Then we have that $A(u_1 - u_2, v) = 0$ for all $v \in V$ and choosing $v = u_1 - u_2$, we conclude that $A(u_1 - u_2, u_1 - u_2) = 0$. Using (3.17) we know that $A(u_1 - u_2, u_1 - u_2) \geq m \left\| u_1 - u_2 \right\|^2$ which implies the contradiction $0 \geq m \left\| u_1 - u_2 \right\|^2$.

We now return to proving the claims necessary to complete the proof of existence. First, we see that $Q(v) \equiv A(u, v)$ is a bounded linear functional on V. Linearity immediately follows from the linearity of $A(\cdot, \cdot)$; the fact that it is bounded follows from (3.16); i.e.,

$$|Q(v)| = |A(u, v)| \leq C \left\| u \right\| \left\| v \right\|$$

and thus $\left\| Q \right\| \leq C \left\| u \right\| < \infty$. It is now required to show that the operator $\mathcal{A}$ is linear. Given $\phi, \psi \in V$

$$\begin{aligned}(v, \mathcal{A}(\alpha\phi + \beta\psi)) &= A(\alpha\phi + \beta\psi, v) = \alpha A(\phi, v) + \beta A(\psi, v) \\ &= \alpha\Phi(v) + \beta\Psi(v) \quad \forall\, v \in V\,.\end{aligned}$$

Using the same argument as we did for $Q(v)$, we see $\Phi(v)$ and $\Psi(v)$ are bounded linear functionals on V and so we can apply the Riesz representation theorem and the definition of $\mathcal{A}$ to write

$$\Phi(v) = (v, \hat{\phi}) = (v, \mathcal{A}\phi)$$

and similarly for $\Psi(v)$. Combining these results we obtain

$$(v, \mathcal{A}(\alpha\phi + \beta\psi)) = \alpha\,(v, \mathcal{A}\phi) + \beta\,(v, \mathcal{A}\psi) \quad \forall\, v \in V$$

and hence $\mathcal{A}(\alpha\phi + \beta\psi) = \alpha\mathcal{A}(\phi) + \beta\mathcal{A}(\psi)$; i.e., linearity is proved. It remains to show that $\mathcal{R}(\mathcal{A})$ is a closed subspace of V and, in fact, $\mathcal{R}(\mathcal{A}) = V$. The fact that $\mathcal{R}(\mathcal{A})$ is a subspace is obvious from its definition; to show that it is closed we choose a sequence $\{\hat{\phi}_n\} \in \mathcal{R}(\mathcal{A})$ which converges to $\hat{\phi} \in V$ and demonstrate that $\hat{\phi} \in \mathcal{R}(\mathcal{A})$. Since $\hat{\phi}_n \in \mathcal{R}(\mathcal{A})$ we can write $\hat{\phi}_n = \mathcal{A}\phi_n$ for $\phi_n \in V$; we want to demonstrate that $\{\phi_n\}$ is a Cauchy sequence in V. Now by the definition of $\mathcal{A}$, $(v, \mathcal{A}\phi_n) = A(\phi_n, v)$ for all $v \in V$ and thus $A(\phi_n - \phi_m, v) = (v, \mathcal{A}(\phi_n - \phi_m))$ for all $v \in V$. Choosing $v = \phi_n - \phi_m$ and using (3.17) we have that

$$m \left\| \phi_n - \phi_m \right\|^2 \leq a(\phi_n - \phi_m, \phi_n - \phi_m) = (\phi_n - \phi_m, \mathcal{A}(\phi_n - \phi_m))\,.$$

Using the Cauchy-Schwartz inequality and the linearity of A we have $\|\phi_n - \phi_m\| \leq \frac{1}{m} \|\mathcal{A}\phi_n - \mathcal{A}\phi_m)\| = \frac{1}{N}\|\hat{\phi}_n - \hat{\phi}_m\|$; we thus conclude that $\{\phi_n\}$ is a Cauchy sequence in V. Since V is complete, there exists a $\phi \in V$ such that $\phi_n \to \phi$. If we now show that $\hat{\phi} = \mathcal{A}\phi$ we have demonstrated that the limit of the sequence $\{\hat{\phi}_n\}$ is in $\mathcal{R}(\mathcal{A})$ and thus $\mathcal{R}(\mathcal{A})$ is closed. To see this we note that by using the linearity of $A(\cdot,\cdot)$ and (3.16) we have

$$|A(\phi_n, v) - A(\phi, v)| \leq M \|\phi_n - \phi\| \|v\| \quad \forall\, v \in V\,.$$

Thus $A(\phi_n, v) \to A(\phi, v)$ as $n \to \infty$ for all $v \in V$. In terms of an inner product, this yields $(v, \mathcal{A}\phi_n) \to (v, \mathcal{A}\phi)$ as $n \to \infty$. But $(\mathcal{A}\phi_n, v) = (\hat{\phi}_n, v) \to (\hat{\phi}, v)$. So $(\mathcal{A}\phi_n, v) \to (\hat{\phi}, v)$ and $(\mathcal{A}\phi_n, v) \to (v, \mathcal{A}\phi)$; thus $\hat{\phi} = \mathcal{A}\phi$ and the $\mathcal{R}(\mathcal{A})$ is closed. To show that $\mathcal{R}(\mathcal{A}) = V$ we assume that $\mathcal{R}(\mathcal{A}) \subset V$; i.e., there exists a $z \in \mathcal{R}(\mathcal{A})^{\perp}$. This implies $(z, \hat{v}) = 0$ for all $\hat{v} \in \mathcal{R}(\mathcal{A})$; or equivalently for all $v \in V$, $(z, \mathcal{A}v) = 0$. In particular, if we set $v = z$ we have $A(z,z) = (z, \mathcal{A}z) = 0$, but from (3.17)) $A(z,z) \geq N \|z\|^2$ implying that $z = 0$, a contradiction.

To conclude the proof we must demonstrate (3.18)). Since $A(u,u) = F(u)$ we have that

$$m \|u\|^2 \leq |A(u,u)| = |F(u)|$$

from which we have for $u \neq 0$

$$\|u\| \leq \frac{1}{m} \frac{|F(u)|}{\|u\|}\,.$$

Therefore

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

■

3.3 Galerkin approximations

{a

In the previous section we defined a general weak problem, (3.13), which is posed on an infinite-dimensional Hilbert space V. We then stated and proved the Lax-Milgram theorem which gave conditions guaranteeing existence and uniqueness of its solution. Since in finite elements, our objective is to approximate the solution of this weak problem, we want to state a general discrete weak problem, give conditions which guarantee existence and uniqueness of its solution, and finally to bound the error between the solution of (3.13) and the discrete solution.

We begin by letting $\{V^h\}$, $0 < h < 1$, be a family of finite dimensional subspaces of the Hilbert space V. Then the discrete problem corresponding to (3.13) for a fixed h is to

$$\left\{ \begin{array}{l} \text{seek } u^h \in V^h \text{ satisfying} \\ \qquad A(u^h, v^h) = F(v^h) \quad \forall\, v^h \in V^h\,. \end{array} \right. \tag{3.20}$$

{a

If the conditions of the Lax-Milgram theorem hold over the whole space V, then clearly they hold over any subspace V^h. Consequently, existence and uniqueness of (3.20) is automatically guaranteed by the Lax-Milgram theorem. The following result, known as Galerkin's or Cea's theorem, provides us with an error estimate for $\|u - u^h\|$ where $u \in V$ satisfies (3.13), $u^h \in V^h \subset V$ satisfies 3.20, and $\|\cdot\|$ denotes the norm on V. Simply stated, this result says that the error in the solution to the weak problem and its Galerkin approximation is less than or equal to a constant (which is ≥ 1) times the best approximation to the solution of (3.13) in S^h.

{la_thm_galerkin} **Lemma 3.6. (Galerkin's or Cea's Lemma)** *Let $A(\cdot,\cdot)$ be a bilinear form on V satisfying (3.16) and (3.17), and let $F(\cdot)$ be a bounded linear functional on V. Let u be the unique solution of*

$$A(u, v) = F(v) \quad \forall\, v \in V$$

guaranteed by the Lax-Milgram theorem. Let $\{V^h\}$, $0 < h < 1$, be a family of finite dimensional subspaces of V. Then for every h there exists a unique $u^h \in V^h$ such that

$$A(u^h, v^h) = F(v^h) \quad \forall\, v^h \in V^h$$

and moreover,

{la_errorestimate}
$$\|u - u^h\| \leq \frac{M}{m} \inf_{\chi^h \in V^h} \|u - \chi^h\| \,, \tag{3.21}$$

where M, m are the constants appearing in the Lax-Milgram theorem and $\|\cdot\|$ denotes the norm on V.

Proof. As indicated in the discussion preceding the theorem, the existence and uniqueness of (3.20) is guaranteed by the Lax-Milgram theorem. In order to prove our error estimate, we begin by establishing the so-called Galerkin *orthogonality condition.* We note that (3.13) holds for all $v \in V$ so, in particular, it holds for all $v^h \in V^h \subset V$; i.e.,

$$A(u, v^h) = F(v^h) \quad \forall\, v^h \in V^h \,.$$

Subtracting this expression from (3.20) we have that

{abs_galerkinorth}
$$A(u - u^h, v^h) = 0 \quad \forall\, v^h \in V^h \tag{3.22}$$

which says that the error $u - u^h$ is orthogonal to V^h. Using the coercivity property of $A(\cdot,\cdot)$ given in (3.17) we have

{abs_gal1}
$$m \left\|u - u^h\right\|^2 \leq A(u - u^h, u - u^h) \,; \tag{3.23}$$

adding and subtracting an arbitrary element $\chi^h \in V^h$ and using the linearity of $A(\cdot,\cdot)$ gives

$$A(u - u^h, u - u^h) = A(u - u^h, u - \chi^h + \chi^h - u^h) = A(u - u^h, u - \chi^h) + A(u - u^h, \chi^h - u^h) \,.$$

Now the orthogonality condition (3.22) tells us that the last term is zero since $\chi^h - u^h \in V^h$. Combining this result with (3.23) and using the bound on $A(\cdot,\cdot)$ given in (3.16) we have

$$m \left\|u - u^h\right\|^2 \leq A(u - u^h, u - \chi^h) \leq M \left\|u - \chi^h\right\| \left\|u - u^h\right\| \quad \forall\, \chi^h \in V^h$$

and thus

$$\left\|u - u^h\right\| \leq \frac{M}{m} \left\|u - \chi^h\right\| \quad \forall\, \chi^h \in V^h\,.$$

Taking the infimum over all $\chi^h \in V^h$ provides the final result. ■

As an immediate corollary to this result we have that if the family of subspaces V^h has the property that the norm of u minus its best approximation in V^h approaches zero as $h \to 0$ then we have convergence of u^h to u as $h \to 0$.

Corollary 3.7. *If* $\{V^h\}$, $0 < h < 1$, *is a family of subspaces of* V *which satisfy*

$$\lim_{h\to 0} \inf_{\chi^h \in V^h} \left\|u - \chi^h\right\| = 0 \tag{3.24}$$

{a

then $\left\|u - u^h\right\| \to 0$ *as* $h \to 0$.

It is important to note that if w^h is any element of V^h then

$$\inf_{\chi^h \in V^h} \left\|u - \chi^h\right\|_1 \leq \left\|u - w^h\right\|_1\,.$$

This is particularly useful when we want to bound the error $\left\|u - u^h\right\|_1$ in terms of powers of h. From the study of approximation theory, we know that bounds are not readily available for the best approximation but bounds are easy to obtain for particular elements of V^h such as the V^h-interpolant of u. Thus if we can bound the error in u and its V^h-interpolant in terms of powers of h, then we have a useful bound for $\left\|u - u^h\right\|_1$. We return to this when we consider particular examples in the next chapter.

The discrete weak problem (3.20) results in a linear algebraic system of equations once a basis is chosen for the n-dimensional space V^h. In particular, let $\{\phi_i(x)\}$, $i = 1, \ldots, n$ be a basis for V^h. Then $u^h \in V^h$ can be written as a linear combination of these basis vectors, i.e.,

$$u^h = \sum_{j=1}^{n} \xi_j \phi_j(x)$$

and thus (3.20) becomes

$$A\big(\sum_{j=1}^{n} \xi_j \phi_j(x), v^h\big) = F(v) \quad \forall\, v^h \in V^h\,. \tag{3.25}$$

{a

Now testing (3.25) against each $v^h \in V^h$ is equivalent to testing it against each element in the basis for V^h so that we have

$$\sum_{j=1}^{n} \xi_j A\big(\phi_j(x), \phi_i(x)\big) = F\big(\phi_i(x)\big) \quad i = 1, 2, \ldots, n$$

or in matrix form $Ac = b$ where A is an $n \times n$ matrix, $c, b \in \mathbb{R}^n$ with

$$A_{ij} = A(\phi_j, \phi_i)\,, \qquad c_i = \xi_i \quad \text{and} \quad b_i = F(\phi_i)\,. \tag{3.26}$$ {a

Properties of the bilinear form $A(\cdot,\cdot)$ are inherited by the matrix A. From numerical linear algebra, we know that a symmetric, positive definite matrix is easily solved by Cholesky factorization or by an iterative method. Consequently, it is worthwhile to note the conditions on $A(\cdot,\cdot)$ which guarantee that the resulting matrix is symmetric, positive definite.

{abs_thm_spd} **Lemma 3.8.** *Let $A(\cdot,\cdot)$ be a symmetric bilinear form defined on $V \times V$. If $A(\cdot,\cdot)$ satisfies the coercivity condition (3.17), then the matrix defined by (3.26) is symmetric and positive definite.*

Proof. See exercises. ∎

Of course we have not discussed choices of the finite dimensional subspaces V^h; we address some simple choices in the next chapters when we consider specific examples; Chapter 6 is devoted entirely to the study of finite element spaces. However, it is important to keep in mind that of all possible choices for V^h, finite element methods usually employ continuous piecewise polynomial spaces.

We have seen that if our bilinear form is symmetric and coercive, then the resulting matrix is symmetric, positive definite. However, since the size of our linear system can be quite large, especially in two and three dimensions, we would also like to have a sparse, banded matrix. The choice of basis for V^h governs this sparsity. In particular, we choose basis functions which have *compact support*, i.e., are zero outside of a compact set. So, for example, in one dimension we choose basis functions which are nonzero on as few intervals as possible.

Example 3.9 Returning to (3.15), the variational formulation in Example 3.1, we see that the corresponding discrete weak problem is to seek $u^h \in V^h \subset H_0^1(0,1)$ satisfying

$$\int_0^1 \frac{\partial u^h}{\partial x}\frac{\partial v^h}{\partial x}\,dx = \int_0^1 \sin \pi x \; v^h\,dx \quad \forall\; v^h \in V^h$$

and Galerkin's lemma provides us with an error bound using the norm on $H^1(0,1)$. In particular we have that

$$\big\|u - u^h\big\|_1 \le \inf_{\chi^h \in V^h} \big\|u - \chi^h\big\|_1$$

where

$$\big\|u - u^h\big\|_1 = \left(\int_0^1 (u - u^h)^2\,dx + \int_0^1 (\frac{du}{dx} - \frac{du^h}{dx})^2\,dx\right)^{1/2}.$$

■

3.4 The Rayleigh-Ritz problem

Recall from linear algebra that determining an $x \in \mathbb{R}^n$ satisfying the linear system $Ax = b$, where A is an $n \times n$ *symmetric, positive definite* matrix, $b \in \mathbb{R}^n$, is equivalent to solving the minimization problem

$$\min_{y \in \mathbb{R}^n} \left(\frac{1}{2} y^T A y - y^T b \right) . \tag{3.27}$$

See exercises. Although we rarely solve a linear system as a minimization problem, the equivalence between the two problems is often useful. In this section we want to show that an analogous relationship exists between the solution of the weak problem (3.13) and an appropriate minimization problem; this minimization problem is often called the Ritz problem or the Rayleigh-Ritz problem.

Consider the minimization problem

$$\min_{v \in V} \mathcal{J}(v) \tag{3.28}$$

where $\mathcal{J} : V \to \mathbb{R}$ is the functional defined by

$$\mathcal{J}(v) = \frac{1}{2} A(v, v) - F(v) \quad \forall\, v \in V . \tag{3.29}$$

It turns out that if $A(\cdot, \cdot)$ satisfies the hypotheses of the Lax-Milgram theorem and is *symmetric* then solving the minimization problem (3.28) is equivalent to solving the weak problem (3.13). Consequently, once we discretize a symmetric problem, we have the choice of solving it as a system of linear algebraic equations or as a minimization problem. The following result demonstrates the equivalence of the two problems.

Theorem 3.10. *Let $A(\cdot, \cdot)$ be a symmetric bilinear form satisfying the hypotheses of the Lax-Milgram Theorem. Then the problem of finding a u satisfying the weak problem (3.13) and finding a solution to the minimization problem (3.28) are equivalent.*

Proof. First assume that $u \in V$ satisfies the weak problem (3.13) and let $w \in V$ be artibrary. Then using the definition (3.29) of $\mathcal{J}$ and the linearity of $A(\cdot, \cdot)$ and $F(\cdot)$, we obtain

$$\begin{aligned}
\mathcal{J}(u + w) &= \frac{1}{2} A(u + w, u + w) - F(u + w) \\
&= \frac{1}{2} A(u, u) + \frac{1}{2} \big(A(w, u) + A(u, w) \big) + A(w, w) - F(u) - F(w) \\
&= \mathcal{J}(u) + A(u, w) - F(w) + A(w, w) ,
\end{aligned}$$

where in the last step we have used the symmetry of $A(\cdot, \cdot)$ and the definition of $\mathcal{J}$. Since $w \in V$ and u satisfies (3.13), $A(u, w) - F(w) = 0$. Also since $A(\cdot, \cdot)$ is coercive,

$A(w,w) > 0$ for $w \neq 0$. Therefore $\mathcal{J}(u+w) > \mathcal{J}(u)$ and thus u is a minimizer of (3.28).

Now assume that u minimizes $\mathcal{J}(v)$ for all $v \in V$. Then for any scalar σ and $v \in V$, $u + \sigma v \in V$ and so $\mathcal{J}(u+\sigma v) \geq \mathcal{J}(u)$. Then the function $g(\sigma) = \mathcal{J}(u+\sigma v)$ has a minimum at $\sigma = 0$. From calculus, we know that

$$\left.\frac{dg}{d\sigma}\right|_{\sigma=0} = 0\,.$$

Since

$$\begin{aligned}
\frac{dg}{d\sigma} &= \frac{d}{d\sigma}\left(\frac{1}{2}A(u+\sigma v, u+\sigma v) - F(u+\sigma v)\right) \\
&= \frac{d}{d\sigma}\left(\frac{1}{2}A(u,u) + \sigma A(u,v) + \frac{1}{2}\sigma^2 A(v,v) - F(u) - \sigma F(v)\right) \\
&= A(u,v) + \sigma A(v,v) - F(v)
\end{aligned}$$

where we have used the properties of $A(\cdot,\cdot)$ and the inner product. Evaluating this derivative at $\sigma = 0$, we arrive at $A(u,v) - F(v) = 0$ for all $v \in V$, i.e., if u minimizes (3.28) then u satisfies (3.13). ∎

Exercises

3.1. Let P^h be the projection operator $P^h : V \to V^h$. Demonstrate that

$$\left\|u - u^h\right\| \leq \frac{M}{m} \left\|u - P^h u\right\| \,, \tag{3.30}$$

where M, m are the constants appearing in the Lax Milgram Theorem 3.5.

3.2. Prove Lemma 3.8.

3.3. Show that on $H_0^1(\Omega)$ the H^1-norm and the H^1-seminorm are equivalent norms.

3.4. Show that determining an $x \in \mathbb{R}^n$ satisfying the linear system $Ax = b$, where A is an $n \times n$ *symmetric, positive definite* matrix, $b \in \mathbb{R}^n$, is equivalent to solving the minimization problem (3.27).

3.5. Give an example of a weak formulation for a linear two-point boundary value problem on $[0, 1]$ which is not equivalent to a Rayleigh-Ritz minimization problem. Explain your reasoning.

Chapter 4

Finite Element Method for Ordinary Differential Equations

{c

In this chapter we consider some simple examples of the finite element method for the approximate solution of ordinary differential equations. Although the principal use of finite element methods is for approximating solutions to partial differential equations, it is instructive to look at one-dimensional problems for their simplicity and ease of understanding. In addition, when we approximate PDEs using rectangular elements, then we take tensor products of one-dimensional elements.

In the first three examples we consider a two-point boundary value problem for a second-order linear ordinary differential equation. Each of these examples is constructed so that the approach for handling different boundary data is made evident. The fourth example is a higher order differential equation.

In each example we define an appropriate weak formulation, either prove or indicate how the hypotheses of the Lax-Milgram theorem can be established, discuss the finite element approximation of the weak problem, and present error estimates. In addition, we provide computational results for some examples.

4.1 A two-point BVP with homogeneous Dirichlet boundary data

{1

We begin by considering the following two-point boundary value problem on $[0,1]$ where we seek a function $u(x)$ satisfying

$$\begin{aligned} -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)+q(x)u &= f(x) \quad \text{for } 0<x<1 \\ u(0) &= 0 \\ u(1) &= 0\,, \end{aligned} \tag{4.1}$$

{1

where $p(x)$, $q(x)$, and $f(x)$ are given functions defined on $[0,1]$. In the sequel we assume that $0 < p_{\min} \leq p(x) \leq p_{\max}$ and $q_{\min} = 0 \leq q(x) \leq q_{\max}$ where $p_{\min}$, $p_{\max}$, and $q_{\max}$ are positive constants and $f \in L^2(0,1)$. This problem is often referred to as a Sturm-Liouville problem.

It is well-known that whenever $f, q \in C[0,1]$ and $p \in C^1[0,1]$ the boundary value problem (4.1) possesses a unique *classical solution* $u(x) \in C^2(0,1)$ which satisfies (4.1) for every $x \in [0,1]$. We are interested in a *weak* or *generalized solution* of (4.1); i.e.,in a function $u(x)$ that satisfies (4.1) in some sense even when f, p, q are not continuous; if f, p, q are sufficiently smooth then we want the weak solution to coincide with the classical solution.

4.1.1 Weak formulation

In choosing the underlying Hilbert space for our weak formulation of (4.1), we know that multiplication of the differential equation by an appropriate test function, integrating over the domain and then integrating by parts to balance the order of the derivatives results in both the test and trial functions having one derivative. Consequently we require our solution to be in $L^2(0,1)$ and to possess at least one weak L^2-derivative. In addition, we constrain our space so that we only consider functions which satisfy the homogeneous Dirichlet boundary conditions. Thus we choose $H_0^1(0,1)$ to be the underlying Hilbert space in which we seek a solution $u(x)$ and for our test space. On $H_0^1(0,1)$ we define the bilinear form $A(\cdot,\cdot)$ by

{1d_homodir_bilinear}

$$A(v,w) = \int_0^1 p(x)v'(x)w'(x)\,dx + \int_0^1 q(x)v(x)w(x)\,dx = (pv', w') + (qv, w)\ , \quad (4.2)$$

where $(\cdot,\cdot)$ denotes the standard $L^2(\Omega)$-inner product. The weak problem is stated as:

{1d_homodir_weak}

$$\left\{ \begin{array}{l} \text{seek } u \in H_0^1(0,1) \text{ satisfying} \\ \qquad A(u,v) = (f,v) \quad \forall\, v \in H_0^1(0,1)\,. \end{array} \right. \quad (4.3)$$

Note that if u is the classical solution of (4.1) then $u(x)$ also satisfies the weak problem because for $v \in H_0^1(0,1)$

$$\begin{aligned} (f,v) = \int_0^1 fv\,dx &= \int_0^1 \big(-(pu')' + qu\big)v\,dx \\ &= \int_0^1 pu'v'\,dx + \int_0^1 quv\,dx - \big[pu'v\big]\,\big|_0^1 \\ &= \int_0^1 pu'v'\,dx + \int_0^1 quv\,dx = A(u,v)\,. \end{aligned}$$

Conversely, if $u \in H_0^1(0,1)$ satisfies (4.3) and is u is sufficiently smooth, i.e.,$u \in C^2(0,1)$, a situation which can be guaranteed if p, q and f are themselves sufficiently smooth, then u coincides with the classical solution of (4.1). The homogeneous Dirichlet boundary conditions are satisfied because $u \in H_0^1(0,1)$ and the differential equation holds because

$$\begin{aligned} A(u,v) - (f,v) &= \int_0^1 pu'v'\,dx + \int_0^1 quv\,dx - \int_0^1 fv\,dx \\ &= \int_0^1 \big[(-pu')' + qu - f\big]v\,dx = 0 \quad \forall\, v \in H_0^1(0,1) \end{aligned}$$

and $v \in H_0^1(0,1)$ is arbritary. Recall that if we can find a function $u \in H_0^1(0,1)$ which is the unique solution of (4.3), then we call u the *weak solution* of (4.1) in $H_0^1(0,1)$.

To prove the existence and uniqueness of $u \in H_0^1(0,1)$ satisfying (4.3) we use the Lax-Milgram theorem (Theorem 3.5) and verify that $A(\cdot,\cdot)$ and $F(v)$ satisfy the hypotheses of this theorem. Clearly, $A(\cdot,\cdot)$ is a bilinear form on $H_0^1(0,1)\times H_0^1(0,1)$. We first show that it is bounded on the space $H_0^1(0,1)$, i.e., $|A(v,w)| \le M\,\|v\|_1\,\|u\|_1$. To do this we use properties of integrals, the given bounds on p, q and the Cauchy-Schwartz inequality to obtain

$$\begin{aligned}|A(v,w)| &\le \Big|\int_0^1 p(x)v'w'\,dx\Big| + \Big|\int_0^1 q(x)vw\,dx\Big| \\ &\le p_{\max}\Big|\int_0^1 v'w'\,dx\Big| + q_{\max}\Big|\int_0^1 vw\,dx\Big| \\ &= p_{\max}\Big|(v',w')\Big| + q_{\max}\Big|(v,w)\Big| \\ &\le p_{\max}\|v'\|_0\|w'\|_0 + q_{\max}\|v\|_0\|w\|_0\,.\end{aligned}$$

To complete the result, we note that by the definition of the L^2-norm and the H^1-norm and seminorm, $\|w'\|_0 = |w|_1$, $\|\cdot\|_0 \le \|\cdot\|_1$, $|\cdot|_1 \le \|\cdot\|_1$. Thus

$$|A(v,w)| \le p_{\max}\,\|v\|_1\,\|w\|_1 + q_{\max}\,\|v\|_1\,\|w\|_1 \le C\,\|v\|_1\,\|w\|_1\,,$$

where $C = p_{\max} + q_{\max}$. Therefore, condition (3.16) of the Lax-Milgram theorem is satisfied.

In general, demonstrating coercivity of the bilinear form usually requires more finesse than proving continuity. We must prove that $A(v,v) \ge m\,\|v\|_1^2$. In our case we have

$$A(v,v) = \int_0^1 p(v')^2\,dx + \int_0^1 qv^2\,dx \ge p_{\min}\|v'\|_0^2 + q_{\min}\|v\|_0^2\,.$$

But we have assumed $q_{\min} = 0$ so

$$A(v,v) \ge p_{\min}\|v'\|_0^2\,.$$

We must now bound $\|v'\|_0 = |v|_1$ below by a constant times $\|v\|_1$ for all $v \in H_0^1(0,1)$. The fact that $v \in H_0^1(0,1)$ allows us to use the Poincaré inequality (**??**) to bound $|v|_1 \ge \frac{1}{C_p}\|v\|_0$. Using this bound for the entire term $\|v'\|_0^2 = |v'|_1^2$ does not give us the desired result so we use the approach of breaking this term into two parts; we have

$$p_{\min}\|v'\|_0^2 = p_{\min}|v|_1^2 = p_{\min}\Big(\frac{1}{2}|v|_1^2 + \frac{1}{2}|v|_1^2\Big) \ge \frac{1}{2}p_{\min}\Big(|v|_1^2 + \frac{1}{C_p^2}\|v\|_0^2\Big)\,.$$

Then

$$A(v,v) \ge \frac{1}{2}p_{\min}\Big[\min\Big(1,\frac{1}{C_p^2}\Big)\Big]\Big(|v|_1^2 + \|v\|_0^2\Big) = m\,\|v\|_1^2\,,$$

where we have used the definition of the H^1-norm, $\|\cdot\|_1^2 = \|\cdot\|_0^2 + |\cdot|_1^2$; thus the coercivity condition (3.17) is satisfied. Clearly $F(v) = (f, v)$ is a bounded linear functional on $H_0^1(0,1)$. Thus the Lax-Milgram theorem guarantees the existence of a unique $u \in H_0^1(0,1)$ which satisfies (4.3).

In this problem we constrained our Hilbert space to consist of functions which satisfy the homogenous Dirichlet boundary conditions. We recall that boundary conditions which are satisfied by constraining the admissible or trial space are called *essential.*

4.1.2 Approximation using piecewise linear polynomials

We now turn to approximating u, the solution of the weak problem (4.3), by its Galerkin approximation u^h in a finite dimensional subspace S_0^h of $H_0^1(0,1)$. The approximate solution is required to satisfy (4.3) but only for all $v^h \in S_0^h$; the discrete weak problem is

{1d_homodir_dweak}

$$\begin{cases} \text{seek } u^h \in S_0^h \text{ satisfying} \\ \quad A(u^h, v^h) = \big(f, v^h\big) \quad \forall\, v^h \in S_0^h\,. \end{cases} \tag{4.4}$$

Because $S_0^h \subset H_0^1(0,1)$ the conditions of the Lax Milgram theorem are automatically satisfied on S_0^h and so we are guaranteed that there exists a unique $u^h \in S_0^h$ which satisfies (4.4). Moreover, Galerkin/Cea's Lemma gives us the error estimate

{1d_homodir_errorba}

$$\left\|u - u^h\right\|_1 \le C \inf_{\chi^h \in S_0^h} \left\|u - \chi^h\right\|_1\,. \tag{4.5}$$

First we choose S_0^h to be the space of continuous linear piecewise polynomials defined on a partition of $[0,1]$ which satisify the homogeneous Dirichlet boundary conditions. In particular, we consider the following partition of $[0,1]$:

{1d_subdiv}

$$0 = x_0 < x_1 < \cdots < x_{N+1} = 1 \quad \text{where} \quad x_i = x_{i-1} + h_i, \quad 1 \le i \le N+1\,, \tag{4.6}$$

and where h_i, $1 \le i \le N+1$ are given numbers such that $0 < h_i < 1$ and $\sum_{i=1}^{N+1} h_i = 1$. We define $h = \max\limits_{1\le i\le N+1} h_i$; if $h_i = h$ for all i then we call the subdivision *uniform.* A *continuous piecewise linear function* with respect to the given subdivision on $[0,1]$ is a function $\phi(x)$ defined on $[0,1]$ which is linear on each subinterval; i.e.,$\phi(x) = \alpha_i x + \beta_i$ on $[x_i, x_{i+1}]$, $0 \le i \le N$. To impose continuity we require that the constants satisfy α_i, β_i where $\alpha_{i-1}x_i + \beta_{i-1} = \alpha_i x_i + \beta_i$, $i = 1, \ldots, N$; We define

{1d_homodir_sh}

$$\begin{aligned} S_0^h = \{&\phi(x) \;:\; \phi \in C[0,1], \\ &\phi(x) \text{ linear on } [x_i, x_{i+1}] \text{ for } 0 \le i \le N, \phi(0) = \phi(1) = 0\}\,. \end{aligned} \tag{4.7}$$

As we discussed in Chapter **??** we want to choose a basis whose functions have as small support as possible so that the resulting coefficient matrix is sparse. For $1 \le i \le N$ we consider again the "hat" functions (see Figure 1.3)

{1d_homodir_basis}

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i} & \text{for } x_{i-1} \le x \le x_i \\ \dfrac{x_{i+1} - x}{h_{i+1}} & \text{for } x_i \le x \le x_{i+1} \\ 0 & \text{elsewhere.} \end{cases} \tag{4.8}$$

Clearly $\phi_i(x) \in S_0^h$ for $1 \le i \le N$. Moreover, we easily see that

{1d_homodir_propbasis}

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

for $1 \le i \le N$ and $0 \le j \le N+1$. Here δ_{ij} denotes the Kronecker delta function. The following proposition justifies our intuition that the functions defined in (4.8) form a basis for S_0^h.

Proposition 4.1. *S_0^h defined by (4.7) is an N-dimensional subspace of $H_0^1(0,1)$. The functions $\{\phi_i(x)\}_{i=1}^N$ defined in (4.8) form a basis for S_0^h.* {1

Proof. Every function $\phi(x) \in S_0^h$ also belongs to $L^2(0,1)$ and each function is piecewise linear, so analogous to the function $|x|$, each has a weak derivative (which is piecewise constant) in $L^2(0,1)$. Also $\phi(0) = \phi(1) = 0$ for all $\phi \in S_0^h$ so that $S_0^h \subset H_0^1(0,1)$. We now show that $\{\phi_i(x)\}$, $i = 1, \ldots, N$ are linearly independent and span the space S_0^h. To see that we have a linearly independent set, let $\psi(x) = \sum_{i=1}^N c_i\phi_i(x)$; we want to show that the only way $\psi(x) = 0$ for all x is if $c_i = 0$ for $i = 1, \ldots, N$. Using (4.9), we see that $\psi(x_i) = c_i$ for $1 \le i \le N$. Thus if $\psi(x) = 0$ for all x we have that $c_i = 0$ for $i = 1, \ldots, N$; in addition if $c_i = 0$ for all $1 \le i \le N$ then the nodal values of ψ are zero and since it is piecewise linear, it is zero everywhere. Hence we conclude that the functions are linearly independent. To show that the set spans S_0^h we let $\psi(x)$ be an arbitrary element of S_0^h and show that we can write $\psi(x)$ as a linear combination of the $\phi_i(x)$, $i = 1, \ldots, N$; i.e.,$\psi(x) = \sum_{i=1}^N c_i\phi_i(x)$. But this can be done by letting $c_i = \psi(x_i)$, i.e.,setting c_i to be the *nodal values* of ψ. ■

Once we have chosen a basis for S_0^h, the problem (4.4) reduces to solving a system of N algebraic equations in N unknowns. Since $u^h \in S_0^h$, we let $u^h(x) = \sum_{j=1}^N \xi_j\phi_j(x)$ and write (4.4) as

$$\sum_{j=1}^N \xi_j A(\phi_j, \phi_i) = (f, \phi_i) \quad \text{for } 1 \le i \le N\,.$$

Then $\vec{c} = (\xi_1, \xi_2, \ldots, \xi_N)^T$ satisfies the matrix system

$$\mathcal{A}\vec{c} = \vec{b}\,, \tag{4.10}$$ {1

where $\vec{b} = \big((f,\phi_1), (f,\phi_2), \ldots, (f,\phi_N)\big)^T$ and $\mathcal{A}$ is the $N \times N$ matrix whose elements are given by

$$\mathcal{A}_{ij} = A(\phi_j, \phi_i) = \left(p\phi_j', \phi_i'\right) + (q\phi_j, \phi_i)$$

or

$$\mathcal{A}_{ij} = \mathcal{S}_{ij} + \mathcal{M}_{ij}$$

with $\mathcal{S}_{ij} = \left(p\phi_j', \phi_i'\right)$ and $\mathcal{M}_{ij} = (q\phi_j, \phi_i)$. The matrix $\mathcal{A}$ is symmetric, positive definite (see the exercises) and tridiagonal. If $p(x) = q(x) = 1$ on $[0,1]$ and we use a uniform mesh, then the matrices $\mathcal{S}$ and $\mathcal{M}$ are explicitly given by

$$\mathcal{S} = \frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & & 0 & -1 & 2 \end{pmatrix} \tag{4.11}$$

{1

and

{1d_homodir_mass}

$$\mathcal{M} = \frac{h}{6}\begin{pmatrix} 4 & 1 & 0 & & \cdots & 0 \\ 1 & 4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 & 4 & 1 \\ 0 & \cdots & & 0 & 1 & 4 \end{pmatrix}. \tag{4.12}$$

In the case $p = 1$ the matrix $\mathcal{S}$ is called the *stiffness matrix* of the basis $\{\phi_i\}_{i=1}^N$ while in the case $q = 1$, the matrix $\mathcal{M}$ is called the *Gram matrix* or the *mass matrix* associated with the basis $\{\phi_i\}_{i=1}^N$.

Solution of the linear system

Our coefficient matrix is a symmetric, positive-definite, tridiagonal matrix. If we choose a direct solver, then a Cholesky tridiagonal solver should be used because it takes advantage of these properties of the matrix. Recall that in a Cholesky factorization we write $\mathcal{A} = LL^T$ where L is a lower triangular matrix with positive elements on the diagonal. If $\mathcal{A}$ is the tridiagonal matrix

$$\mathcal{A} = \begin{pmatrix} a_1 & b_2 & & \\ b_2 & a_2 & b_3 & \\ & \ddots & \ddots & \ddots \\ & & b_N & a_N \end{pmatrix} = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_N & \alpha_N \end{pmatrix}\begin{pmatrix} \alpha_1 & \beta_2 & & \\ & \alpha_2 & \beta_3 & \\ & & \ddots & \ddots \\ & & & \alpha_N \end{pmatrix}$$

then

{1d_trisolver_factor}

$$\begin{array}{rrcl} & \alpha_1 & = & \sqrt{a_1} \\ \text{for } i = 2, \dots, N & \beta_i & = & b_i/a_{i-1} \quad \text{and} \quad \alpha_i = \sqrt{a_i - \beta_i^2}\,. \end{array} \tag{4.13}$$

Note that we can not determine all the β_i first and then determine the α_i but rather for each i we must determine β_i and then α_i before incrementing i. To solve the system $\mathcal{A}\vec{c} = \vec{f}$ we write $LL^T\vec{c} = \vec{f}$ and solve $L\vec{y} = \vec{f}$ and $L^T\vec{c} = \vec{y}$. Doing this we have the equations

{1d_trisolver_forward}

$$\begin{array}{rrcl} & y_1 & = & f_1/\alpha_1 \\ \text{for } i = 2, \dots, N & y_i & = & \dfrac{f_i - \beta_i y_{i-1}}{\alpha_i} \end{array} \tag{4.14}$$

and

{1d_trisolver_backward}

$$\begin{aligned} c_N &= y_N/\alpha_N \\ \text{for } i = N-1,\ldots,1 \qquad c_i &= \frac{y_i - \beta_{i+1} y_{i+1}}{\alpha_i}\,. \end{aligned} \tag{4.15}$$

For simplicity of exposition we have defined new variables, α_i, β_i, y_i, and c_i while in practice the entries of $\mathcal{A}$ and $\vec{f}$ are overwritten and no new arrays need be defined.

Error estimates and interpolation results

The bound for the error (4.5) in terms of the error in u and its best approximation in the subspace is not particularly useful in computations; what we would like is to measure the error in terms of powers of h. In order to have a quantitative estimate in terms of powers of h we need to estimate the H^1-error in u and its best approximation in S_0^h but this is difficult to do. However, we note that

$$\inf_{\chi^h \in S_0^h} \left\| u - \chi^h \right\|_1 \le \left\| u - w^h \right\|_1 \quad \text{for any } w^h \in S_0^h$$

is always true by the definition of the best approximation. So we immediately have

$$\left\| u - u^h \right\|_1 \le C \left\| u - w^h \right\|_1 \quad \text{for any } w^h \in S_0^h. \tag{4.16}$$ {1

Thus we need to find an element of S_0^h for which an approximation result is available.

Recall from elementary numerical analysis that one way to approximate a function is to use a polynomial interpolant; i.e.,to find a polynomial which agrees with the given function or its derivatives at a set of points. One such example is a *Lagrange interpolant* which interpolates given data or function values. Due to the fact that one cannot guarantee that the norm of the difference in the function and the Lagrange interpolating polynomial approaches zero as the degree of the polynomial increases, one often considers piecewise polynomial interpolation. In piecewise Lagrange interpolation we put together Lagrange polynomials of a fixed degree to force them to interpolate the given function values or data. For example, a piecewise linear Lagrange polynomial is a continuous function which is a linear polynomial over each subinterval. Clearly, a piecewise linear Lagrange polynomial over the subdivision of $[0,1]$ given in (4.6) which is zero at $x=0$ and $x=1$ is an element of S_0^h.

We state the estimates for the error in a function in $H^1(0,1)$ and its S^h-interpolant where S^h is the space of piecewise linear functions defined over the given partition with no boundary conditions imposed; i.e.,

$$S^h = \{\phi(x) \in C[0,1] \ : \ \phi(x) \text{ linear on } [x_i, x_{i+1}] \text{ for } 0 \le i \le N\}\,. \tag{4.17}$$ {1

Then these results also hold for $S_0^h \subset H_0^1(0,1)$. If $v(x)$ is a continuous function on $[0,1]$ then we can find a unique element which agrees with $v(x)$ at each of the points x_i, $i = 0,\ldots,N+1$; we call this element of S^h the S^h-interpolant of v and denote it by $I^h v$. Once we have the standard estimate for the approximation of a function by its piecewise linear Lagrange interpolant measured in the H^1-norm, then, we can use it in (4.16) to obtain an estimate in terms of powers of h. The following lemma gives standard results for approximating a function by its piecewise linear Lagrange interpolant in the L^2 and H^1 norms; see [Prenter] for details.

Lemma 4.2. *Let $f \in H^1(0,1)$ and $S^h \subset H^1(0,1)$ be defined by (4.17); let $I^h f$ denote the S^h-interpolant of f. Then there exists positive constants C_1, C_2, and C_3, independent of h and f, such that*

{1d_interl2orderh}

$$\left\|f - I^h f\right\|_0 \leq C_1 h \left\|f\right\|_1 . \qquad (4.18)$$

In addition, if $f \in H^2(0,1)$ then

{1d_interl2orderhsq}

$$\left\|f - I^h f\right\|_0 \leq C_2 h^2 \left\|f\right\|_2 \qquad (4.19)$$

and

{1d_interh1}

$$\left\|f - I^h f\right\|_1 \leq C_3 h \left\|f\right\|_2 . \qquad (4.20)$$

It is important to note that if the solution u to our problem is not smooth enough, i.e.,$u \in H^1(0,1)$ and $u \notin H^2(0,1)$, then (4.19) and (4.20) do not hold. In this situation we only have (4.18) and $\left\|u - I^h u\right\|_1 \leq C \left\|u\right\|_1$; the latter implying that there is no convergence in h; i.e.,as $h \to 0$, $\left\|u - I^h u\right\|_1$ does not approach zero. We say that the rate of convergence in (4.19) is order h squared, which is quadratic convergence, and denote it $\mathcal{O}(h^2)$; similarly the rate of convergence in (4.20) is $\mathcal{O}(h)$ which is linear convergence. From (4.19) and (4.20) we see a pattern arising that the error in the interpolant measured in the L^2 norm is one order higher than the error measured in the H^1 norm; this is due to the fact that the H^1 norm measures errors in the derivatives as well as the function values.

We can now use Lemma 4.2 to state an estimate for the error in u and u^h measured in the H^1-norm in terms of powers of h. We require $u \in H^2(0,1) \cap H^1_0(0,1)$; note that this can be guaranteed if $f, q, p \in L^2(0,1)$. In this case we get the *optimal* rate; this means that we get the same rate of convergence as $h \to 0$ as the interpolant.

{1d_thm_h1error} **Theorem 4.3.** *Let $u \in H^2(0,1) \cap H^1_0(0,1)$ and let u^h be the Galerkin approximation of u in the space S^h_0 defined by (4.7); i.e.,u^h satisfies (4.4). Then there exists a positive constant C, independent of u, h, or u^h such that*

{1d_homodir_h1error}

$$\left\|u - u^h\right\|_1 \leq C h \left\|u\right\|_2 . \qquad (4.21)$$

Proof. The proof is an obvious consequence of (4.20) and (4.16). ∎

It is often the case that we are interested in estimating the error in just the function itself and not its derivatives; in this case we want an estimate for the error in the L^2-norm. From the definition of the L^2- and H^1-norms we immediately have that

$$\left\|u - u^h\right\|_0 \leq \left\|u - u^h\right\|_1 \leq C h \left\|u\right\|_2 ,$$

the latter inequality holding if $u \in H^2(0,1) \cap H^1_0(0,1)$. However, Lemma 4.2 suggests that we should be able to improve the error to $\mathcal{O}(h^2)$; in addition, computations indicate that $O(h^2)$ is attainable. In order to obtain an optimal L^2-estimate, we must assume sufficient smoothness on u and use a technique known as "Nitsche's trick".

Theorem 4.4. *Let $u \in H^2(0,1) \cap H^1_0(0,1)$ be the solution of (4.3) and let u^h be the Galerkin approximation of u in the space S^h_0 defined by (4.7) satisfying (4.4). Then there exists a positive constant C, independent of u, h, or u^h such that* {1

$$\left\|u - u^h\right\|_0 \leq Ch^2 \left\|u\right\|_2 . \tag{4.22}$$ {1

Proof. Let $e = u - u^h$ and let ψ be the unique function in $H^1_0(0,1)$ (whose existence and uniqueness is guaranteed by the Lax-Milgram theorem) satisfying

$$A(\psi, \phi) = (e, \phi) \quad \forall\, \phi \in H^1_0(0,1) . \tag{4.23}$$ {1

Since $e \in H^1_0(0,1)$ we can set $\phi = e$ in the above expression to obtain

$$\left\|e\right\|^2_0 = (e,e) = A(\psi, e) .$$

Now Galerkin orthogonality for this problem guarantees that $A(u - u^h, v^h) = 0$ for all $v^h \in S^h_0$ and thus $A(e, v^h) = 0$ for all $v^h \in S^h_0$ and we can add this term without impunity. We know that $A(\cdot,\cdot)$ is linear and symmetric so we have

$$\left\|e\right\|^2_0 = A(\psi, e) - A(e, v^h) = A(e, \psi - v^h) \quad \forall\, v^h \in S^h_0 .$$

Using the boundedness of the bilinear form gives us

$$\left\|e\right\|^2_0 \leq C \left\|e\right\|_1 \left\|\psi - v^h\right\|_1 \quad \forall\, v^h \in S^h_0 .$$

We can use Theorem 4.3 to bound $\left\|e\right\|_1$ by $Ch\left\|u\right\|_2$. If we set v^h to be the S^h_0-interpolant of ψ then if $\psi \in H^1_0(0,1) \cap H^2(0,1)$ the estimate (4.20), along with Theorem 4.3 implies

$$\left\|e\right\|^2_0 \leq Ch^2 \left\|\psi\right\|_2 \left\|u\right\|_2 .$$

From the theory of elliptic partial differential equations one can show that if ψ is the solution to (4.23) and $\psi \in H^2(0,1) \cap H^1_0(0,1)$ then we can bound ψ by the L^2-norm of the data; i.e.,$\left\|\psi\right\|_2 \leq C\left\|e\right\|_0$. Substituting this bound for ψ into the above expression gives the desired result from

$$\left\|e\right\|^2_0 \leq Ch^2 \left\|e\right\|_0 \left\|u\right\|_2 .$$

■

It is important to realize that in order to get the optimal estimates in the L^2- and H^1-norms, we must have additional smoothness on our solution. This is a consequence of approximation theory, not an artifact of our finite element analysis.

When we present some numerical simulations, we see that a loss in accuracy occurs if our solution is not smooth enough.

We have now completed our analysis of a finite element solution of (4.1) using continuous, piecewise linear polynomials. Before turning our attention to implementing the method to obtain some numerical results we consider approximating using higher degree polynomials and then remind ourselves how the entries in the matrix and right-hand side of (4.10) are obtained.

4.1.3 Approximation using higher degree polynomials

{1d_sec_higher}

From the error estimate (4.21) we see that the rate of convergence is linear in the H^1 norm. If we want our calculations to converge at a higher rate, such as quadratically, then we have to choose a higher degree polynomial for our approximating space S_0^h. In this section we give some general results for the error in the interpolating polynomial for a kth degree polynomial and then use these to get optimal error estimates for our problem. We also consider a basis for quadratic polynomials and the structure of the resulting linear system which is no longer tridiagonal as it was when we used linear polynomials. The case of continuous, cubic polynomials is left to the exercises.

We now define S^h to be the space of continuous, piecewise polynomials of degree k or less over the partition of $[0,1]$ defined in (4.6), i.e.,

{1d_shk}

$$\begin{aligned} S^h = \{\phi(x) \ : \ \phi \in C[0,1], \phi(x) \text{ polynomial of} \\ \text{degree} \le k \text{ on } [x_i, x_{i+1}] \text{ for } 0 \le i \le N\} \,. \end{aligned} \tag{4.24}$$

S_0^h is defined in the same way except we require $\phi(x)$ to be zero at the endpoints;

{1d_shzerok}

$$\begin{aligned} S_0^h = \{\phi(x) \ : \ \phi \in C[0,1], \phi(x) \text{ polynomial of} \\ \text{degree} \le k \text{ on } [x_i, x_{i+1}] \text{ for } 0 \le i \le N, \phi(0) = \phi(1) = 0\} \,. \end{aligned} \tag{4.25}$$

A theorem for the S^h-interpolant of functions in H^1 is provided in the following lemma.

{1d_thm_interpolant} **Lemma 4.5.** *Let $f \in H^{k+1}(0,1)$ and $S^h \subset H^1(0,1)$ where S^h is defined by (4.24); let $I^h f$ denote the S^h-interpolant of f. Then there exists positive constants C_1, C_2, independent of h and f, such that*

{1d_interl2orderhkp1}

$$\left\| f - I^h f \right\|_0 \le C_1 h^{k+1} \left\| f \right\|_{k+1} \tag{4.26}$$

and

{1d_interh1orderk}

$$\left\| f - I^h f \right\|_1 \le C_2 h^{k} \left\| f \right\|_{k+1} \,. \tag{4.27}$$

Note that (4.26) reduces to (4.19) and (4.27) reduces to (4.20) when $k = 1$. These are the best rates of convergence possible with a kth degree polynomial. If f is not in $H^{k+1}(0,1)$ then there is a loss in the rates of convergence. For example, if $f \in H^k(0,1)$ and not in $H^{k+1}(0,1)$, then a power of h is lost in each estimate. If our finite element solution is in $H^{k+1}(0,1)$ then optimal rate of convergence in the H^1 norm are given in the following theorem.

{1d_thm_h1errork} **Theorem 4.6.** *Let $u \in H^{k+1}(0,1) \cap H_0^1(0,1)$ be the solution of (4.3) and let u^h be the Galerkin approximation of u in the space S_0^h defined by (4.25) satisfying (4.4). Then there exists a positive constant C, independent of u, h, or u^h such that*

$$\left\| u - u^h \right\|_1 \leq C h^k \left\| u \right\|_{k+1} . \tag{4.28}$$

{1

We note that this estimate says that if the solution is sufficiently smooth, then increasing the degree of the polynomial by one increases the rate of convergence by one.

As before, we are often interested in the L^2 norm of the error. We can mimic the proof of Theorem 4.4 to get the following result when S_0^h is defined by (4.25). See the exercises for details.

Theorem 4.7. *Let $u \in H^{k+1}(0,1) \cap H_0^1(0,1)$ be the solution of (4.3) and let u^h be the Galerkin approximation of u in the space S_0^h defined by (4.25) satisfying (4.4). Then there exists a positive constant C, independent of u, h, or u^h such that* {1

$$\left\| u - u^h \right\|_0 \leq C h^{k+1} \left\| u \right\|_{k+1} . \tag{4.29}$$

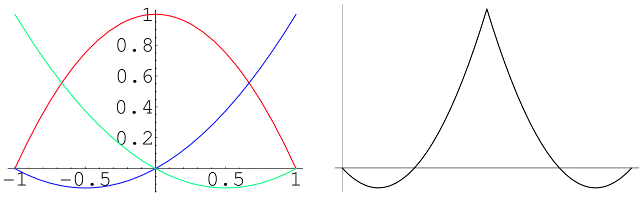
{1

We note that the optimal rate of convergence in the L^2 norm is one power of h higher than in the H^1 norm which measures the error in the derivatives of the solution as well as the solution itself.

We now turn to the concrete problem of finding a basis for S^h or S_0^h when we choose quadratic polynomials, i.e., $k = 2$. In this case we know that the rates of convergence are $\mathcal{O}(h^2)$ in the H^1 norm and $\mathcal{O}(h^3)$ in the L^2 norm, if the solution is sufficiently smooth. We use the same partition of $[0,1]$ as before, i.e., that given in (4.6). The problem now is that over each element $[x_{i-1}, x_i]$ the basis function must be a quadratic; however, it takes three points to uniquely determine a quadratic. To this end, we add a node in each subinterval; the easiest thing to do is add a node at the midpoint of each subinterval, $x_{i-\frac{1}{2}} = (x_{i-1} + x_i)/2$. We still have $N+1$ elements, but now have the $N+2$ points from the endpoints of the intervals plus the $N+1$ midpoints giving a total of $2N+3$ points. Analogous to the continuous, piecewise linear case, we expect that a basis for S^h for $k = 2$ consists of $2N+3$ elements and for S_0^h we don't need the endpoints so we have $2N+1$ elements in basis.

For simplicity of exposition, we renumber our $2N+3$ nodes as x_i, $i = 0, \ldots, 2N+2$. However, we must remember that the elements are $[x_{2j-2}, x_{2j}]$ for $j = 1, \ldots, N+1$. To determine a nodal basis for S^h we require each ϕ_i in the basis to have the property that it is one at node x_i and zero at all other nodes. In the basis for piecewise linear polynomials we were able to make the support of the basis functions to be two adjacent elements; the same is true in this case. However, now we have two different formulas for the basis functions determined by whether the function is centered at an endpoint of an interval or the midpoint.

To easily get an idea what these quadratic functions look like, we first write the polynomials on $[-1,1]$ with nodes $x=-1,0,1$; we can then translate them to the desired interval. From these we can determine the shape of our basis functions. For the quadratic function which is one at the midpoint, i.e., $x=0$, and zero at $x=\pm 1$ we have $\phi(x)=1-x^2$. For the quadratic function which is one at $x=-1$ and zero at $x=0,1$ we have $\phi(x)=\frac{1}{2}(x^2-x)$. Similarly for a quadratic function which is one at $x=1$ and zero at $x=-1,0$ we have $\phi(x)=\frac{1}{2}(x^2+x)$. These functions are illustrated in Figure 4.1 and have the same shape as the ones on $[x_{2j-2},x_{2j}]$. We can splice together the two functions centered at the endpoints of the interval to get a complete picture of the basis function centered at an endpoint which has support over two intervals; this is demonstrated in the right plot in Figure 4.1. Note that analogous to the case of continuous piecewise linear polynomials the quadratic basis functions will be in C^0 but not C^1.



{1d_fig_quad_basis}

Figure 4.1. *Plot on left shows nodal quadratic functions on* $[-1,1]$ *and plot on right shows shape of quadratic basis function centered at endpoint of an interval having support over two intervals.*

To find the analogous polynomials on $[x_{2j-2},x_{2j}]$ we need to translate our functions on $[-1,1]$ to the desired interval or equivalently solve linear systems. For example, a straightforward way to find the quadratic which is one at x_{2j-1} and zero at the endpoints is to solve

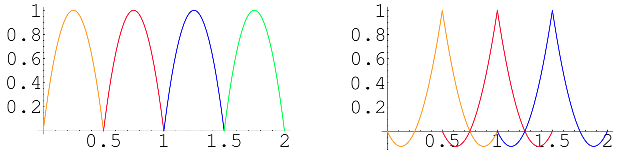
$$\begin{aligned} 0 &= a + b\big(x_{2j-2}\big) + c\big(x_{2j-2}\big)^2 \\ 1 &= a + b\big(x_{2j-1}\big) + c\big(x_{2j-2}\big)^2 \\ 0 &= a + b\big(x_{2j}\big) + c\big(x_{2j}\big)^2 . \end{aligned}$$

In later chapters we discuss more efficient approaches to finding basis functions. The support of the quadratic basis functions for S_0^h on a uniform partition of $[0,2]$ with $h=0.5$ are illustrated in Figure 4.2.

We have seen that once a basis for the finite dimensional space is chosen, the discrete problem can be converted to solving a linear system of equations. The (i,j) entry of the coefficient matrix $\mathcal{A}$ is given by the same expression as in the case of piecewise linear functions except we are using a different basis; specifically, we have

$$\mathcal{A}_{ij} = (p\phi_j', \phi_i') + (q\phi_j, \phi_i)$$

where ϕ_i is now a quadratic polynomial. We recall that when the standard "hat" functions were used as a basis for S_0^h the resulting matrix was $N\times N$, symmetric,



{1d_fig_quad_basis2} **Figure 4.2.** *Support of quadratic basis functions on the uniform partition of* $[0,2]$ *with* $h = .5$ *assuming homogeneous Dirichlet boundary conditions.*

positive definite and tridiagonal. In the case of our quadratic basis functions in S_0^h, the matrix is still symmetric and positive definite but we note that the size of our matrix has increased to $2N+1$. Also, it is no longer tridiagonal. To determine the bandwidth of the matrix, we need to ascertain where the zero entries begin in each row. We return to Figure 4.2 and note that for a basis function ϕ_i centered at a midpoint node x_i, the integral $\int_0^1 \phi_i \phi_j \, dx$ is zero when $j > i+1$ or $j < i-1$, i.e., outside of the interval; the same is true for the term $\int_0^1 \phi_i' \phi_j' \, dx$. However, for a basis function ϕ_i centered at the right endpoint node x_i, the integral $\int_0^1 \phi_i \phi_j \, dx$ is potentially nonzero in that interval and the next which includes a total of five basis functions, counting itself. Thus the integral is zero when $j > i+2$ or $j < i-2$ and the maximum bandwidth of the matrix is five. This system can be efficiently solved by a direct method such as a banded Cholesky algorithm or an iterative method such as conjugate gradient or one of its variants.

If we desire to have a method which converges cubically in the H^1 norm, then we can choose continuous, piecewise cubic polynomials for S^h. Because we need four points to uniquely determine a cubic, we add two points to each interval in our original partition given in (4.6). For S_0^h we now have $N + 2(N+1) = 3N+2$ points and we expect that this is the dimension of the space and thus the dimension of the resulting matrix. The shape of the basis functions and the structure of the resulting matrix is explored in the exercises.

4.1.4 Numerical quadrature

{1

If we are implementing our example given in (4.1) in the case $p = q = 1$ with continuous, piecewise linear polynomials for S_0^h and where we are using a uniform grid, then (4.11) and (4.12) explicitly give the coefficient matrices. However, entries in the right-hand side of (4.10) must be computed and also entries for the coefficient matrix for general p, q. For some choices of f we could evaluate the integrals exactly. However, if we want to write a general finite element program then we should be able to do problems where the integrals can not be evaluated exactly. In this case, we must use *quadrature rules* to approximate the integrals. Recall that in our error analysis, we have assumed that the integrals are computed exactly; the effects of numerical integration are discussed in a later chapter. For now, we present some widely used quadrature formulas in one-dimension and give general rules for choosing a formula.

In numerical integration we approximate the integral by the sum of the integrand evaluated at a prescribed set of points multiplied by weights; i.e.,

{numint}

$$\int_a^b f(x)\,dx \approx \sum_k f(q_k)w_k\,, \tag{4.30}$$

where q_k represent the quadrature points and w_k the quadrature weights. Of particular interest in one dimension are the Gauss quadrature rules; in these rules the quadrature points and weights are chosen so that the rule integrates exactly as high a degree polynomial as possible. Specifically, if we use n Gaussian quadrature points then the rule integrates polynomials of degree $2n-1$ exactly. The Gaussian quadrature rule for one point is the well known midpoint rule. The following table gives the Gaussian quadrature points and weights on the interval $[-1,1]$.

{gaussquad1d}

Table 4.1. *Gauss quadrature formulas on* $[-1,1]$

n	nodes	weights
1	0.0000000000	2.0000000000
2	$\pm\frac{1}{\sqrt{3}} = \pm 0.5773502692$	1.0000000000
3	±0.7745966692 0.0000000000	0.5555555556 0.8888888889
4	±0.8611363116 ±0.3399810436	0.3478548451 0.6521451549
5	±0.9061798459 ±0.5384693101 0.0000000000	0.2369268850 0.4786286701 0.5688888889

If the domain of integration is different from $(-1,1)$, then a change of variables is needed. For example, to compute the integral $\int_a^b f(\hat{x})\ d\hat{x}$ we use the linear mapping $\hat{x} = a + \frac{b-a}{2}(x+1)$ to map to the integral over $(-1,1)$. In this case we have

$$\int_a^b f(\hat{x})\ d\hat{x} = \frac{b-a}{2}\int_{-1}^1 f(a + \frac{b-a}{2}(x+1)\ dx\,.$$

Then we apply the quadrature rule to the integral over $(-1,1)$. Note that we have just modified the quadrature weight by multiplying by $\frac{b-a}{2}$ and mapping the quadrature point to the interval (a,b).

When choosing a quadrature rule, we want to use as low a degree rule as possible for efficiency but as high a degree rule as necessary for accuracy. It is not necessary to evaluate the integrals exactly, even if this is possible; however, we must assure that the error in the numerical quadrature does not contaminate the power of h accuracy in our estimate. When using piecewise linear polynomials for the finite element space in one-dimension for the problem (4.1), it is adequate to use a one-point Gauss quadrature rules; for piecewise quadratic polynomials a two-point rule is adequate.

4.1.5 Computational examples

In this section we implement two specific examples of the boundary value problem given in (4.1) where we know the exact solution so that errors and rates of convergence can be calculated. These problems differ in the choice of p, q and f. The choice of f is especially important because a lack of smoothness in f results in the solution not being smooth enough to guarantee the optimal rates of convergence. In all computations we use continuous, piecewise polynomials on a uniform grid, an appropriate Gauss quadrature rule to evaluate the integrals in the coefficient matrix and the right-hand side, and a direct solver for the linear system. For the error computation we use a higher order quadrature rule to evaluate the integrals. The reason for the higher order rule in the error computation is to make absolutely sure that no error from the numerical integration contaminates the calculation of the error. The computations are performed using $h = 1/4, 1/8, 1/16$, and $1/32$ with linear, quadratic and cubic elements; the H^1- and L^2-errors are computed for each grid.

For each example we are interested in calculating the numerical rate of convergence and comparing it with the theoretical results presented in Theorems 4.3, 4.4,**??** and **??**. The errors for each grid can be used to compute an approximate rate of convergence. For example, we have $\|u - u^h\| \approx Ch^r$ where we expect r to approach some value as the grid size decreases. If we have the error, E_i, on two separate meshes then we have that $E_1 \approx Ch_1^r$ and $E_2 \approx Ch_2^r$ where E_1 and E_2 represent $\|u - u^h\|$ on the grid with mesh spacing h_1 and h_2, respectively. If we solve for C and set the two relationships equal, we have $E_1/h_1^r \approx E_2/h_2^r$; solving for r we obtain

$$r \approx \frac{\ln E_1/E_2}{\ln h_1/h_2}\,. \tag{4.31}$$ {1

We note that if the grid spacing is halved, i.e., $h_2 = h_1/2$ then the error should be approximately decreased by $\left(\frac{1}{2}\right)^r$ since $E_2 \approx \left(\frac{h_2}{h_1}\right)^r E_1$. This implies that if $r = 1$ the error is approximately halved when the grid spacing is halved; if the rate is two, then the error is reduced by a factor of one-fourth when the grid spacing is halved, etc.

{1

Example 4.8 We first consider the problem

$$\begin{aligned} -u'' + \pi^2 u &= 2x\pi^2 \sin \pi x - 2\pi \cos \pi x \quad \text{for } 0 < x < 1 \\ u(0) = u(1) &= 0\,, \end{aligned} \tag{4.32}$$ {1

whose exact solution is given by $u = x \sin \pi x$. Since our solution $u(x) = x \sin \pi x$ is actually in $C_0^\infty(0,1)$ we expect the optimal rates of convergence; in particular if we use continuous, piecewise linear polynomials then the rate r, calculated from (4.31), should approach two as $h \to 0$ for the L^2-norm and approach one for the H^1-norm. These values for r are calculated in Table 4.2 along with the errors and rates using continuous, piecewise quadratic and cubic polynomials; in the table we computed the rate using the errors at $h = 1/4$ and $1/8$, at $h = 1/8$ and $1/16$, and at $h = 1/16$ and $h = 1/32$. Note that, in fact, $r \to 1$ in the H^1 error and $r \to 2$ in the L^2-error as Theorems 4.3 and 4.4 predict when piecewise linear polynomials are used; the optimal rates for quadratic and cubic polynomials are also obtained. In these calculations we used a one-point Gauss rule for linear polynomials, a two-point Gauss rule for quadratic polynomials, and a three-point Gauss rule for cubic polynomials. In Table 4.3 we illustrate what happens if we use continuous quadratic polynomials using a one-point, a two-point and a three-point Gauss quadrature rule. Note that the rates of convergence using a two-point and a three-point rule are essentially the same, but when we use the one-point rule the results are meaningless. ■

Table 4.2. *Numerical results for Example 4.8 using continuous, piecewise linear polynomials.*

_homodir_example1_linear}

p^k	h	$\lVert u-u^h \rVert_1$	rate	$\lVert u-u^h \rVert_0$	rate
linear	1/4	0.47700		0.28823×10^{-1}	
linear	1/8	0.23783	1.0041	0.69831×10^{-2}	2.0459
linear	1/16	0.11885	1.0007	0.17313×10^{-2}	2.0120
linear	1/32	0.059416	1.0002	0.43199×10^{-3}	2.0028
quadratic	1/4	0.49755×10^{-1}		0.15707×10^{-2}	
quadratic	1/8	0.12649×10^{-1}	1.9758	0.20227×10^{-3}	2.9570
quadratic	1/16	0.31747×10^{-2}	1.9940	0.25553×10^{-4}	2.9847
quadratic	1/32	0.79445×10^{-3}	1.9986	0.32031×10^{-5}	2.9960
cubic	1/4	0.51665×10^{-2}		0.10722×10^{-3}	
cubic	1/8	0.64425×10^{-3}	3.003	0.67724×10^{-5}	3.985
cubic	1/16	0.80496×10^{-4}	3.001	0.42465×10^{-6}	3.9953
cubic	1/32	0.10061×10^{-4}	3.000	0.26564×10^{-7}	3.9987

{1d_example_2}

Example 4.9 The next problem we want to consider is

{1d_homodir_example2}

$$\begin{aligned} -u'' &= -\alpha(\alpha-1)x^{\alpha-2} \quad \text{for } 0<x<1 \\ u(0)=u(1) &= 0\,, \end{aligned} \tag{4.33}$$

where $\alpha > 0$; the exact solution u is given by $u(x) = x^\alpha - x$. The results for various values of α are presented in Table 4.4 using continuous, piecewise linear polynomials and a one-point Gauss quadrature rule. Recall that the optimal rates in this case are $\mathcal{O}(h)$ in the H^1 norm and $\mathcal{O}(h^2)$ in the L^2 norm. Note that for $\alpha = 7/3$ we get the optimal rates of convergence. However, for $\alpha = 4/3$ we have less than optimal rates and for $\alpha = 1/3$ the H^1-error is almost constant and the rate in the L^2-norm is less than one. Of course, the reason for this is that when $\alpha = 3/2$ the exact solution $u = x^{4/3} - x \notin H^2(0,1)$ and when $\alpha = 1/3$ the exact solution $u = x^{1/3} - x \notin H^1(0,1)$. Thus the interpolation results (4.19) and (4.20) do *not* hold and hence Theorems 4.3 and 4.4 do not apply. ■

Table 4.3. *Numerical results for Example 4.8 using continuous, piecewise quadratic polynomials with three different quadrature rules.*

le_homodir_example1_quad}

Gauss Quadrature Rule	h	$\left\|u-u^h\right\|_1$	rate	$\left\|u-u^h\right\|_0$	rate
one-point	1/4	8.885		0.3904	
one-point	1/8	18.073		0.3665	
one-point	1/16	36.391		0.3603	
one-point	1/32	72.775		0.3587	
two-point	1/4	0.49755×10^{-3}		0.15707×10^{-4}	
two-point	1/8	0.12649×10^{-3}	1.9758	0.20227×10^{-5}	2.9570
two-point	1/16	0.31747×10^{-4}	1.9940	0.25553×10^{-6}	2.9847
two-point	1/32	0.79445×10^{-5}	1.9986	0.32031×10^{-7}	2.9960
three-point	1/4	0.49132×10^{-3}		0.18665×10^{-4}	
three-point	1/8	0.12109×10^{-3}	1.9620	0.24228×10^{-5}	2.9456
three-point	1/16	0.31724×10^{-4}	1.9911	0.30564×10^{-6}	2.9868
three-point	1/32	0.79430×10^{-5}	1.9978	0.38292×10^{-7}	2.9967

Table 4.4. *Numerical results for Example 4.9.*

d_table_homodir_example2}

α	h	$\left\|u-u^h\right\|_1$	rate	$\left\|u-u^h\right\|_0$	rate
7/3	1/4	0.1747		0.17130×10^{-1}	
7/3	1/8	0.08707	1.0046	0.33455×10^{-2}	1.9726
7/3	1/16	0.04350	1.0012	0.84947×10^{-3}	1.9776
7/3	1/32	0.02174	1.0007	0.21495×10^{-3}	1.9826
4/3	1/4	0.47700		0.28823×10^{-1}	
4/3	1/8	0.23783	0.7690	0.69831×10^{-2}	1.8705
4/3	1/16	0.11885	0.7845	0.17313×10^{-2}	1.8834
4/3	1/32	0.059416	0.7965	0.43199×10^{-3}	1.8005
1/3	1/4	0.43332		0.14594	
1/3	1/8	0.43938		0.10599	0.4615
1/3	1/16	0.46661		0.07922	0.4200
1/3	1/32	0.50890		0.06064	0.3857

4.2 A two-point BVP with Neumann boundary data

{1d_sec_homoneu}

In this section we consider the same differential equation as in the first section but now we impose Neumann boundary data instead of homogeneous Dirichlet data. In particular we seek a function $u(x)$ satisfying

$$\begin{aligned} -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)+q(x)u &= f(x) \quad \text{for } 0<x<1 \\ u'(0) &= 0 \\ u'(1) &= \alpha\,. \end{aligned} \tag{4.34}$$

{1

As before, p and q are bounded functions on $[0,1]$ satisfying $0 < p_{\min} \le p(x) \le p_{\max}$ but now we impose $0 < q_{\min} \le q(x) \le q_{\max}$ for all $x \in [0,1]$. Again if $f, q \in C[0,1]$ and $p \in C^1[0,1]$ the boundary value problem (4.34) possesses a unique *classical solution* $u(x) \in C^2(0,1)$ which satisfies (4.34) for every $x \in [0,1]$. Note that here we require that $q_{\min} > 0$ to guarantee a unique solution; this is because if $q = 0$ and u satisfies (4.34) then so does $u + C$ for any constant C.

In this case our underlying finite element space is $H^1(0,1)$ because we have no boundary conditions to impose on the space. The weak formulation is

{1d_homoneu_weak}

$$\left\{ \begin{array}{l} \text{seek } u \in H^1(0,1) \text{ satisfying} \\ \qquad A(u,v) = (f,v) + \alpha p(1) v(1) \quad \forall\, v \in H^1(0,1)\,, \end{array} \right. \tag{4.35}$$

where

$$A(v,w) = \int_0^1 p(x) v'(x) w'(x)\,dx + \int_0^1 q(x) v(x) w(x)\,dx \quad \forall\, v,w \in H^1(0,1)\,.$$

Clearly, if $u(x)$ satisfies the classical problem (4.34), then $u(x)$ satisfies (4.35) because

$$\begin{aligned} \int_0^1 f(x) v\,dx &= \int_0^1 \big(-(p(x)u'(x))' + q(x)u(x)\big) v(x)\,dx \\ &= -pu'v\big|_0^1 + \int_0^1 p(x)u'(x)v'(x)\,dx + \int_0^1 q(x)u(x)v(x)\,dx \\ &= -p(1)u'(1)v(1) + p(0)u'(0)v(0) + A(u,v) \\ &= A(u,v) - \alpha p(1)v(1)\,, \end{aligned}$$

where we have imposed the homogenous Neumann boundary condition $u'(0) = 0$ and the inhomogeneous condition $u'(1) = \alpha$. Note that these boundary conditions are *not* imposed on the space, but rather on the weak formulation; these are called *natural* boundary conditions.

In a manner similar to the example in Section 4.1, we can show that the hypotheses of the Lax-Milgram theorem are satisfied. Recall that in proving coercivity for the previous example, we used the Poincaré inequality to relate the L^2 norm with the H^1 seminorm. We can not longer do this because our function is not zero on any portion of the boundary. However, coercivity can be proved in a straightforward manner; the details are left to the exercises. Thus we are guaranteed the existence and uniqueness of a solution to (4.35).

If we want to seek an approximation to $u(x)$ in the space of continuous, piecewise linear functions defined over the subdivision (4.6) then we cannot use the space S_0^h defined in (4.7) since this space was designed to approximate functions in $H_0^1(0,1)$. Instead we consider S^h where

{1d_homoneu_sh}

$$S^h = \{\phi(x) \in C[0,1], \phi(x) \text{ linear on } (x_i, x_{i+1}) \text{ for } 0 \le i \le N\}\,. \tag{4.36}$$

Similar to the homogeneous Dirichlet case, it can be shown that S^h is an $N+2$ dimensional subspace of $H^1(0,1)$; a basis for S^h is given by the standard "hat" functions that we used for S_0^h along with one defined at each endpoint. Specifically, we have the functions ψ_i, $i = 1, \ldots, N+2$ defined by

{1d_homoneu_basis}

$$\psi_i(x) = \begin{cases} \phi_0(x) & \text{for } j = 1 \\ \phi_{i-1}(x) & \text{for } 2 \leq i \leq N+1 \\ \phi_{N+1}(x) & \text{for } j = N+2 \end{cases} \tag{4.37}$$

where $\phi_i(x)$, $i = 1, \ldots N$ are given by (4.8) and

$$\phi_0(x) = \begin{cases} \dfrac{x_1 - x}{h_1} & \text{for } 0 \leq x \leq x_1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.38}$$ {1

and

$$\phi_{N+1}(x) = \begin{cases} \dfrac{x - x_N}{h_{N+1}} & \text{for } x_N \leq x \leq 1 \\ 0 & \text{elsewhere}\,. \end{cases} \tag{4.39}$$ {1

Galerkin's theorem guarantees that there is a unique $u^h \in S^h \subset H^1(0,1)$ satisfying

$$A(u^h, v^h) = (f, v^h) + \alpha p(1) v^h(1) \quad \forall\, v^h \in S^h\,. \tag{4.40}$$ {1

The problem of finding a $u^h \in S^h$ which satisfies (4.40) reduces to solving a linear system of equations; in this case the coefficient matrix has dimension $N+2$. In addition, we can use the interpolation results given in Lemma 4.2 to obtain the following optimal error estimates. See the exercises for a proof.

{1

Theorem 4.10. *Let $u \in H^2(0,1)$ be the solution of (4.34) and let u^h be the Galerkin approximation in S^h defined by (4.36) given by (4.40). Then for some constant C, independent of h, u, and u^h we have*

$$\left\| u - u^h \right\|_k \leq C h^{2-k} \left\| u \right\|_2$$

for $k = 0, 1$.

One purpose of the following computations is to demonstrate the difference in satisfying a boundary condition by imposing it on the space (an essential boundary condition) and imposing it weakly through the weak formulation (a natural boundary condition).

{1

Example 4.11 We consider the problem

$$\begin{aligned} -u'' + \pi^2 u &= 2x\pi^2 \sin \pi x - 2\pi \cos \pi x \quad \text{for } 0 < x < 1 \\ u'(0) &= 0 \\ u'(1) &= -\pi\,, \end{aligned} \tag{4.41}$$ {1

whose exact solution is given by $u = x \sin \pi x$. Note that this is the same differential equation as in Example 4.8 but now we are imposing Neumann boundary conditions. Since our solution $u(x) = x \sin \pi x$ is actually in $C^\infty(0,1)$ we expect the optimal rates of convergence which we can see are obtained from Table 4.5. The approximate solutions using uniform grids of $h = \frac{1}{4}$, $\frac{1}{8}$ and $\frac{1}{16}$ along with the exact solution are plotted in Figure 4.3. Note that although our exact solution is zero at the endpoints, our approximate solution is not because we imposed Neumann boundary conditions. However, the approximate solution does not satisfy the exact derivative boundary condition because we have satisfied it weakly. In the last plot in Figure 4.3 we have blown up the approximate solutions at the right end point which should have a slope of $-\pi$. The approximate derivative at the right boundary is -1.994, -2.645, -2.917 and -3.036 for $h = 1/4$, 1/8, 1/16, and 1/32 respectively. These correspond to errors of 1.147, 0.4968, 0.2244 and 0.1055. As $h \to 0$ the derivative of the approximate solution at $x = 1$ approaches the exact value of $-\pi$ linearly; this is expected because the rate of convergence in the H^1 norm is one. Note that this is in contrast to Example 4.8 where our approximate solution exactly satisfied the homogeneous Dirichlet boundary condition because we imposed it on our space. ■

{1d_table_example3}

Table 4.5. *Numerical results for Example 4.11 using continuous, piecewise linear polynomials.*

	h	$\|u-u^h\|_1$	rate	$\|u-u^h\|_0$	rate
1/4	0.48183		0.22942×10^{-1}		
1/8	0.23838	1.0153	0.56235×10^{-2}	2.0281	
1/16	0.11892	1.0033	0.13988×10^{-2}	2.0073	
1/32	0.059425	1.0009	0.34924×10^{-3}	2.0019	

4.3 A two-point BVP with inhomogeneous boundary data

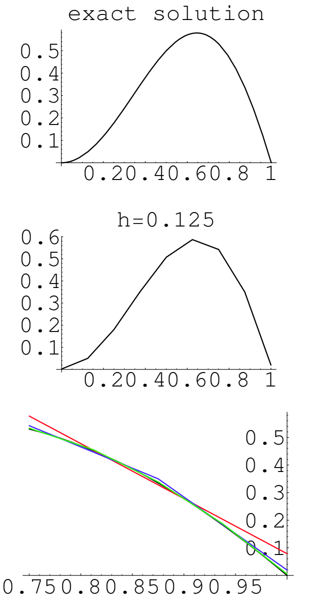
{1d_sec_inhomo}

In the previous two sections we considered two-point boundary values problems with homogeneous Dirichlet boundary data and homogeneous and inhomogeneous Neumann data. Consequently, the only type of boundary conditions that are left to see how to handle are inhomogeneous Dirichlet data and mixed, or Robin, boundary conditions. In this section we demonstrate how an inhomogeneous Dirichlet boundary condition can be handled; the mixed boundary condition is handled similarly to the inhomogeneous Neumann boundary condition. In particular we seek a function $u(x)$ satisfying

{1d_inhom_2ptbvp}

$$\begin{aligned} -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u &= f(x) \quad \text{for } 0 < x < 1 \\ u(0) = \alpha \qquad u'(1) + \sigma u(1) &= \beta\,, \end{aligned} \tag{4.42}$$

where α, β, and σ are constants. Note that if we choose $\sigma = 0$ then we just have an inhomogeneous Neumann boundary condition at the right endpoint as we did in (4.34).



{1d_fig_neu} **Figure 4.3.** *Plots of the exact solution and three piecewise linear approximations. The last plot gives a blow-up of the right endpoint demonstrating that the natural boundary condition is only satisfied weakly.*

We know that the underlying Hilbert space for the weak formulation should be $H^1(0,1)$ or some subspace.. For the Dirichlet boundary condition in Section 4.1 we imposed the boundary condition on the space; i.e.,we sought our solution in the subspace of $H^1(0,1)$ consisting of all functions that were zero on the boundary. However, we can not constrain our space to be all functions $\phi \in H^1(0,1)$ which satisfy $\phi(0)=\alpha$. The reason is that this is *not* a subspace of $H^1(0,1)$ since, for example, if $v(0)=\alpha$ and $w(0)=\alpha$ then $(v+w)(0)=2\alpha$.

Inhomogeneous Dirichlet boundary conditions can be handled in several ways. One of the easiest ways to handle them *theoretically* is to transform the problem into one which has homogeneous Dirichlet boundary data. In our problem we choose a function $g(x)\in H^1(0,1)$ such that $g(0)=\alpha$ and such that $g(x)$ is nonzero only on $[0,\xi]$ where $\xi<1$; the reason for the latter requirement is so that the boundary condition at $x=1$ is unaffected. We then define $w(x)=u(x)-g(x)$ so that $w(0)=u(0)-g(0)=0$. Because we have converted the problem to one for $w=u-g$ with $g(x)$ zero outside $[0,\xi]$, $\xi<1$ we have the same boundary condition for $w'(1)$ as for $u'(1)$. The differential equation is now modified as

$$-\frac{d}{dx}\left(p(x)\frac{d(w+g)}{dx}\right)+q(x)(w+g)=f(x)\,.$$

Because $g(x)$ is a known function, the two-point boundary value problem for $w(x)$

becomes

$$\begin{aligned} -\frac{d}{dx}\left(p(x)\frac{dw}{dx}\right)+q(x)w &= f(x)+(p(x)g'(x))'-q(x)g(x) \quad \text{for } 0<x<1 \\ w(0) &= 0 \\ w'(1)+\sigma w(1) &= \beta\,, \end{aligned} \tag{4.43}$$

{1

The mixed boundary condition at the right boundary is handled in a similar manner to the inhomogeneous Neumann. In this case, instead of $w'(1)$ being set to a constant, we have $w'(1)=\beta-\sigma w(1)$. When we substitute this value in the weak form, the constant β goes to the right hand side of the equation because it is known whereas the term $\sigma w(1)$ is unknown and is incorporated in the bilinear form.

We now define a weak problem for the function $w(x)=u(x)-g(x)$. Let $\hat{H}^1(0,1)$ be the subspace of $H^1(0,1)$ consisting of all functions in $H^1(0,1)$ which are zero at $x=0$. Then we seek a $w\in\hat{H}^1(0,1)$ satisfying

{1d_inhom_weak}

$$\left\{\begin{array}{l} \text{seek } u\in\hat{H}^1(0,1) \text{ satisfying} \\ A(w,v)=(f,v)-A(g,v)+\beta p(1)v(1) \quad \forall\, v\in\hat{H}^1(0,1)\,, \end{array}\right. \tag{4.44}$$

where

{1d_inhom_bilinear}

$$A(w,v)=(pw',v')+(qw,v)+\sigma p(1)w(1)v(1)\,. \tag{4.45}$$

To demonstrate that a solution to (4.43) is also a solution to (4.44) we first note that

$$\begin{aligned} \int_0^1 \big[-(pw')'+qw\big]v\,dx &= \int_0^1 pw'v'\,dx+\int_0^1 qwv\,dx-p(1)w'(1)v(1)+p(0)w'(0)v(0) \\ &= \int_0^1 pw'v'\,dx+\int_0^1 qwv\,dx-p(1)\big(\beta-\sigma w(1)\big)v(1) \\ &= A(w,v)-\beta p(1)v(1)\,. \end{aligned}$$

Now the right-hand side of (4.43) can be written as

$$\begin{aligned} &\int_0^1 \big(f(x)+(p(x)g'(x))'\big)v(x)\,dx-\int_0^1 q(x)g(x)v(x)\,dx \\ &\qquad = (f,v)-\int_0^1 p(x)g'(x)v'(x)\,dx+p(1)g'(1)v(1)-p(0)g'(0)v(0) \\ &\qquad\qquad -\int_0^1 q(x)g(x)v(x)\,dx \\ &\qquad = (f,v)-\Big(\int_0^1 p(x)g'(x)v'(x)\,dx+\int_0^1 q(x)g(x)v(x)\,dx\Big) \\ &\qquad = (f,v)-A(g,v) \end{aligned}$$

where we have used the fact that $v\in\hat{H}^1(0,1)$ implies $v(0)=0$ and $g(1)=g'(1)=0$ because $g=0$ in $(\xi,1]$. Combining these two results demonstrates that if w satisfies the classical two-point boundary value problem (4.43) then w satisfies the weak problem (4.44).

Using similar techniques as before, we can demonstrate that $A(\cdot,\cdot)$ defined by (4.45) satisfies the conditions of the Lax-Milgram theorem and that the right-hand side of (4.44) denotes a bounded linear functional on the Hilbert space $\hat{H}^1(0,1)$. Then we have that there exists a unique solution $w \in \hat{H}^1(0,1)$ to (4.44). The generalized or weak solution u to (4.42) is given by $u = w + g$.

To find an approximate solution to (4.44) in the space of piecewise linear functions which are zero at $x = 0$ we define the $(N+1)$-dimensional subspace of $\hat{H}^1(0,1)$

$$\hat{S}^h = \{\phi \in \hat{H}^1(0,1) \ : \ \phi \text{ is piecwise linear on each subinterval and } \phi(0) = 0\}\,,$$

where we are using the mesh defined by

$$0 = x_0 < x_1 < \cdots < x_{N+1} = 1 \quad \text{where} \quad x_i = x_{i-1} + h_i, \quad 1 \leq i \leq N+1\,,$$

A basis for $\hat{S}^h$ is given by ϕ_i, $i = 1, \ldots, N+1$ where $\phi_1, \ldots, \phi_N$ are defined by (4.8) and ϕ_{N+1} is defined by (4.39). We choose $g(x) = a\phi_0(x)$ where $\phi_0(x)$ is the basis function defined by (4.38). Then the resulting linear system is given by $\mathcal{A}\vec{c} = \vec{b}$ where $\mathcal{A}_{ij} = A(\phi_j, \phi_i)$ for $1 \leq i, j \leq N+1$ and $\vec{b}_i = -A(a\phi_0, \phi_i) + (f, \phi_i) + bp(1)\phi_i(1)$ for $1 \leq i \leq N+1$. We note that the coefficient matrix in the resulting algebraic system has the same formulas for the entries as in our previous examples, the dimension is just $N+1$. However, the right-hand side has an additional contribution in the first entry due to the boundary condition at $x = 0$ and in the last position due to the inhomogeneous mixed boundary condition.

We should note that in more complicated problems in higher dimensions, it may not be so easy to construct the function g. To handle the problem theoretically, we can always assume such a function but implementing in a computer program may be more difficult. In later chapters we see different ways to implement inhomogeneous Dirichlet boundary data.

Summarizing, we see that the mixed boundary condition at $x = 1$ required no adjustment of the underlying Hilbert space but rather was "automatically" satisfied by our choice of the weak formulation. As before, such a boundary condition is called *natural.* On the other hand, the Dirichlet boundary condition required that we constrain our underlying Hilbert space so that the boundary condition is satisfied. This is another example of an *essential* boundary condition.

4.4 A fourth order example

{1

In this section we consider approximating the solution of a fourth order boundary value problem. In particular, we consider

$$\begin{aligned} \frac{d^2}{dx^2}\left(r(x)\frac{d^2u}{dx^2}\right) - \frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u(x) &= f(x) \quad 0 < x < 1 \\ u(0) = u(1) = 0 \qquad u''(0) = u''(1) &= 0\,, \end{aligned} \tag{4.46}$$

{1

where $r_{\max} \geq r(x) \geq r_{\min} > 0$ and $p_{\max} \geq p(x) \geq 0$, $q_{\max} \geq q(x) \geq 0$ for all $x \in [0,1]$. Other boundary conditions which can be applied are explored in the exercises.

This problem differs from the previous second order problem because when we perform a a single integration by parts we have three derivatives on the trial function and two on the test function. To balance the derivatives we need to perform a second integration by parts. An obvious choice for the bilinear form $A(\cdot,\cdot)$ is

{1d_fourth_bilinear}

$$A(v,w) = \int_0^1 \Big(r v'' w'' + p v' w' + q v w \Big)\, dx\,. \tag{4.47}$$

In this situation we immediately realize that due to the appearance of second derivatives we can no longer use $H^1(0,1)$ as our underlying Hilbert space; we must now use $H^2(0,1)$ which is the space of all functions in $L^2(0,1)$ which possess weak L^2 derivatives up to order two. The notation $H^2_0(0,1)$ is used for the space

{1d_htwozero}

$$H^2_0(0,1) = \{ v \in H^2(0,1) \ :\ v(0) = v(1) = v'(0) = v'(1) = 0 \}\,. \tag{4.48}$$

Because we have boundary conditions on u'' we don't need $H^2_0(0,1)$ so we consider the space $H^2(0,1) \cap H^1_0(0,1)$ which is the set of all functions v in $H^2(0,1)$ which satisfy $v(0) = 0$ and $v(1) = 0$. Then our weak formulation is to

{1d_fourth_weak}

$$\left\{ \begin{array}{l} \text{seek } u \in H^2(0,1) \cap H^1_0(0,1) \text{ satisfying} \\ \qquad A(u,v) = (f,v) \quad \forall\, v \in H^2(0,1) \cap H^1_0(0,1)\,. \end{array} \right. \tag{4.49}$$

If u is the classical solution of (4.46) then

$$\begin{aligned} (f,v) &= \int_0^1 \Big((ru'')'' - (pu')' + qu \Big) v\, dx \\ &= \int_0^1 \big(-(ru'')'v' + (pu')v' + quv \big)\, dx + ru''v'|_0^1 - pu'v|_0^1 \\ &= \int_0^1 \big(ru''v'' + pu'v' + quv \big)\, dx \\ &= A(u,v)\,, \end{aligned}$$

where we have imposed the boundary conditions $u''(0) = u''(1) = 0$ on the weak form and used the fact that $v \in H^2(0,1) \cap H^1_0(0,1)$ implies $v(0) = v(1) = 0$. In this case the boundary conditions $u''(0) = u''(1) = 0$ are *natural* boundary conditions and $u(0) = u(1) = 0$ are *essential* boundary conditions.

The proof that the bilinear form defined by (4.47) satisfies the hypotheses of the Lax-Milgram theorem is left to the exercises. In the sequel we assume that a unique solution to the weak problem can be guaranteed.

We now consider the approximate problem. An immediate consequence of having $H^2(0,1)$ as the underlying Hilbert space is that we can no longer approximate using continuous piecewise linear polynomials or even continuous piecewise polynomials of degree k. A space S^h consisting of piecewise polynomials satisfies $S^h \subset H^1(0,1)$ if and only if the functions in S^h are continuous; for $S^h \subset H^2(0,1)$ we require the functions and the first derivatives to be continuous. These results are formally proved in a general setting in a later chapter. As a consequence of using $H^2(0,1)$ as the underlying space we must now investigate piecewise polynomials which are in $C^1(0,1)$ so that we can guarantee them to be subspaces of $H^2(0,1)$. We consider two spaces: piecewise cubic Hermite polynomials and piecewise cubic splines.

4.4.1 Piecewise cubic Hermite polynomials

{1

In this section we consider a space of piecewise polynomials which are $C^1(0,1)$ and which are cubic on each subinterval of a partition of $[0,1]$; for simplicity of exposition we take a uniform partition. We define the space of piecewise cubic Hermite polynomials over the subdivision

$$0 = x_0 < x_1 < \cdots < x_{N+1} = 1 \quad \text{where } x_i = x_{i-1} + h, \quad h = \frac{1}{N+1}$$

to be all polynomials $\phi(x) \in C^1(0,1)$ which are cubic on each subinterval $[x_i, x_{i+1}]$. The dimension of this space is easily determined by considering the number of degrees of freedom and number of constraints we have. On each of the $N+1$ subintervals there are four degrees of freedom to determine a cubic yielding a total of $4N+4$ degrees of freedom; for the piecewise polynomial to be $C^1(0,1)$ we require continuity of the polynomial and its derivative at each of the N interior nodes yielding a total of $2N$ constraints. Combining these results, we see that this space is a $(2N+4)$-dimensional subspace of $H^2(0,1)$. We define the space $\mathcal{H}^h$ to be the space of piecewise cubic Hermite polynomials over the given partition; i.e.,

$$\begin{aligned}\mathcal{H}^h = \{\phi(x) \ : \ & \phi \in C^1(0,1), \\ & \phi(x) \text{ is a cubic polynomial on } [x_i, x_{i+1}],\ 0 \le i \le N\}\,. \end{aligned} \tag{4.50}$$ {1

Of course for our particular example we have to constrain this space to satisfy the homogeneous Dirichlet boundary conditions. However, we first consider a basis for $\mathcal{H}^h$.

A convenient way to establish a basis for $\mathcal{H}^h$ is to consider translations of functions defined on $[-1,1]$. In particular, we consider the piecewise cubic polynomials $\xi(x)$ and $\eta(x)$ defined by

$$\xi(x) = \begin{cases} (x+1)^2(-2x+1) & -1 \le x \le 0 \\ (x-1)^2(2x+1) & \quad 0 \le x \le 1 \end{cases}$$

and

$$\eta(x) = \begin{cases} x(x+1)^2 & -1 \le x \le 0 \\ x(x-1)^2 & \quad 0 \le x \le 1 \end{cases}$$

on $[-1,1]$ These polynomials are illustrated in Figure 4.4. Note that $\xi(x) \in C^1[-1,1]$, $\xi(\pm 1) = 0$, $\xi'(0) = 0$, and $\xi'(\pm 1) = 0$; also $\eta(x) \in C^1[-1,1]$, $\eta(0) = \eta(\pm 1) = 0$, $\eta'(\pm 1) = 0$, and $\eta'(0) = 1$.

We now translate these cubic polynomials to the interval $[x_{i-1}, x_{i+1}]$ for $i = 1, \ldots, N$ to obtain our basis elements $\xi_i(x)$ and $\eta_i(x)$. Specifically, we define

$$\xi_i(x) = \begin{cases} \xi(\frac{x}{h} - i) & x_{i-1} \le x \le x_{i+1} \\ 0 & \qquad \text{elsewhere} \end{cases} \tag{4.51}$$ {1

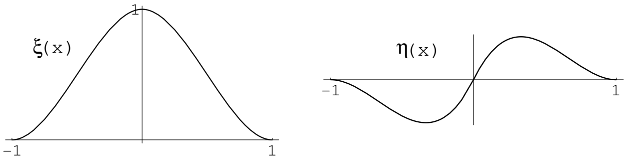


Figure 4.4. *Basis functions for cubic Hermite polynomials on* $[-1,1]$ {1

and

$$\eta_i(x) = \begin{cases} \eta(\frac{x}{h} - i) & x_{i-1} \le x \le x_{i+1} \\ 0 & \text{elsewhere} \end{cases} \tag{4.52}$$ {1

for $i = 1, \ldots, N$. So far we have $2N$ functions and we know that the dimension of $\mathcal{H}^h$ is $2N+4$ so we must define four additional functions. To this end, we define $\xi_0(x)$, $\xi_{N+1}(x)$ and $\eta_0(x)$, $\eta_{N+1}(x)$ by

{1d_fourth_xil}
$$\xi_0(x) = \begin{cases} \xi(\frac{x}{h}) & 0 \le x \le x_1 \\ 0 & \text{elsewhere} \end{cases} \qquad \eta_0(x) = \begin{cases} \eta(\frac{x}{h}) & 0 \le x \le x_1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.53}$$

{1d_fourth_etar}
$$\eta_{N+1}(x) = \begin{cases} \eta(\frac{x}{h} - (N+1)) & x_N \le x \le x_{N+1} \\ 0 & \text{elsewhere.} \end{cases} \tag{4.54}$$

and

{1d_fourth_xir}
$$\xi_{N+1}(x) = \begin{cases} \xi(\frac{x}{h} - (N+1)) & x_N \le x \le x_{N+1} \\ 0 & \text{elsewhere} \end{cases} \tag{4.55}$$

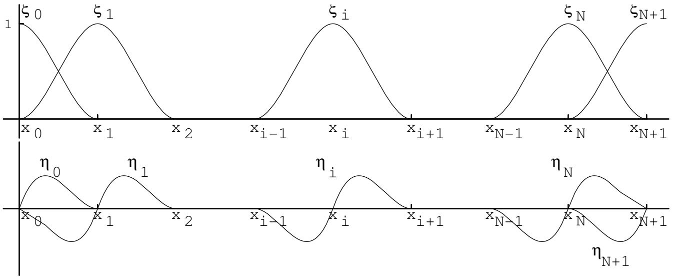
Summarizing, we have that

{1d_fourth_xi_eqn}
$$\xi_i(x_j) = \delta_{ij} \quad \text{and} \quad \xi_i'(x_j) = 0 \quad \text{for } 0 \le i,j \le N+1 \tag{4.56}$$

and

{1d_fourth_eta_eqn}
$$\eta_i(x_j) = 0 \quad \text{and} \quad h\eta_i'(x_j) = \delta_{ij} \quad \text{for } 0 \le i,j \le N+1\ . \tag{4.57}$$

These polynomials are illustrated in Figure 4.5.



{1d_fig_hercubic_basis}

Figure 4.5. *Basis functions for cubic Hermite polynomials*

Clearly these $2N+4$ functions $\{\xi_i\}_0^{N+1}$, $\{\eta_i\}_0^{N+1}$ belong to $\mathcal{H}^h$; moreover, they form a basis for $\mathcal{H}^h$. To see this, let $p(x) \in \mathcal{H}^h$ so that $p \in C^1(0,1)$, $p(x)$ is a cubic on each subinterval $[x_i, x_{i+1}]$, $0 \le i \le N$. Clearly $p(x)$ is uniquely determined by its value and that of its derivative at the $N+2$ nodes $x_0, \dots, x_{N+1}$. Using (4.56) and (4.57) we have

$$p(x) = \sum_{i=0}^{N+1} p(x_i)\xi_i(x) + h \sum_{i=0}^{N+1} p'(x_i)\eta_i(x) \,.$$

Thus the vectors span $\mathcal{H}^h$ and are also clearly linearly independent.

Of course, for our example, we must constrain the space to satisfy the homogeneous Dirichlet boundary conditions. To this end, we define $\widehat{\mathcal{H}}^h$ to be all functions $\phi(x) \in \mathcal{H}^h$ which satisfy $\phi(0) = \phi(1) = 0$. In this case we choose the $2n+2$ functions $\{\xi_i\}_1^N$, $\{\eta_i\}_0^{N+1}$; we do not include ξ_0, ξ_{N+1} since from (4.56) we know that $\xi_0(0) = 1$, $\xi_{N+1}(1) = 1$ so that $\xi_0, \xi_{N+1} \notin \widehat{\mathcal{H}}^h$.

We can now pose our weak problem over $\widehat{\mathcal{H}}^h \subset H^2(0,1) \cap H_0^1(0,1)$. We seek $u^h \in \widehat{\mathcal{H}}^h$ satisfying

$$A(u^h, v^h) = \int_0^1 \Big(r{u^h}'' {v^h}'' + p{u^h}'{v^h}' + q u^h v^h \Big)\, dx = \big(f, v^h\big) \quad \forall\, v^h \in \widehat{\mathcal{H}}^h \,. \qquad (4.58) \quad \{1$$

Once we have chosen a basis for our approximating space, we know that our discrete weak problem reduces to solving a linear system of algebraic equations $\mathcal{A}c = \mathcal{F}$. We let $\{\phi_i\}_{i=1}^{2N+2}$ be the basis functions $\{\xi_i(x), \eta_i(x)\}$ defined by (4.51)–(4.54) and ordered in the sequence $\{\eta_0, \xi_1, \eta_1, \xi_2, \eta_2, \dots, \xi_N, \eta_N, \eta_{N+1}\}$. If we write $u^h = \sum_{i=1}^{2N+2} c_j \phi_j(x)$ then the c_j's represent either the nodal values of u^h or of $h(u^h)'$. The matrix $\mathcal{A}$ whose entries are given by

$$\mathcal{A}_{ij} = \int_0^1 \Big(r(x)\phi_i''(x)\phi_j''(x) + p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x) \Big)\, dx$$

is a symmetric matrix. However, the matrix is no longer tridiagonal as in the case of piecewise linear elements but rather has the block tridiagonal form

$$\begin{pmatrix} A_0 & B_0 & 0 & & \cdots & 0 \\ B_0 & A_1 & B_1 & 0 & \cdots & 0 \\ 0 & B_1 & A_2 & B_2 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & B_{N-2} & A_{N-1} & B_{N-1} \\ 0 & \cdots & & 0 & B_{N-1} & A_N \end{pmatrix}$$

where the A_i's and B_i's are 2×2 matrices. To see this, consider the interval $[x_{i-1}, x_{i+1}]$. The basis functions which are nonzero on this interval are ξ_{i-1}, η_{i-1}, ξ_i, η_i, ξ_{i+1}, and η_{i+1} so that the maximum number of nonzero entries in a single row is six. It can be shown that the coefficient matrix is also positive definite so that the linear system can be efficiently solved using a block Cholesky factorization.

In order to obtain an error estimate we turn to Galerkin's theorem which provides us with the H^2-estimate

$$\big\| u - u^h \big\|_2 \le \inf_{\chi^h \in \widehat{\mathcal{H}}^h} \big\| u - \chi^h \big\|_2 \,, \qquad (4.59) \quad \{1$$

where u and u^h satisfy (4.49) and (4.58), repectively. To bound the term for the error in the best approximation, we consider the $\mathcal{H}^h$ interpolant. The Hermite cubic interpolant of a function $g(x)$ on the uniform partition of $[0,1]$

$$0 = x_0 < x_1 < \cdots < x_{N+1} = 1 \quad \text{where} \quad x_i = x_{i-1} + h_i, \quad 1 \leq i \leq N+1\,,$$

where $g(0) = g(1) = 0$ is given by

fourth_hercubic_definter}

$$I^h g = \sum_{i=1}^{N} g(x_i)\xi_i(x) + h \sum_{i=0}^{N+1} g'(x_i)\eta_i(x)\,. \tag{4.60}$$

From approximation theory we have the following result which tells us how well a function can be approximated by its cubic Hermite interpolant. As in the case of the piecewise linear interpolant, additional smoothness on the function must be assumed in order to get the optimal rates of approximation.

{1d_thm_hercubic_inter} **Lemma 4.12.** *Let $f \in H^s(0,1)$ where $2 \leq s \leq 4$ and let $I^h f$ denote its piecewise cubic Hermite interpolant defined by (4.60). Then for $0 \leq k \leq 2$ and for some constant C, we have that*

{1d_hercubic_inter}

$$\left\| f - I^h f \right\|_k \leq C h^{s-k} \left\| f \right\|_s \,. \tag{4.61}$$

Note that in order to get the optimal accuracy, e.g.,$O(h^4)$ in the L^2-norm, we must have that $f \in H^4(0,1)$.

We can now use Lemma 4.12 to bound the right-hand side of (4.59). We have the following error estimates; the proof is similar to that of Theorems 4.3 and 4.4.

{1d_thm_hercubic_error} **Theorem 4.13.** *Let $u \in H^k(0,1)$, $2 \leq k \leq 4$, be the solution of (4.49) and let $u^h \in \widehat{\mathcal{H}}^h$ be the solution of (4.58). Then*

1d_fourth_hercubic_error}

$$\left\| u - u^h \right\|_j \leq C h^{k-j} \left\| u \right\|_k \tag{4.62}$$

where $j = 0$ or $j = 2$.

In practice, cubic Hermite functions are not often used. The reason for this is twofold. First, there are, in general, $2N + 4$ parameters to compute in one dimension; as we will see in the next section there is an approximating space which maintains the same accuracy as cubic Hermites with only $N + 4$ parameters. Hence for cubic Hermite polynomials we would be solving a $(2N + 4)$-dimensional system in $\mathbb{R}^1$ versus a $(N+4)$-dimensional system; this difference in size of the linear system magnifies as we move to higher dimensions. Secondly, the Hermite cubic interpolant matches the function and its derivative at the nodes. In many situations, especially in more than one-dimension, it is not possible to accurately specify the derivatives at the nodes. The space considered in the next section requires interpolation of the function values only.

4.4.2 Piecewise cubic spline functions

{1d_subsec_splines}

In this section we consider a finite dimensional subspace of $H^2(0,1)$ which has two desirable properties. We will require that only function values (and not derivatives) will be used to interpolate smooth functions and that our space has as small a dimension as possible. To do this, we constrain the space of Hermite cubics to obtain the space

$$\mathcal{C}^h = \{\phi(x) \ : \ \phi(x) \in C^2[0,1], \ \phi(x)\text{is a cubic in each } [x_i, x_{i+1}]\} \,, \tag{4.63}$$

{1

where we are using the same uniform partition defined by $h = 1/(N+1)$ as before. In an analogous manner to the case of $\mathcal{H}^h$ of cubic Hermite polynomials, we can determine that $\mathcal{C}^h$ is a $(N+4)$-dimensional subspace of $H^2(0,1)$.

We must now specify a basis for $\mathcal{C}^h$. In our space $\mathcal{H}^h$ of cubic Hermite polynomials, there was a clear criterion for determining its elements. In fact, to determine $\phi \in \mathcal{H}^h$ we just specified $\phi(x_i)$ and $\phi_i'(x_i)$ at the nodes x_i, $0 \le i \le N+1$, thus defining a unique cubic polynomial on each interval and at the same time assuring that it be in $C^1[0,1]$. For the cubic spline space $\mathcal{C}^h$, the obvious thing to try in order to assure C^2-continuity at the nodes would be to specify $\phi(x)$, $\phi'(x)$, and $\phi''(x)$ there. However, this cannot be done using cubic polynomials because we would be overspecifying them.

Instead of specifying basis functions which interpolate a function and its first and second derivatives at the nodes, we take the approach of constructing a basis for $\mathcal{C}^h$. To find the ith basis function we first note that its support cannot be in the interval $[x_{i-1}, x_{i+1}]$ as was the case with piecewise linear functions and Hermite cubic functions. To see this, we note that there are eight degrees of freedom to determine the cubic polynomials on $[x_{i-1}, x_{i+1}]$ and a total of nine conditions to specify over the three nodes, i.e.,$\phi(x_{i\pm1}) = \phi'(x_{i\pm1}) = \phi''(x_{i\pm1}) = 0$ and the continuity of $\phi(x)$, $\phi'(x)$, and $\phi''(x)$ at $x = x_i$. Consequently, we must extend our interval to $[x_{i-2}, x_{i+2}]$ and attempt to construct a C^2 function which is cubic on each of the four subintervals $[x_{i-s}, x_{i-s+1}]$, for $s = -1, 0, 1, 2$ and which is zero outside the interval $[x_{i-2}, x_{i+2}]$. In this case we have 16 degrees of freedom and 15 conditions to impose so that it is clearly possible; the extra degree of freedom will be used to specify that the function is one at node x_i. A straightforward, but tedious, computation gives such a function on the interval $[-2,2]$; this function is illustrated in Figure 4.6. Translating this function to the interval $[x_{i-2}, x_{i+2}]$ for $2 \le i \le N-1$ we have

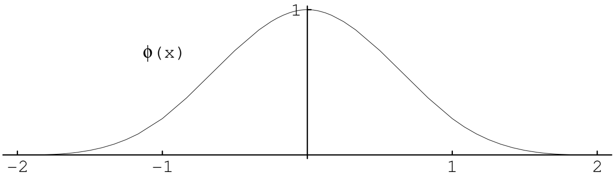
$$\phi_i(x) = \left\{ \begin{array}{ll} \phi(\frac{x}{h} - i) & x_{i-2} \le x \le x_{i+2} \\ 0 & \text{elsewhere} \end{array} \right. \tag{4.64}$$

{F

where

$$\phi(x) = \begin{cases} \frac{1}{4}\,(x+2)^3 & -2 \le x \le -1\,, \\ \frac{1}{4}\left(1+3(1+x)+3(1+x)^2-3(1+x)^3\right) & -1 \le x \le 0\,, \\ \frac{1}{4}\left(1+3(1-x)+3(1-x)^2-3(1-x)^3\right) & 0 \le x \le 1\,, \\ \frac{1}{4}(2-x)^3 & 1 \le x \le 2\,. \end{cases} \tag{4.65}$$ {1

Clearly each $\phi_i(x) \in \mathcal{C}^h$ for $2 \le i \le N-1$; we have a total of $N-2$ functions so



{1d_fig_basis_spline1}

Figure 4.6. *Basis function on* $[-2,2]$ *for cubic splines*

that an additional six basis functions are needed to reach the dimension $N+4$. We add the functions defined below where we have introduced extra nodes $x_{-1} = -h$ and $x_{N+2} = 1+h$:

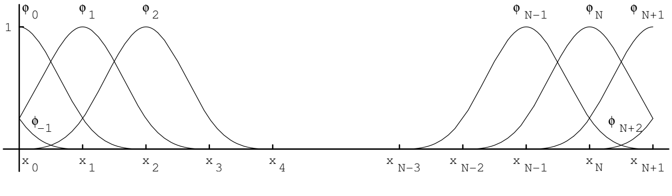
$$\phi_{-1}(x) = \begin{cases} \phi(\frac{x}{h}+1) & 0 \le x \le x_1 \\ 0 & \text{otherwise} \end{cases} \qquad \phi_0(x) = \begin{cases} \phi(\frac{x}{h}+1) & 0 \le x \le x_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_1(x) = \begin{cases} \phi(\frac{x}{h}-1) & 0 \le x \le x_3 \\ 0 & \text{otherwise} \end{cases} \qquad \phi_N(x) = \begin{cases} \phi(\frac{x}{h}-N) & x_{N-2} \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

_fourth_spline_basis_end}

$$\phi_{N+1}(x) = \begin{cases} \phi(\frac{x}{h}-N-1) & x_{N-1} \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.66}$$

$$\phi_{N+2}(x) = \begin{cases} \phi(\frac{x}{h}-N-2) & x_N \le x \le 1 \\ 0 & \text{otherwise}\,. \end{cases}$$



{1d_fig_basis_spline2}

Figure 4.7. *Basis functions for cubic splines*

The set $\{\phi_i(x)\}_{i=-1}^{N+2}$ form a basis for $\mathcal{C}^h$. These basis functions are illustrated in Figure 4.7.

An important difference in the basis functions for cubic splines and those we studied for piecewise linear functions and cubic Hermites is that the unknowns we solve for are no longer the nodal values of u^h. This is because the cubic spline basis functions are no longer zero at all nodes except one.

Since for our problem the underlying Hilbert space is $H^2(0,1) \cap H_0^1(0,1)$, i.e.,all functions in $H^2(0,1)$ which are zero at $x=0$ and at $x=1$, we need only $N+2$ basis functions. We define $\widehat{\mathcal{C}}^h$ as

$$\widehat{\mathcal{C}}^h = \{\phi \in \mathcal{C}^h \ : \ \phi(0) = \phi(1) = 0\}\,.$$

For cubic splines we can not simply omit the specific basis functions which are nonzero at $x=0$ and $x=1$ as we did for the piecewise linear basis. We can equip $\widehat{\mathcal{C}}^h$ with a basis consisting of $\{\phi_i\}_{i=2}^{N-1}$ and the four functions $\tilde{\phi}_0$, $\tilde{\phi}_1$, $\tilde{\phi}_N$, and $\tilde{\phi}_{N+1}$ which are obtained as linear combinations of the remaining $\phi_i(x)$'s so that they vanish at $x=0$ and at $x=1$. For example, $\tilde{\phi}_0 = \phi_0 - 4\phi_{-1}$, $\tilde{\phi}_1 = \phi_1 - \phi_{-1}$. (See exercises.)

As before, in order to obtain error estimates using cubic splines as our approximating space, we need to obtain estimates for the error in the cubic spline interpolant. We note that we have $N+2$ nodes and the space $\mathcal{C}^h$ has dimension $N+4$; thus if the interpolant matches the function value at the $N+2$ nodes, then we have two additional conditions to impose. There are numerous choices; here we consider one type of cubic spline interpolant of a function $f \in H^2(0,1)$. We require the interpolant, denoted $I^h f$, to satisfy the $N+4$ conditions

$$\begin{array}{rcl} I^h f(x_i) & = & f(x_i) \quad \text{for } 0 \le i \le N+1 \\ I^h f'(x_0) & = & f'(x_0)\,, \qquad I^h f'(x_{N+1}) = f'(x_{N+1})\,. \end{array} \tag{4.67}$$ {1

Lemma 4.14. *Let $f \in H^s(0,1)$, $2 \le s \le 4$. Then for $0 \le r \le 2$ we have that* {1

$$\left\| D^r(f - I^h f) \right\|_0 \le C h^{s-r} \left\| D^s f \right\| \,. \tag{4.68}$$ {1

We now state a result analogous to Theorem 4.13 .

Theorem 4.15. *Let $u \in H^k(0,1)$, $2 \le k \le 4$ be the solution of (4.49) and let $u^h \in \widehat{\mathcal{C}}^h$ be the solution of (4.58). Then* {1

$$\left\| u - u^h \right\|_j \le C h^{k-j} \left\| u \right\|_k \tag{4.69}$$ {1

where $j = 0$ or $j = 2$.

Chapter 5

Simple Examples on Rectangular Domains

{c

In this chapter we consider simple elliptic boundary value problems in rectangular domains in $\mathbb{R}^2$ or $\mathbb{R}^3$; our prototype example is the Poisson equation but we also briefly consider the biharmonic equation and the Helmholtz equation. Similar to our exposition of the two-point boundary value problem in Chapter 4, we consider the implementation of different boundary conditions for our prototype equation. Much of this chapter is a straightforward extension of the analysis presented in the previous chapter for the two-point boundary value problem. However, a few important differences are evident.

For the finite element approximation of these elliptic boundary value problems, we only consider approximating with finite elements spaces which are obtained by taking tensor products of one-dimensional finite element spaces. In Chapter 6 we consider the general problem of determining finite element spaces on polygonal domains and in a later chapter we consider isoparametric finite elements for curved domains.

5.1 The Poisson equation with homogeneous Dirichlet boundary data

{2

In this section we consider Poisson's equation defined in a bounded domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ with homogeneous Dirichlet boundary data. We let $\vec{x}$ denote a point in $\mathbb{R}^2$ or $\mathbb{R}^3$. Specifically, we let Ω be an open, connected, bounded set in $\mathbb{R}^2$ or $\mathbb{R}^3$ and let $\partial\Omega$ denotes its boundary. At this point in our discussion of the finite element method, we only have the background to use finite element spaces which are tensor products of the one dimensional finite element spaces discussed in the last chapter. Consequently, when we move to the discretization stage we require that Ω be a rectangular domain. However, the weak formulations that we present hold for more general domains. In the next chapters we address the problem of discretizing using other elements suitable for more general domains. We let $\overline{\Omega}$ denote the closure of Ω; i.e.,$\overline{\Omega} = \Omega \cup \partial\Omega$. Let $f = f(\vec{x})$ be a given function that is continuous on the closure of Ω; i.e.,$f \in C(\overline{\Omega})$. We say that a function $u(\vec{x})$ defined on $\overline{\Omega}$ is a classical solution of the Poisson equation with homogeneous Dirichlet boundary conditions if $u \in C^2(\Omega)$, $u \in C(\overline{\Omega})$ and u satisfies

{2d_poisson}
$$-\Delta u(\vec{x}) = f(\vec{x}) \quad \text{for } \vec{x} \in \Omega \tag{5.1a}$$

{2d_dirbc}
$$u(\vec{x}) = 0 \quad \text{for } \vec{x} \in \partial\Omega\,, \tag{5.1b}$$

where $\Delta u = u_{xx} + u_{yy}$ in $\mathbb{R}^2$ or analogously $\Delta u = u_{xx} + u_{yy} + u_{zz}$ in $\mathbb{R}^3$. It is well known that for sufficiently smooth $\partial\Omega$ there exists a unique classical solution of (5.1).

In the sequel, we assume enough smoothness of the boundary so that the domain admits the application of the divergence theorem. Every polygonal domain or a domain with a piecewise smooth boundary has sufficient smoothness for our purposes.

We make extensive use of Green's formula which is the analog of the integration by parts formula in higher dimensions and is derived from the divergence theorem of vector calculus. Let $\vec{n}$ denote the unit outer normal to $\partial\Omega$ and let dS denote the measure defined on the boundary and dV the measure of volume. We have that for $v \in C^1(\overline{\Omega})$, $w \in C^2(\overline{\Omega})$

$$\int_\Omega v \Delta w \, dV = \int_{\partial\Omega} v(\vec{n} \cdot \nabla w)\, dS - \int_\Omega \nabla w \cdot \nabla v \, dV$$

or equivalently

{2d_green}
$$\int_\Omega v \Delta w \, dV = \int_{\partial\Omega} v \frac{\partial w}{\partial \vec{n}} \, dS - \int_\Omega \nabla w \cdot \nabla v \, dV\,. \tag{5.2}$$

5.1.1 Weak formulation

To define the weak formulation we first determine the underly As before, we impose the homogeneous Dirichlet boundary conditions by constraining our space $H^1(\Omega)$; in particular we have the space

$$H^1_0(\Omega) = \{ v \in H^1(\Omega) \ : \ v = 0 \text{ on } \partial\Omega \}\,.$$

The weak formulation which we consider is

{2d_homodir_weak}
$$\left\{ \begin{array}{l} \text{seek } u \in H^1_0(\Omega) \text{ such that} \\ A(u,v) = \displaystyle\int_\Omega \nabla v \cdot \nabla w \, dV = (f,v) \quad \forall\, v \in H^1_0(\Omega)\,. \end{array} \right. \tag{5.3}$$

The solution $u \in H^1_0(\Omega)$ of (5.3) is called the generalized or weak solution of (5.1).

If u satisfies the classical problem (5.1) then u satisfies the weak formulation (5.1) because

$$\begin{aligned} \int_\Omega f v \, dV &= -\int_\Omega \Delta u v \, dV \quad \forall\, v \in H^1_0(\Omega) \\ &= \int_\Omega \nabla u \cdot \nabla v \, dV - \int_{\partial\Omega} \frac{\partial u}{\partial \vec{n}} v \, dS \end{aligned}$$

$$\begin{aligned} &= \int_\Omega \nabla u \cdot \nabla v \, dV \\ &= A(u,v)\,, \end{aligned}$$

where we have used Green's formula (5.2) and imposed the fact that $v = 0$ on $\partial\Omega$.

The existence and uniqueness of a weak solution to (5.3) can be verified by satisfying the hypotheses of the Lax-Milgram theorem. Recall that the norm on $H^1(\Omega)$ is defined by

$$\|u\|_1^2 = \int_\Omega \big(u^2 + \nabla u \cdot \nabla u\big)\, dV = \|u\|_0^2 + \|\nabla u\|_0^2 = \|u\|_0^2 + |u|_1^2\,.$$

The bilinear form is bounded on all of $H^1(\Omega)$ since

$$\begin{aligned} |A(u,v)| = \Big| \int_\Omega \nabla u \cdot \nabla v \, dV \Big| &= |\,(\nabla u, \nabla v)\,| \\ &\le \|\nabla u\|_0 \|\nabla v\|_0 \le \|u\|_1 \, \|v\|_1 \ , \end{aligned}$$

where we used the Cauchy-Schwarz inequality and the definition of the H^1 and L^2 norms.

We must now show coercivity of the bilinear form, i.e.,there exists a constant $m > 0$ such that

$$A(u,u) = \int_\Omega (\nabla u \cdot \nabla u)\, dV \ge m \, \|u\|_1^2 \quad \forall \, u \in H_0^1(\Omega)\,.$$

Note that the bilinear form $A(u,u)$ can also be written as

$$A(u,u) = |u|_1^2 = \frac{1}{2}\big(|u|_1^2 + |u|_1^2\big)\,.$$

Our underlying Hilbert space is $H_0^1(\Omega)$ so we can use the Poincaré inequality to demonstrate that the standard H^1-norm is norm equivalent to this semi-norm and thus coercivity is guaranteed in an analogous manner to the homogeneous Dirichlet problem for the two-point boundary value problem of the last chapter. Specifically, we have

$$A(u,u) = \frac{1}{2}\big(|u|_1^2 + |u|_1^2\big) \ge \frac{1}{2} \min\{1, \frac{1}{C_p^2}\}\Big(\|u\|_0^2 + |u|_1^2\Big) = m \, \|u\|_1^2 \ ,$$

where C_p is the constant in the Poincaré inequality. We have demonstrated the boundedness and coercivity of the bilinear form defined in (5.3) and thus the Lax-Milgram theorem guarantees the existence and uniqueness of a solution to the weak problem (5.3) because the right-hand side is obviously a bounded linear functional on $H^1(\Omega)$. The bilinear form is symmetric and so we know that approximating the solution of the weak problem is equivalent to the minimization problem

$$\min_{v \in H_0^1(\Omega)} \int_\Omega \Big(\frac{1}{2}\nabla v \cdot \nabla v - fv\Big)\, dV\,.$$

5.1.2 Approximation using bilinear functions

{2

In later chapters we consider finite element spaces over polygonal or curved domains. At present, we restrict the domain so that we can use rectangular elements; therefore, the finite element spaces can be constructed from the spaces used in the previous chapter. As in the one-dimensional case, we must now choose a finite dimensional subspace of $S_0^h(\Omega) \subset H_0^1(\Omega)$ in which to seek the approximate solution. For the discrete problem we have

{2d_homodir_dweak}

$$\begin{cases} \text{seek } u^h \in S_0^h(\Omega) \text{ satisfying} \\ A(u^h, v^h) = \int_\Omega \left(\nabla u^h \cdot \nabla v^h\right) dV = \left(f, v^h\right) \quad \forall\, v^h \in S_0^h\,. \end{cases} \tag{5.4}$$

Existence and uniqueness of the solution to this problem is guaranteed by the Lax-Milgram theorem.

To approximate our finite element solution we consider the concrete case where Ω is the unit square or unit cube. We choose the space $S_0^h(\Omega)$ to be continuous, piecewise bilinear functions defined on $\Omega \subset \mathbb{R}^2$ or continuous, piecewise trilinear functions[9] for $\Omega \subset \mathbb{R}^3$. We formally construct the bilinear basis functions; the trilinear basis functions are defined analogously. Let N, M be positive integers and let $h_x = 1/(N+1)$, $h_y = 1/(M+1)$ and consider the subdivision of Ω into rectangles of size $h_x \times h_y$ where

$$x_i = ih_x, \;\; 0 \le i \le N+1, \quad y_j = jh_y, \;\; 0 \le j \le M+1\,.$$

See Figure 5.1 for a sample grid on a unit square with $h_y = 2h_x$. Let $\phi_i(x)$, $1 \le i \le N$ represent the standard "hat" piecewise linear basis functions in x and let $\phi_j(y)$, $1 \le j \le M$, be similarly defined, i.e.,

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_x} & \text{for } x_{i-1} \le x \le x_i \\ \dfrac{x_{i+1} - x}{h_x} & \text{for } x_i \le x \le x_{i+1} \\ 0 & \text{elsewhere} \end{cases} \quad \phi_j(y) = \begin{cases} \dfrac{y - y_{j-1}}{h_y} & \text{for } y_{j-1} \le y \le y_j \\ \dfrac{y_{j+1} - y}{h_y} & \text{for } y_j \le y \le y_{j+1} \\ 0 & \text{elsewhere.} \end{cases}$$

On $\Omega = (0,1) \times (0,1)$ we now define the NM *bilinear functions*

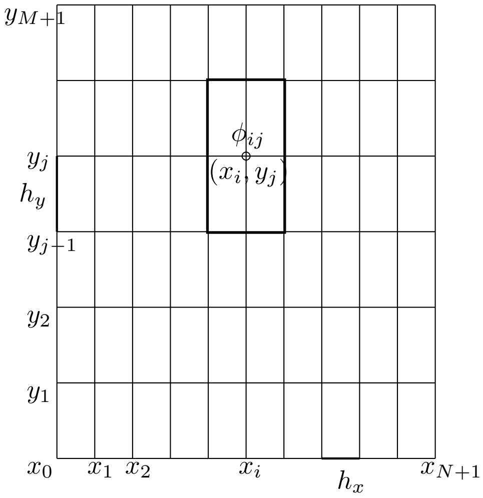
{2d_homodir_basis}

$$\phi_{ij}(x,y) = \phi_i(x)\phi_j(y) \quad \text{for } 1 \le i \le N,\, 1 \le j \le M\,. \tag{5.5}$$

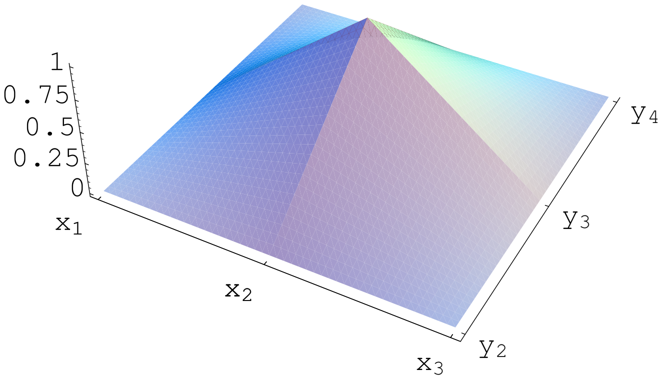
We easily see that $\phi_{ij}(x_i, y_j) = 1$ and $\phi_{ij}(x_k, y_l) = 0$ for $k \neq i$ **or** $l \neq j$. Also $\phi_{ij}(x,y)$ is zero outside of $[(i-1)h_x, (i+1)h_x] \times [(j-1)h_y, (j+1)h_y]$. The support of $\phi_j(x,y)$ is illustrated in Figure 5.1 and the shape of a specific bilinear function $\phi_{2,3}$ which is one at node (x_2, y_3) is given in Figure 5.2.

For Ω the unit square, we choose $S_0^h(\Omega) \equiv S_0^h(0,1) \otimes S_0^h(0,1)$ to be the tensor product of the subspaces $S_0^h(0,1)$ (one each in the $x-$ and $y-$ directions) of one-dimensional piecewise linear, continuous functions which vanish at zero and one.

[9] A bilinear or trilinear function is a function which is linear with respect to its variables because if we hold one variable fixed, it is linear in the other; for example $f(x,y) = xy$ is a bilinear function but $f(x,y) = x^2y$ is not.



{2d_fig_grid} **Figure 5.1.** *Grid on a unit square with support of basis function $\phi_{ij}(x, y)$ indicated.*



{2d_fig_basis_linear} **Figure 5.2.** *Support of bilinear basis function $\phi_{2,3}$.*

$S_0^h(\Omega)$ consists of all functions $v(x, y)$ on $(0, 1) \times (0, 1)$ of the form

{2dfunc}
$$v(x, y) = \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} \phi_i(x) \phi_j(y) = \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} \phi_{ij}(x, y)\,. \tag{5.6}$$

Note that the general form of a bilinear function in $\mathbb{R}^2$ is $a_0 + a_1x + a_2y + a_3xy$ compared with a linear function in two variables which has the general form $a_0 + a_1x + a_2y$. Clearly $S_0^h(\Omega)$ is the space of all continuous, piecewise bilinear functions (with respect to the given subdivision) which vanish on the sides of the unit square. Also, every piecewise bilinear function $w(x, y)$ can be written in the form (5.6) with $c_{ij} = f(x_i, y_j)$; i.e.,it is a linear combination of the $P = NM$ linearly independent functions $\phi_{ij}(x, y)$. $S_0^h(\Omega)$ is an P-dimensional subspace of $H_0^1(\Omega)$; note that for $M = N$, S_0^h is an N^2 dimensional subspace whereas in one dimension, it was an N dimensional subspace. Of course this affects the size of our resulting matrix problem.

From previous discussions we know that once a basis is chosen for the approximating subspace, the discrete problem can be written as a linear system of equations. To investigate the structure of the coefficient matrix for our choice of bilinear basis functions, we let the basis functions $\phi_{ij}(x, y)$ for $S_0^h(\Omega)$ be rewritten in single index notation; for simplicity of exposition we choose $M = N$. We have

$$\{\psi_k(x, y)\}_{k=1}^{N^2} = \{\phi_{ij}(x, y)\}_{i,j=1}^{N} .$$

For example, $\psi_k = \phi_{k1}$ for $1 \le k \le N$, $\psi_{N+k} = \phi_{k2}$ for $1 \le k \le N$, etc. Our discrete weak formulation (5.4) is equivalent to seeking $u^h \in S_0^h$ satisfying

$$A(u^h, \psi_i) = (f, \psi_i) \quad \text{for } 1 \le i \le N^2 .$$

We now let $u^h = \sum_{j=1}^{N^2} c_j\psi_j$ and substitute into the above expression. The result is a linear system of N^2 equations in the N^2 unknowns $\{c_j\}_{j=1}^{N^2}$; i.e.,$\mathcal{A}\vec{c} = \vec{\mathcal{F}}$ where $\vec{c} = (c_1, \ldots, c_{N^2})^T$, $\mathcal{F}_i = (f, \psi_i)$ and $\mathcal{A}_{ij} = A(\psi_i, \psi_j)$. Note that with the numbering scheme we are using for the basis functions, we are numbering our unknowns which correspond to the coefficients c_j across rows. Because we have assumed the same number of points in the x and y directions we could have easily numbered them along columns of the grid.

To determine the structure of the resulting matrix we consider the ith row of the matrix and decide how many nonzero entries are in the row. Because we know the matrix is symmetric, we only consider terms above the diagonal. Clearly there can be nonzero entries in columns i and $i + 1$. The next nonzero entries occur for unknowns corresponding to basis functions in the next row of the grid. Specifically we can have nonzero entries in columns $i + N - 1$, $i + N$ and $i + N + 1$ where N is the number of unknowns across the row. The coefficient matrix $\mathcal{A}$ is an $N^2 \times N^2$ symmetric, positive definite matrix which has a block tridiagonal structure of the form

{2dPDmatrix}

$$\mathcal{A} = \begin{pmatrix} A_0 & A_1 & 0 & & \cdots & 0 \\ A_1 & A_0 & A_1 & 0 & \cdots & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & A_1 & A_0 & A_1 \\ 0 & \cdots & & 0 & A_1 & A_0 \end{pmatrix} , \tag{5.7}$$

where A_0 and A_1 are $N \times N$ tridiagonal matrices. (See exercises.) A matrix of this form can be solved efficiently by a banded Cholesky algorithm, a block tridiagonal solver or an iterative solver.

Error estimates

{2

Galerkin's theorem provides us with the estimate

$$\left\|u-u^h\right\|_1 \le \inf_{\chi^h\in S_0^h}\left\|u-\chi^h\right\|_1 \,. \tag{5.8}$$ {2

As before, we turn to the interpolant of u in our finite dimensional space $S_0^h(\Omega)$ to obtain an estimate in terms of powers of h. Specifically for $\mathbb{R}^2$, we denote $I^h v$ as the unique function in $S_0^h(\Omega)$ which satisifes $(I^h v)(x_i,y_j) = v(x_i,y_j)$ for $0\le i,j\le N+1$. We can write $I^h v$ as a linear combination of our basis functions; i.e.,$(I^h v)(x,y) = \sum_{i,j=1}^N v(x_i,y_j)\phi_{ij}(x,y)$. For v defined on $\Omega\subset\mathbb{R}^2$, we let $I_x^h v$ and $I_y^h v$ denote the interpolation operators in the x- and y-directions; i.e.,

$$(I_x^h v)(x,y) = \sum_{i=1}^N v(x_i,y)\phi_i(x) \quad\text{and}\quad (I_y^h v)(x,y) = \sum_{j=1}^N v(x,y_j)\phi_j(y)\,.$$

Then we have that

$$\begin{aligned}(I_y^h I_x^h v)(x,y) &= I_y^h\left(\sum_{i=1}^N v(x_i,y)\phi_i(x)\right) = \sum_{j=1}^N\left(\sum_{i=1}^N v(x_i,y_j)\phi_i(x)\right)\phi_j(y)\\ &= (I^h v)(x,y)\,.\end{aligned}$$

Similarly, $I^h v = I_x^h I_y^h v$. For $\Omega\subset\mathbb{R}^3$ clearly $I^h v = I_x^h I_y^h v I_z^h$. This result can be used to prove the following theorem which gives us an estimate of the error in $v - I^h v$ when v is sufficiently smooth.

Lemma 5.1. *Let $v\in H^2(\Omega)$. Then if $I^h v$ is the interpolant of v in $S^h(\Omega)$, the space of continuous, piecewise bilinear functions, then there exist constants C_i, $i=1,2$ independent of v and h such that*

$$\left\|v-I^h v\right\|_0 \le C_1 h^2\left\|v\right\|_2 \tag{5.9}$$ {2

and

$$\left\|v-I^h v\right\|_1 \le C_2 h\left\|v\right\|_2\,. \tag{5.10}$$ {2

As in the case in one-dimension, we can now make use of the interpolation result to prove an optimal error estimate in the H^1-norm. To obtain a result for the L^2-norm we again use "Nitsche's trick" in a manner completely analogous to that in the one-dimensional case where now we make use of elliptic regularity. The details of the proof are left to the exercises.

Theorem 5.2. *Let $u \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution of (5.3) where $\Omega = (0,1) \times (0,1)$. Let $S_0^h(\Omega)$ be the space of piecewise bilinear functions which vanish on $\partial\Omega$ and let u^h be the Galerkin approximation to u in $S_0^h(\Omega)$ defined by (5.4). Then* {2

$$\left\| u - u^h \right\|_1 \le Ch \left\| u \right\|_2 \tag{5.11}$$ {2

and

{2dPDerrl2}
$$\left\| u - u^h \right\|_0 \le Ch^2 \left\| u \right\|_2 \tag{5.12}$$

for some constants C independent of h and u.

5.1.3 Higher order elements

Our discussion of approximating the problem (5.1) posed on $\Omega = (0,1) \times (0,1)$ has so far included only piecewise bilinear function spaces. Of course, we can also use tensor products of higher order spaces such as the quadratic or cubic functions in one space dimension. Note that a general biquadratic function has the form $a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2 + a_6 x^2 y + a_7 xy^2 + a_8 x^2 y^2$ compared with a general quadratic function in two dimensions which has the form $a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2$. As in the one-dimensional case, for a smooth enough solution, these spaces yield higher rates of convergence then that achieved with piecewise bilinear approximations. The construction of the basis functions in two or three dimensions is done analogous to the piecewise bilinear case; the details are left to the exercises.

5.1.4 Numerical quadrature

Once again, the entries in the matrix and right-hand side of our linear system are calculated using a numerical quadrature rule which has the form

$$\int_\Omega f(\vec{x})\, d\Omega \approx \sum_i f(\vec{q}_i)\omega_i \,,$$

where the points $\vec{q}_i$ are the quadrature points and ω_i are the quadrature weights. Because we are using rectangular elements with basis functions obtained by taking the tensor product of one-dimensional basis functions, the most straightforward approach is to use tensor products of the quadrature rules in one spatial dimension. Typically, we use the same quadrature rule in each spatial dimension. For example, if we have the rule

$$\int_a^b f(x)\; dx = \sum_i f(q_i) w_i$$

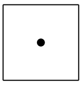
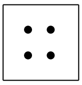
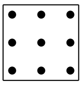
then we can write

$$\int_a^b \int_c^d f(x,y)\; dydx \approx \int_a^b \Big(\sum_j f(x, q_j) w_j \Big) \approx \sum_i \sum_j f(q_i, q_j) w_j w_i \,.$$

In one dimension we employed the Gauss-Legendre quadrature rules on $[-1,1]$. If we take the tensor products of a p-point Gauss rule in each direction then we would

{2d_table_quad}

Table 5.1. *Tensor product of Gauss quadrature rules in two dimensions*

	1-D rule	# points in $\mathbb{R}^2$	points q_i & weights w_i
	1 point Gauss	1	$q_1 = (0,0) \quad w_1 = 4$
	2 point Gauss	4	$q_i = \frac{1}{\sqrt{3}}\{(-1,-1),(1,-1),(-1,1$ $w_i = 1$
	3 point Gauss	9	$q_i = \sqrt{\frac{3}{5}}\Big\{(-1,-1),(0,-1),(1,-1$ $((0,0),(1,0),(-1,1),(0,1),(1$ $w_i = \frac{1}{81}\{25,40,25,40,64,40,25,$

have one point for the tensor product of the one-point rule, four points for the tensor product of the two-point rule, etc. The quadrature points in two dimensions formed by the tensor product of one-point through three-point Gauss quadrature rules are described in Table 5.1. Note that in three dimensions we have 1, 8, and 27 quadrature points for tensor products of these three quadrature rules. To apply these rules to an integral over an arbitrary rectangular domain, we must perform a change of variables in both the x and y directions analogous to the one-dimensional case. For our example, if we are using bilinear or trilinear elements, then the tensor product of the one-point Gauss rule is adequate; for biquadratics or triquadratics we need to use the tensor product of the two-point Gauss rule.

5.2 The Poisson equation with Neumann boundary data

In this section we consider solving Poisson's equation on an open, bounded domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ where we specify Neumann data on a portion of the boundary and Dirichlet data on the remainder of the boundary. In particular, we seek a function u satisfying

$$\begin{aligned} -\Delta u(\vec{x}) &= f(\vec{x}) \quad \text{for } \vec{x} \in \Omega \\ u(\vec{x}) &= 0 \quad \text{for } \vec{x} \in \Gamma_1 \\ \frac{\partial u}{\partial \vec{n}}(\vec{x}) &= g(\vec{x}) \quad \text{for } \vec{x} \in \Gamma_2\,, \end{aligned} \tag{5.13}$$

{2

where $\Gamma_1 \cup \Gamma_2 = \partial\Omega$, $\Gamma_1 \cap \Gamma_2$ is a set of measure zero, and $\partial u/\partial \vec{n}$ denotes the directional derivative of u in the direction of the unit outward normal to the boundary of the domain. We note that if $\Gamma_1 = \partial\Omega$ then we have the purely Dirichlet problem discussed in Section 5.1; in the case $\Gamma_2 = \partial\Omega$ we have a purely Neumann problem. As expected, in the latter case the problem does not have a unique solution. It is well known that for sufficiently smooth $\partial\Omega$ there exists a unique classical solution of (5.13) provided, of course, that Γ_1 is measurable.

5.2.1 Weak Formulation

For this problem we define $H^1_B(\Omega)$ as

{2dh1b}
$$H^1_B(\Omega) = \{u \in H^1(\Omega) \ : \ u = 0 \text{ on } \Gamma_1\}\,. \tag{5.14}$$

Our weak formulation is

{2d_PDNw}
$$\begin{cases} \text{seek } u \in H^1_B(\Omega) \text{ satisfying} \\ A(u,v) \equiv \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega = (f,v) + \int_{\Gamma_2} gv \quad \forall\, v \in H^1_B(\Omega)\,. \end{cases} \tag{5.15}$$

If u is a solution of the classical problem (5.13) then by Green's theorem u satisfies

$$\begin{aligned} (f,v) = -\int_\Omega \Delta uv \, d\Omega &= \int_\Omega \nabla u \cdot \nabla v \, d\Omega - \int_\Gamma \frac{\partial u}{\partial \vec{n}} v \, ds \\ &= A(u,v) - \int_{\Gamma_1} \frac{\partial u}{\partial \vec{n}} v \, ds - \int_{\Gamma_2} \frac{\partial u}{\partial \vec{n}} v \, ds \\ &= A(u,v) - \int_{\Gamma_2} g(\vec{x}) v \, ds \quad \forall\, v \in H^1_B(\Omega)\,, \end{aligned}$$

where we have used the fact that the boundary integral over Γ_1 is zero since $v \in H^1_B(\Omega)$ and for the boundary integral over Γ_2 we have used $\partial u/\partial \vec{n} = g(\vec{x})$. In this problem the Dirichlet boundary condition on Γ_1 is essential whereas the Neumann boundary condition on Γ_2 is natural. It's interesting to compare the weak formulation (5.15) with the analogous weak formulation (**??**) for the two-point boundary value problem. In the one-dimensional case, we simply have the value of the derivative at a point times the test function at the same point. In two spatial dimensions with inhomogeneous Neumann boundary conditions we have a line integral on the right-hand side of the weak form and in three spatial dimensions we have a surface integral. This complicates the implementation of the method but it is straightforward; for example, for $\Omega \subset \mathbb{R}^2$ we have a line integral on the boundary which can be approximated using a Gauss quadrature rule. The existence and uniqueness of a solution to (5.15) is demonstrated in an analogous manner to the purely Dirichlet problem discussed in Section 5.1. The only complication is demonstrating that the right-hand side, which now contains a boundary integral, is a bounded linear functional on $H^1(\Omega)$.

When the classical problem is a purely Neumann problem, i.e.,when $\Gamma_2 = \partial\Omega$, it is clear that there is not a unique solution. Thus, we can not expect the hypotheses of the Lax-Milgram theorem to be satisfied. In particular, we are unable to demonstrate coercivity of the bilinear form.

5.2.2 Approximation using bilinear functions

As a concrete example we once again take $\Omega = (0,1) \times (0,1)$; we choose Γ_1 to be the top and bottom portions of the boundary, i.e., when $y = 0$ and $y = 1$; Γ_2 is the remainder of the boundary. We subdivide our domain into rectangles of size $h \times h$ where $h = 1/(N+1)$, $x_i = ih$, $y_j = jh$, $i, j = 0, \ldots, N+1$. If we approximate using continuous, piecewise bilinear functions as in Section 5.1, then we seek our solution in the space $\hat{S}^h(\Omega$ which is the space of all continuous, piecewise bilinear functions on Ω which are zero at $y = 0$ and $y = 1$. In the x-direction we have the $N+2$ basis functions $\phi_i(x)$, $i = 0, 1, \ldots, N+1$ and N basis functions in the y-direction $\phi_j(y)$, $j = 1, \ldots, N$. In this case we have the $N(N+2)$ basis functions $\phi_{ij}(x,y)$ which are the tensor products of the one-dimensional basis functions. The basic structure of the matrix is the same as in the previous example. Optimal error estimates are derived in a completely analogous manner to the previous section when $u \in H^2(\Omega) \cap H^1_B(\Omega)$.

We note that if we attempt to discretize the purely Neumann problem, i.e.,when $\Gamma_2 = \partial\Omega$, then the resulting $(N+2)^2$ matrix would be singular. This is to be expected because we could not prove uniqueness of the solution to the weak problem. A unique solution to the system can be found by imposing an additional condition on u^h such as specifying u^h at *one* point or requiring the solution to have zero mean, i.e., $\int_\Omega u\, dV = 0$.

5.3 Other examples

In this section we make a few brief remarks concerning some additional examples. In particular, we consider Poisson's equation with inhomogeneous Dirichlet boundary data and with a mixed boundary condition, a purely Neumann problem for the Helmholtz equation and a fourth order equation.

5.3.1 Other boundary conditions

As in the one-dimensional case, we can consider problems with inhomogeneous Dirichlet boundary conditions such as

$$\begin{aligned} -\Delta u &= f \quad \vec{x} \in \Omega \\ u(\vec{x}) &= q(\vec{x}) \quad \text{on } \Gamma\,. \end{aligned} \tag{5.16}$$

{2

To treat the inhomogeneous Dirichlet condition we proceed formally as before and define a function $g(\vec{x}) \in H^1(\Omega)$ such that $g(\vec{x}) = q(\vec{x})$ on Γ. Then we convert the problem into one which has homogeneous Dirichlet boundary conditions. Then our solution is $u(\vec{x}) = w(\vec{x}) + g(\vec{x})$ where w is the unique solution in $H^1_0(0,1)$ of

$$A(w,v) = (f,v) - A(q,v) \quad \forall\, v \in H^1_0(0,1)\,,$$

where, as before,

$$A(w,v) = \int_\Omega \nabla w \cdot \nabla v\, dV\,.$$

The most serious difficulty arises when we try to approximate. In the one-dimensional case, it was easy to determine a function g^h in our approximating subspace; however, for higher dimensions this is not the case. When we choose a finite dimensional approximating subspace, in general, we are not able to find a function g^h in this subspace to use for the function g above. If g^h is not in the approximating subspace then $u^h = w^h + g^h$ is not in the approximating space. We postpone discussion of this problem until a later chapter.

A problem with a mixed boundary condition such as

{2d_Pi2}
$$\begin{aligned} -\Delta u &= f \quad \vec{x} \in \Omega \\ \frac{\partial u}{\partial \vec{n}} + \alpha(\vec{x})u(\vec{x}) &= q(\vec{x}) \quad \text{on } \Gamma \end{aligned} \tag{5.17}$$

can be handled analogous to the one-dimensional case; i.e.,we merely include a term $\int_\Gamma \alpha uv\, ds$ in the bilinear form and add the boundary integral $\int_\Gamma qv\, ds$ to the right-hand side. Recall that in the one-dimensional case, we had to add point values of the solution and/or test function to the bilinear form or right-hand side whereas in the two-dimensional case we are modifying the bilinear form and the right-hand side by a boundary integral.

5.3.2 A Neumann problem for the Helmholtz equation

We have seen that the purely Neumann problem for Poisson's equation; i.e.,when $\Gamma_2 = \partial\Omega$, does not have a unique solution and if we attempt to discretize then we are lead to a singular matrix. If, however, we consider the Neumann problem for the Helmholtz equation

$$\begin{aligned} -\Delta u + \sigma^2 u = f \quad &\text{in } \Omega \\ \frac{\partial u}{\partial \vec{n}} = 0 \quad &\text{on } \partial\Omega \end{aligned}$$

then the problem possesses a unique solution for $u \in C^2$ and sufficiently smooth $\partial\Omega$. In this case the weak formulation is to find $u \in H^1(\Omega)$ such that

$$A(u,v) \equiv \int_\Omega \big(\nabla u \cdot \nabla v + uv\big)\, dV = (f,v) \quad \forall\, v \in H^1(\Omega)\,.$$

This bilinear form is coercive on $H^1(\Omega)$ as well as bounded. In fact for $k^2 = 1$

$$A(u,u) = \int_\Omega \nabla u \cdot \nabla u + u^2\, dV = \|u\|_1^2$$

and in general

$$A(u,u) = \int_\Omega \nabla u \cdot \nabla u + u^2\, dV A(u,u) = \int_\Omega \nabla u \cdot \nabla u + u^2\, dV = \|u\|_1^2 gemin1, k^2\, \|u\|_1^2\,.$$

5.3.3 A fourth order problem

The biharmonic equation is the fourth order partial differential equation

$$\Delta\Delta u = \Delta^2 u = f \quad \text{in } \Omega \ .$$

We may impose boundary conditions such as

$$u = \frac{\partial u}{\partial \vec{n}} = 0 \quad \text{on } \partial\Omega \, .$$

We define $H_0^2(\Omega)$ to be the space

$$H_0^2(\Omega) = \{v \in H^2(\Omega) \mid v = \frac{\partial v}{\partial n} = 0 \ \text{ on } \partial\Omega\}.$$

A weak formulation is to seek $u \in H_0^2(\Omega)$ such that

$$A(u,v) = F(v) \quad \forall \, v \in H_0^2(\Omega)$$

where

$$A(u,v) = \int_\Omega \Delta u \Delta v \, d\Omega \quad \forall \, u,v \in H_0^2(\Omega)$$

and

$$F(v) = \int_\Omega f v \, d\Omega \quad \forall \, v \in H_0^2(\Omega)$$

It can be shown that if u is the classical solution of the biharmonic problem then u satisfies this weak problem; moreover, the weak problem has a unique solution u in $H_0^2(\Omega)$. (See exercises.) To discretize this problem we must use a subspace of $H_0^2(\Omega)$ such as bicubic splines or bicubic Hermites in $\Omega \subset \mathbb{R}^2$. However, in the next chapter we see that when we use a triangular element, it is not so easy to obtain a subspace of H^2.

5.4 Computational examples

Before looking at a specific example, we first compare the number of nodes, the number of unknowns, and the number of quadrature points required to approximate the solution of the problem $-\Delta u + u = f$ with homogeneous, Neumann boundary conditions in one, two and three dimensions. Note that in this purely Neumann problem the number of unknowns is the same as the number of nodes. Specifically we compare the number of unknowns for various values of h for linear, bilinear and trilinear elements as well as for tensor products of quadratic and cubic spaces. We also provide the minimum number of quadrature points that are used in each case. Recall that the number of unknowns corresponds to the size of the matrix and the number of quadrature points influences the amount of work required to compute the entries in the matrix and right-hand sides. In all cases we assume a uniform grid with spacing h in each dimension. The “curse of dimensionality” can clearly be seen from Table 5.2.

Table 5.2. *Comparison of number of unknowns for solving a problem on a domain* $(0,1)^n$, $n = 1,2,3$ *using tensor products of one-dimensional elements.* {2

	Number of unknowns			Number of
	$h = 0.1$	$h = 0.01$	$h = 0.001$	quadrature pts.
linear	11	101	1001	1
bilinear	121	10,201	1.030×10^6	1
trilinear	1331	1.030×10^6	1.003×10^9	1
quadratic	21	201	2001	2
biquadratic	441	40,401	4.004×10^6	4
triquadratic	9261	8.121×10^6	8.012×10^9	8
cubic	31	301	3001	3
bicubic	961	90,601	9.006×10^6	9
tricubic	29,791	2.727×10^7	2.703×10^{10}	27

We now turn to providing some numerical results for the specific problem

$$\begin{aligned} -u''(x) &= (x^2+y^2)\sin(x,y) \qquad \forall\,(x,y)\in\Omega \\ u &= \sin(xy) \quad \text{on } \partial\Omega \end{aligned} \tag{5.18}$$ {2

where $\Omega = \{(x,y)\ :\ 0\le x\le 3,\quad 0\le y\le 3\}$. The exact solution to this problem is $u(x,y) = \sin(xy)$ whose solution is plotted in Figure 5.3 along with a contour plot of the solution. Note that we are imposing inhomogeneous Dirichlet boundary conditions in this example. The results presented here use bilinear and biquadratic elements on a uniform grid of size h in each dimension; for the quadrature rule we use the tensor product of the one point Gauss rule for bilinears and the tensor product of the two point Gauss rule for biquadratics. As usual, a higher order quadrature rule is used to calculate the error. The numerical rates of convergence are obtained using (**??**). The results are presented in Table 5.3 and some results are plotted for the bilinear case in Figure **??**. Note that as expected, the optimal rates of convergence are obtained.

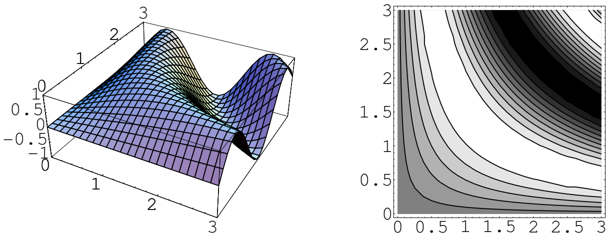


Figure 5.3. {2

Table 5.3. *Numerical results for (5.18) using bilinear and biquadratic elements.*

{2

element	h	No. of unknowns	$\left\|u - u^h\right\|_1$	rate	$\left\|u - u^h\right\|_0$	rate
bilinear	1/4	144	0.87717		0.76184×10^{-1}	
bilinear	1/8	529	0.0.43836	1.0007	0.19185×10^{-1}	1.9895
bilinear	1/16	2209	0.21916	1.0001	0.48051×10^{-2}	1.9973
bilinear	1/32	9216	0.0.10958	1.0000	0.12018×10^{-3}	1.9994
biquadratic	1/4	529	0.70737×10^{-1}		0.22488×10^{-2}	
biquadratic	1/8	2209	0.17673×10^{-1}	1.9758	0.28399×10^{-3}	2.9853
biquadratic	1/16	9025	0.44175×10^{-2}	1.9940	0.35604×10^{-4}	2.9957
biquadratic	1/32	36,491	0.11043×10^{-2}	1.9986	0.44539×10^{-5}	2.9990

Chapter 6

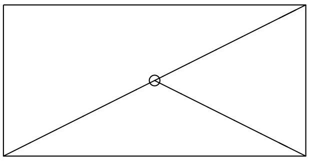
Finite Element Spaces

{c

One of the advantages of the finite element method is that it can be used with relative ease to find approximations to solutions of differential equations on general domains. So far we have only considered approximating in one dimension or in higher dimensions using rectangular elements. The goal of this chapter is to formally define a finite element, present some examples of commonly used elements and to establish a taxonomy for describing elements. Isoparametric elements, which are used for domains with curved boundaries, are discussed in a later chapter.

To precisely describe a particular finite element, it is not enough to give the geometric figure, e.g., a triangle, rectangle, etc. One must also specify the degree of polynomial that is used. Does describing these two pieces of information uniquely determine the choice? In fact, no. If we recall in $\mathbb{R}^1$ using an interval as the geometric element and specifying a cubic polynomial on each interval does not completely describe the finite element because we can determine the cubic by function values at four points or by function and derivative values (as in Hermite cubic) at two points. Consequently, three pieces of information must be provided to give an adequate description of a finite element; we must specify the geometric element, the degree of polynomial, and the degrees of freedom which are used to uniquely determine the polynomial.

Once we have chosen a particular finite element, we subdivide the domain into a finite number of geometric elements; this meshing must be “admissible”, i.e., satisfy certain properties. We want to construct a finite element space, S^h, over this mesh which possesses specific properties. A basic property which we said is a distinguishing feature of the finite element method is that we use a piecewise polynomial which is a kth degree polynomial when restricted to the specific element. For conforming finite elements we require our finite element space to be a subspace of the underlying Hilbert space. For second order problems this space was $H^1(\Omega)$ or a subspace and for fourth order problems the underlying space was $H^2(\Omega)$. Consequently a second property we require is a global smoothness requirement on the space. Finally, for the finite element method to be computationally efficient we must be able to construct a basis which has small support. Before addressing some



{spaces_fig_inadmissible} **Figure 6.1.** *Inadmissible triangulation due to "hanging node"*

of these issues we consider the admissible "triangulations" of a domain.

{spaces_sec_construction}

6.1 Construction of a finite element space

paces_sec_triangulations}

6.1.1 Admissible triangulations

Once a specific geometric element is chosen, we subdivide the domain $\bar{\Omega}$ (here $\bar{\Omega}$ denotes the closure of Ω, i.e., the interior plus the boundary) into a finite number of individual subsets or geometric elements. We will use the terminology *triangulation* to refer to a subdivision of $\bar{\Omega}$ even if the specific geometric element is not a triangle. The subsets form a triangulation of $\bar{\Omega}$, denoted $\mathcal{T}^h$, which must satisfy certain properties. Some of these properties are obvious, such as the fact that their union is $\bar{\Omega}$, while others may not be as obvious. For example, we must add a condition which guarantees there are no "hanging nodes" as indicated in Figure 6.1.

{spaces_def_admissible} **Definition 6.1.** *A subdivision $\mathcal{T}^h$ of Ω into subsets $\{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_M\}$ is an admissible triangulation of Ω if it satisfies the following properties:*

(i) $\overline{\Omega} = \cup_{j=1}^{M} \mathcal{K}_j$ *;*

(ii) for each j, $j = 1, 2, \ldots, M$, the set $\mathcal{K}_j$ is closed and the interior of $\mathcal{K}_j$ is non-empty;

(iii) for each $\mathcal{K}_j$, $j = 1, 2, \ldots, M$, the boundary $\partial\mathcal{K}_j$ is Lipschitz continuous[10] *;*

(iv) if the intersection of two elements $\mathcal{K}_j$ and $\mathcal{K}_\ell$ is nonempty then the intersection must be a common vertex of the elements if the intersection is a single point; otherwise the intersection must be an entire edge or face common to both $\mathcal{K}_\ell$ and $\mathcal{K}_j$.

[10] A domain in Euclidean space with Lipschitz boundary is one whose boundary is "sufficiently regular". Formally, this means that the boundary can be written as, e.g., $z = f(x, y)$ where f is Lipschitz continuous. Recall that a function g is Lipschitz continuous if $\|g(p) - g(q)\| \leq C\,\|p - q\|$. for all p, q.

The penultimate condition in Definition 6.1 allows the application of Green's formula over each element.

The parameter h in the triangulation $\mathcal{T}^h$ is related to the size of the geometric elements and generally gives a measure of the coarseness or fineness of the mesh. It is usually taken to be the diameter of the largest element, i.e., for $\vec{p}, \vec{q} \in \mathbb{R}^n$

$$h_j = \max_{\vec{p},\vec{q} \in \mathcal{K}_j} \left(\sum_{i=1}^{n} |p_i - q_i|^2 \right)^{1/2}, \quad j = 1, 2, \ldots, M$$

and

$$h = \max\{h_1, h_2, \ldots, h_M\}\,.$$

If we have a mesh where all the geometric elements are congruent, then the triangulation is *uniform* if all the elements are the same size; otherwise the triangulation is called *nonuniform.*

Clearly, we are interested in obtaining approximations on successively finer triangulations. For this reason, it is important to look at properties of families of triangulations. For example, we know that when we refine a mesh we can't just make the elements smaller in one portion of the domain but rather refine in some uniform way. To define the concept of a shape regular triangulation we introduce the parameter $\rho = \min_j \rho_j$ where ρ_j denotes the diameter of the largest ball contained in an element $\mathcal{K}_j$. Then a triangulation is called *shape regular* provided there exists a constant σ such that

$$\sigma = \frac{h}{\rho}\,. \tag{6.1}$$

{s

A family of triangulations is called *shape regular* (or just *regular*) provided σ is uniform over the triangulations.

6.1.2 Formal definition of a finite element

{s

From our previous examples in one and two dimensions, we saw that to completely describe a finite element we had to give more information than simply the choice of the geometric element and the degree of the polynomial. In fact, we need three pieces of information – the geometric element, the specific polynomial space defined over the geometric element, and the degrees of freedom needed to uniquely determine the polynomial. We follow Ciarlet's approach for the formal definition of a finite element.

Definition 6.2. *A finite element in $\mathbb{R}^n$ is a triple $(\mathcal{K}, \mathcal{P}_\mathcal{K}, \Theta_\mathcal{K})$ where* {s

(i) $\mathcal{K}$ is a closed subset of $\mathbb{R}^n$ with nonempty interior and a Lipschitz continuous boundary.

(ii) $\mathcal{P}_\mathcal{K}$ is a space of dimension s of real-valued functions over the set $\mathcal{K}$;

(iii) $\Theta_\mathcal{K}$ is a set of s linearly independent functionals, θ_i, $1 \le i \le s$, defined over the space $\mathcal{P}_\mathcal{K}$.

It is assumed that every $p \in \mathcal{P}_{\mathcal{K}}$ is uniquely determined by the values of the s functionals in $\Theta_{\mathcal{K}}$.

The set $\mathcal{K}$ is the specific geometric element in an admissible triangulation. The space $\mathcal{P}_{\mathcal{K}}$ usually consists of a polynomial defined over $\mathcal{K}$; however, we allow a broader definition so as to include some less common elements. In practice, we take these functions to be our basis for the space $\mathcal{P}_{\mathcal{K}}$. The set $\Theta_{\mathcal{K}}$ consists of the degrees of freedom which uniquely determine an element of $\mathcal{P}_{\mathcal{K}}$.

We can not arbitrarily choose a triple $(\mathcal{K}, \mathcal{P}_{\mathcal{K}}, \Theta_{\mathcal{K}})$ to define a finite element because $p \in \mathcal{P}_{\mathcal{K}}$ may not be uniquely determined by the degrees of freedom specified by $\Theta_{\mathcal{K}}$. An obvious example is the case where we don't have enough degrees of freedom specified; however, even if we have enough constraints they still may not uniquely determine the polynomials. To demonstrate that the degrees of freedom uniquely determine the polynomial several approaches can be taken. One approach is to show that the system of equations which results from imposing the degrees of freedom on an arbitary $p \in \mathcal{P}_{\mathcal{K}}$ has a unique solution. An alternate approach is to actually construct a basis for the space $\mathcal{P}_{\mathcal{K}}$. We will demonstrate both techniques when we consider specific finite elements.

6.1.3 Properties of finite element spaces

{spaces_sec_properties}

We subdivide the domain into a finite number of individual elements $\mathcal{K}_j$. On each $\mathcal{K}_j$ the polynomial space $\mathcal{P}_{\mathcal{K}_j}$ is specified along with the degrees of freedom which uniquely determine a polynomial $p \in \mathcal{P}_{\mathcal{K}_j}$ on the element $\mathcal{K}_j$. Then, an associated *finite element space* is defined through a systematic process. In every instance, this space is a finite-dimensional space of functions defined over $\overline{\Omega}$. An outline of the process is given as follows.

First, one defines the *local* properties with respect to each set $\mathcal{K}_j$ of the finite element space S^h. Restricted to each subset $\mathcal{K}_j \subset \overline{\Omega}$, functions belonging to S^h belong to $\mathcal{P}_{\mathcal{K}_j}$. Furthermore, over each $\mathcal{K}_j$, the functions in S^h are determined by the specified degrees of freedom.

Second, one defines the *global* properties with respect to $\overline{\Omega}$ of the finite element space. In particular, the desired order of global continuity and differentiability for S^h must be specified. For example, one could merely require that $S^h \subset C^0(\overline{\Omega})$ or it may be necessary to require that $S^h \subset C^1(\overline{\Omega})$.

The global properties are dictated by the differential equation which is being approximated. We have seen that for second order differential equations the underlying global smoothness of the finite element space is $S^h \subset H^1(\Omega)$ whereas for fourth order problems we require $S^h \subset H^2(\Omega)$. The question then arises how we can guarantee these global properties. Clearly the choice of local properties of S^h influences the global properties.

The following two propositions give conditions which guarantee the standard global smoothness conditions on S^h. The significance of the first proposition is that imposing the global smoothness $S^h \subset H^1(\Omega)$ does not require the functions in S^h to be continuously differentiable but merely continuous; this should be contrasted with the smoothness requirements for the classical solution of a second order boundary value problem. Similarly, the requirement $S^h \subset H^2(\Omega)$ only requires functions $v^h \in S^h$ to be in $C^1(\Omega)$. In the proposition, the additional assumption that $\mathcal{P}_{\mathcal{K}_j} \subset H^1(\mathcal{K}_j)$ is automatically satisfied when $\mathcal{P}_{\mathcal{K}_j}$ is a polynomial space on $\mathcal{K}_j$.

Proposition 6.3. *Assume that $\mathcal{T}^h$ is an admissible triangulation of $\overline{\Omega} \subset \mathbb{R}^n$ into the subsets $\{\mathcal{K}_j\}$. Let $\mathcal{P}_{\mathcal{K}_j} \subset H^1(\mathcal{K}_j)$ for all j, let $S^h \subset C^0(\overline{\Omega})$, and let $v^h|_{\mathcal{K}_j} \in \mathcal{P}_{\mathcal{K}_j}$ for all $v^h \in S^h$. Then $S^h \subset H^1(\Omega)$. Moreover, if S_0^h consists of those functions in S^h which vanish on the boundary of Ω, then* {s

$$S_0^h \equiv \{v^h \in S^h \ : \ v^h = 0 \ \text{on} \ \partial\Omega\} \subset H_0^1(\Omega)\,.$$

Proof. Let $v^h \in S^h$; we must show that $v^h \in H^1(\Omega)$, i.e., that $v^h \in L^2(\Omega)$ and that its first-order weak derivatives belong to $L^2(\Omega)$. Since $v^h \in C^0(\overline{\Omega})$ we have that $v^h \in L^2(\Omega)$. To demonstrate that its first-order weak derivatives are in $L^2(\Omega)$, we must find functions w_i^h, $i = 1, \dots, n$, such that

$$\int_\Omega v^h \frac{\partial \phi}{\partial x_i}\, d\Omega = -\int_\Omega w_i^h \phi \, d\Omega \quad \forall\, \phi \in C_0^\infty(\Omega)\,.$$

For each i, we choose the function w_i^h to be the function whose restriction on each finite element $\mathcal{K}_j$ is the function $\partial(v^h|_{\mathcal{K}_j})/\partial x_i$; this is possible since $\mathcal{P}_{\mathcal{K}_j} \subset H^1(\mathcal{K}_j)$. Since each finite element $\mathcal{K}_j$ has a Lipschitz-continuous boundary $\partial\mathcal{K}_j$, we may apply Green's formula to obtain

$$\int_{\mathcal{K}_j} \frac{\partial}{\partial x_i}\left(v^h|_{\mathcal{K}_j}\right) \phi \, dx = -\int_{\mathcal{K}_j} \left(v^h|_{\mathcal{K}_j}\right) \frac{\partial \phi}{\partial x_i}\, dx + \int_{\partial\mathcal{K}_j} v^h|_{\mathcal{K}_j}\, \phi \, n_{i,\mathcal{K}_j}\, dS\,,$$

where $n_{i,\mathcal{K}_j}$ is the ith component of the unit outer normal along the boundary of $\mathcal{K}_j$. Summing over all the elements, we obtain

$$\int_\Omega w_i^h \phi \, d\Omega = -\int_\Omega v^h \frac{\partial \phi}{\partial x_i}\, d\Omega + \sum_j \int_{\partial\mathcal{K}_j} v^h|_{\mathcal{K}_j} \phi \, n_{i,\mathcal{K}_j}\, dS\,.$$

We are done if we can show that the last term vanishes. The boundary of the elements $\partial\mathcal{K}_j$ can be broken up into segments that are part of $\partial\Omega$ and segments that are also part of the boundary of an adjacent subset, say $\mathcal{K}_\ell$. In the first case, $\phi = 0$ so that clearly those terms vanish. In the other case, the boundary integrals from the two adjacent elements cancel since, by hypothesis, $v^h \in C^0(\overline{\Omega})$ and if two elements $\mathcal{K}_j$ and $\mathcal{K}_\ell$ are adjacent then on their common boundary, $n_{i,\mathcal{K}_j} = -n_{i,\mathcal{K}_\ell}$.

The fact that $S_0^h \subset H_0^1(\Omega)$ follows since $\partial\Omega$ is Lipschitz continuous and if $v^h \in S_0^h$, $v^h = 0$ on $\partial\Omega$. ∎

Proposition 6.4. *Assume that $\mathcal{T}^h$ is an admissible triangulation of $\overline{\Omega} \subset \mathbb{R}^n$ into the subsets $\{\mathcal{K}_j\}$. Let $\mathcal{P}_{\mathcal{K}_j} \subset H^2(\mathcal{K}_j)$ for all j, let $S^h \subset C^1(\Omega)$, and let $v^h|_{\mathcal{K}_j} \in \mathcal{P}_{\mathcal{K}_j}$* {s *for all $v^h \in S^h$. Then, $S^h \subset H^2(\Omega)$. Moreover, if S^h_b consists of all functions that vanish on the boundary, then*

$$S^h_b \equiv \{v^h \in S^h \ : \ v^h = 0 \ \textit{on} \ \partial\Omega\} \subset H^2(\Omega) \cap H^1_0(\Omega) \tag{6.2}$$

{s

and if S^h_0 consists of all functions that vanish on the boundary and whose derivative in the direction of the unit outer normal also vanish on the boundary, then

$$S^h_0 \equiv \{v^h \in S^h \ : \ v^h = \frac{\partial v^h}{\partial \vec{n}} = 0 \ \textit{on} \ \partial\Omega\} \subset H^2_0(\Omega)\,. \tag{6.3}$$

Proof. The proof is analogous to the proof of Proposition 6.3. The details are left to the exercises. ■

6.2 Examples of finite elements on n-simplices

In $\mathbb{R}^2$ the common choices for a geometric element are a triangle and a quadrilateral. If the domain is polygonal and not rectangular, then triangular elements are needed to discretize. In $\mathbb{R}^3$ the commonly used elements are tetrahedra and cubes or bricks. In a later chapter we consider isoparametric elements to handle domains with curved boundaries. In this section we look at some of the more commonly used triangular elements and their variants.

We have seen that to completely specify a finite element, it is not enough to just choose a geometric element. We must also specify the degree of polynomial on the element and the degrees of freedom which uniquely determine the polynomial. To use the element we must also specify a basis which has small support. In the last chapter we saw that for rectangular elements we could simply use tensor products of the basis in one-dimension. For triangles or tetrahedra, this approach does not work. In the following section we see that barycentric coordinates are a useful tool in writing basis functions on a triangle or tetrahedron. In addition, one can consider an approach of determining the basis functions on a reference element and mapping them to the desired element.

In this section and the next we develop a taxonomy for identifying finite elements whether in one, two or three dimensions. We identify the element by its geometric shape which is called an n-simplex or an $n-$rectangle; by its type which indicates the polynomial space, and by whether it is a Lagrange or Hermite element which indicates the kind of degrees of freedom used.

{spaces_sec_simplices}

6.2.1 n-simplices

The first class of finite elements we consider uses subsets $\mathcal{K}$ of $\mathbb{R}^n$ that are *simplices*, e.g., line segments in $\mathbb{R}^1$, triangles in $\mathbb{R}^2$ or tetrahedra in $\mathbb{R}^3$. Formally, we define an *n-simplex* in the following way.

Definition 6.5. *Let z_k, $k = 1, \ldots, n+1$, denote $n+1$ points in $\mathbb{R}^n$. The convex hull of these $n+1$ points, i.e., the intersection of all convex sets* [11] *containing z_k, $k = 1, \ldots, n+1$, is called an n-simplex and the points z_k, $k = 1, \ldots, n+1$, are called the vertices of the n-simplex.*

For example, for $n = 2$ we specify three points $\{z_1, z_2, z_3\}$ and a 2-simplex is simply a triangle with vertices (z_{i_1}, z_{i_2}), $i = 1, 2, 3$, provided the three points are not collinear. To enforce the noncollinearity of the points, we require that the matrix

$$\begin{pmatrix} z_{1_1} & z_{2_1} & z_{3_1} \\ z_{1_2} & z_{2_2} & z_{3_2} \\ 1 & 1 & 1 \end{pmatrix}$$

is nonsingular. Note that the magnitude of the determinant of this matrix is just the area of the parallelogram formed by the vectors $z_2 - z_1$ and $z_3 - z_1$. For $n = 3$, we specify four points $\{z_1, z_2, z_3, z_4\}$ and a 3-simplex is just a tetrahedron with vertices z_i, $i = 1, \ldots, 4$, provided the four points are not coplanar, i.e., provided the matrix

$$\begin{pmatrix} z_{1_1} & z_{2_1} & z_{3_1} & z_{4_1} \\ z_{1_2} & z_{2_2} & z_{3_2} & z_{4_2} \\ z_{1_3} & z_{2_3} & z_{3_3} & z_{4_3} \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

is nonsingular. Note that the magnitude of the determinant of this matrix is the volume of the parallelepiped formed by the vectors $z_i - z_1$, $i = 2, 3, 4$.

For an integer j such that $1 < j \leq n$, any j-simplex whose vertices are a subset of the $(n+1)$ vertices of a given n-simplex is called a *j-face* of the n-simplex. An (n-1)-face is simply called a *face* and any 1-face is called an *edge.* In $\mathbb{R}^2$, triangles have edges, i.e., 1-faces. In $\mathbb{R}^3$, tetrahedra have faces (2-faces) and edges (1-faces.)

6.2.2 Barycentric coordinates

{s

A geometric concept which is useful in easily writing polynomial basis functions on an n-simplex is the idea of barycentric coordinates which were first defined by Möbius in 1827 (Coexeter 1969, p 27; Fauvel 1993). We know that if we are given a frame in $\mathbb{R}^n$, then we can define a local coordinate system with respect to the frame; e.g., Cartesian coordinates. If we are given a set of $n+1$ points in $\mathbb{R}^n$ then we can also define a local coordinate system with respect to these points; such coordinate systems are called *barycentric coordinates.*

Suppose we are given a set of $n+1$ points $z_k \in \mathbb{R}^n$, $k = 1, \ldots, n+1$, such that

[11] Recall that a set $\mathcal{S}$ is convex if given any two points x and y in $\mathcal{S}$ then the line segment joining x and y lies entirely in $\mathcal{S}$.

the determinant of the matrix

$$\begin{pmatrix} z_{1_1} & z_{2_1} & \cdots & z_{n+1_1} \\ z_{1_2} & z_{2_2} & \cdots & z_{n+1_2} \\ \vdots & \vdots & \vdots & \vdots \\ z_{1_n} & z_{2_n} & \cdots & z_{n+1_n} \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{6.4}$$

{s

is nonzero. As we have seen, this is just the condition which guarantees in $\mathbb{R}^2$ that the points are not collinear and in $\mathbb{R}^3$ that the points are not coplanar. Consider the set of all linear combinations of these points of the form

$$q = \lambda_1 z_1 + \lambda_2 z_2 + \cdots \lambda_{n+1} z_{n+1}$$

where

$$\sum_{j=1}^{n+1} \lambda_j = 1 \, .$$

Then the coordinates $(\lambda_1, \lambda_2, \ldots, \lambda_{n+1})$ are called the *barycentric coordinates* of points of the space with respect to the given points z_k, $k = 1, \ldots, n+1$.

As an example of barycentric coordinates consider three specific points in $\mathbb{R}^2$, $z_1 = (0,0)$, $z_2 = (1,0)$ and $z_3 = (1,1)$; the points form a triangle. Any linear combination of these three points such that $\lambda_1 + \lambda_2 + \lambda_3 = 1$ gives the barycentric coordinates (with respect to z_1, z_2, z_3) of a point in $\mathbb{R}^2$. For example, the barycentric coordinates $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ is the point in space with Cartesian coordinates $(\frac{1}{2}, \frac{1}{4})$ since

$$\frac{1}{2}(0,0) + \frac{1}{4}(1,0) + \frac{1}{4}(1,1) = (\frac{1}{2}, \frac{1}{4}) \, .$$

Similarly, the barycentric coordinates $(1,-1,1)$ is the point in space with Cartesian coordinates $(0,1)$ since

$$1 \cdot (0,0) + (-1) \cdot (1,0) + 1 \cdot (1,1) = (0,1) \, .$$

We notice that the point $(\frac{1}{2}, \frac{1}{4})$ with barycentric coordinates $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ lies within the triangle formed by the given z_i, $i = 1,2,3$ whereas the point $(0,1)$ with barycentric coordinates $(1,-1,1)$ is not inside the triangle. In general, one can demonstrate that if $0 \leq \lambda_1, \lambda_2, \lambda_3 \leq 1$ then the point $q = \lambda_1 z_1 + \lambda_2 z_2 + \lambda_3 z_3$ lies inside the triangle. If any λ_k, $k = 1,2,3$, is less than zero or greater than one, then the point q lies outside the triangle. If, for example, $\lambda_1 = 0$, then the point q lies on the edge of the triangle through z_2 and z_3. The justifications of these statements are left to the exercises.

Suppose now we are given a point $(x_1, x_2, \ldots, x_n)$, in a Cartesian coordinate system or some other frame and want to determine the barycentric coordinates of the point with respect to a given set of $n+1$ points. The barycentric coordinates $(\lambda_1, \lambda_2, \ldots, \lambda_{n+1})$ of the point with respect to the prescribed points $n+1$ points z_1,

$z_2,\dots,z_{n+1}$ are found by solving the system

{spaces_bary_system}

$$\begin{aligned}\sum_{j=1}^{n+1} z_{j_i}\lambda_j &= x_i \qquad i=1,\dots,n\\ \sum_{j=1}^{n+1}\lambda_j &= 1\,.\end{aligned} \tag{6.5}$$

Here z_{j_i} denotes the ith component of the point z_j. The coefficient matrix of (6.5) is just the matrix in (6.4) and hence we are guaranteed a unique solution. If we solve this system for the barycentric coordinates, then we see that the $\lambda_j(x)$, $j = 1,\cdots,n{+}1$, are linear functions of the coordinates of the point $x=(x_1,x_2,\dots,x_n)\in\mathbb{R}^n$, i.e.,

$$\lambda_j = \sum_{k=1}^{n}\zeta_{j,k}x_k + \zeta_{j,n+1} \qquad j=1,\dots,n+1\,, \tag{6.6}$$ {s

where $\zeta_{i,j}$ denotes the i,j entry of the inverse of the matrix given in (6.4). For example, the barycentric coordinates with respect to the points $z_1=(0,0)$, $z_2=(1,0)$ and $z_3=(1,1)$ for the Cartesian point $(\frac{3}{4},\frac{1}{2})$ are found by solving the system

$$\begin{pmatrix}0&1&1\\0&0&1\\1&1&1\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_2\end{pmatrix}=\begin{pmatrix}\frac{3}{4}\\\frac{1}{2}\\1\end{pmatrix}$$

to get $(\frac{1}{4},\frac{1}{4},\frac{1}{2})$. We can write the barycentric coordinates as

$$\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_2\end{pmatrix}=\begin{pmatrix}-1&0&1\\1&-1&0\\0&1&0\end{pmatrix}\begin{pmatrix}\frac{3}{4}\\\frac{1}{2}\\1\end{pmatrix}$$

or in the form of (6.6) as $\lambda_1=(-1)x_1+(0)x_2+1$, $\lambda_2=(1)x_1+(-1)x_2+0$, etc.

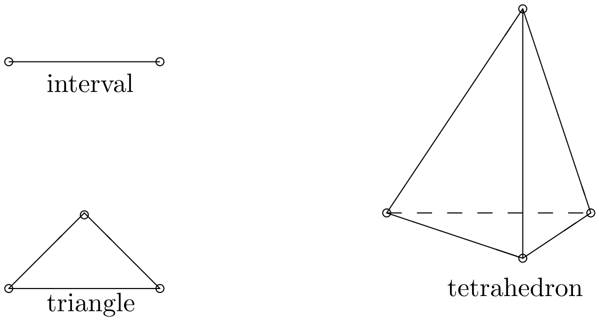
We now want to see how barycentric coordinates can assist us in writing a basis for a linear polynomial space defined on a triangle or tetrahedron where we require that the basis is a nodal basis, i.e., it has the property that it is one at one vertex and is zero at the other vertices. Consider the example of a 2-simplex, i.e., a triangle, with vertices $\{z_1,z_2,z_3\}$. Then, the barycentric coordinates of a point $x=(x_1,x_2)\in\mathbb{R}^2$ are determined by solving the linear system

$$\begin{aligned} z_{1_1}\lambda_1 &+ z_{2_1}\lambda_2 + z_{3_1}\lambda_3 = x_1\\ z_{1_2}\lambda_1 &+ z_{2_2}\lambda_2 + z_{3_2}\lambda_3 = x_2\\ \lambda_1 &+ \lambda_2 + \lambda_3 = 1\,.\end{aligned}$$

It is easy to see that if x is one of the vertices of the 2-simplex, say $x=z_k$, , then $\lambda_j(z_k)=\delta_{jk}$ where $\delta_{jk}=0$ if $j\neq k$ and is equal to one if $j=k$. For example, if $x=z_1$ then the barycentric coordinates of x are $(1,0,0)$. Note also that $\lambda_1(x)$ is zero along the edge formed by z_2 and z_3 since it is a linear function which is zero at z_2 and z_3; thus the side of the triangle formed by the vertices z_2 and z_3 can be described by the equation $\lambda_1=0$.

{spaces_fig_simplex1}

Figure 6.2. *n-simplicies of type(1)*



Summarizing, we have that the barycentric coordinates $(\lambda_1, \lambda_2, \lambda_3)$ are linear functions of x (from (6.6)) which take on the values $(1, 0, 0)$ at $x = z_1$, the values $(0, 1, 0)$ at $x = z_2$ and $(0, 0, 1)$ at $x = z_3$. Consequently, λ_1 is a linear function of x which is one at the vertex z_1 and is zero at the other two vertices z_2 and z_3; similar conditions hold for λ_2 and λ_3. Hence these barycentric coordinates can serve as basis functions for the space of linear polynomials over the triangle formed by the points z_1, z_2, and z_3. When we consider quadratic or higher order basis functions we see that we can simply take appropriate products of the λ_j, $j = 1, \ldots, n + 1$.

6.2.3 Lagrange finite elements on n-simplices

When all the specified degrees of freedom are function values, then the finite element is referred to as a *Lagrange finite element.* Lagrange finite elements on n-simplices lead to finite element spaces that are subspaces of $C^0(\Omega)$ and hence by Proposition 6.3 they are subspaces of $H^1(\Omega)$. Such finite elements are often referred to as "C^0-elements". In the taxonomy of finite elements such elements are called *n-simplices of type (ℓ)* where the qualifier "type (ℓ)" refers to the degree of polynomial specified on the n-simplex.

Lagrange finite element on n-simplices of type (1)

We first consider an n-simplex of type (1); i.e., we are using a linear polynomial defined over an interval in $\mathbb{R}^1$, a triangle in $\mathbb{R}^2$ or a tetrahedron in $\mathbb{R}^3$. These are illustrated in Figure 6.2. We choose $\mathcal{P}(\mathcal{K}) = P_1(\mathcal{K})$ to be linear polynomials defined over $\mathcal{K}$. Since the $\dim(P_1(\mathcal{K})) = 3$ in $\mathbb{R}^2$ and $\dim(P_1(\mathcal{K})) = 4$ in $\mathbb{R}^3$, we expect a linear function on $\mathcal{K}$ to be uniquely determined by its values at the $n + 1$ nodes of the n-simplex. This can be proved in several ways; in the following proposition we prove the result using a linear algebra argument and then following the proof we outline an alternate argument.

{spaces_thm_linpk} **Proposition 6.6.** *Let $\mathcal{K}$ be an n-simplex in $\mathbb{R}^n$, $n = 1, 2, 3$, with vertices $z_1, \ldots, z_{n+1}$. A polynomial $p(x) \in P_1(\mathcal{K})$ is uniquely determined by its values at the vertices.*

Proof. We present a proof for the case $n = 2$ and leave the case $n = 3$ to the exercises; we have already addressed the case of an interval in $\mathbb{R}^1$. Let $p = c_0 + c_1 x_1 + c_2 x_2$ where c_0, c_1, c_2 are constants and let η_i, $i = 1, 2, 3$ be the prescribed values of $p(x)$ at the vertices. Then we must show that there is a unique function $p(x) \in P_1(\mathcal{K})$ such that $p(z_i) = \eta_i$, $i = 1, 2, 3$; i.e., that the linear system

$$\eta_i = c_0 + c_1 z_{i_1} + c_2 z_{i_2} \quad \text{for } i = 1, 2, 3$$

has a unique solution. Note that the requirement that this coefficient matrix be nonsingular is equivalent to the condition which guaranteed that the vertices were not collinear in $\mathbb{R}^2$. ■

Alternately, we could have shown that any polynomial $p(x) \in P_1(\mathcal{K})$ can be written in terms of its values η_i at the vertices. Recall that the barycentric coordinates satisfy $\lambda_i(z_k) = \delta_{ik}$ for $1 \le i, k \le 3$ so that in $\mathbb{R}^2$ the polynomial

$$\eta_1 \lambda_1(x) + \eta_2 \lambda_2(x) + \eta_3 \lambda_3(x)$$

has the desired property; i.e., when we evaluate it at the vertices we get the nodal values. Thus any linear polynomial on an n-simplex with vertices $\{z_1, \ldots, z_{n+1}\}$ can be written as

$$p(x) = \sum_{i=1}^{n+1} p(z_i) \lambda_i(x)\,.$$

Summarizing, we define the *2-simplex of type(1)* to be the set $\mathcal{K}$ where $\mathcal{K}$ is a triangle with vertices z_i, $i = 1, 2, 3$ together with the space $P_1(\mathcal{K})$ and the degrees of freedom of the finite element consisting of the values at the three vertices. A *3-simplex of type(1)* is a set $\mathcal{K}$, where $\mathcal{K}$ is a tetrahedron with vertices z_i, $i = 1, 2, \ldots, 4$, together with the space $P_1(\mathcal{K})$ and the degrees of freedom of the finite element being the values at the four vertices.

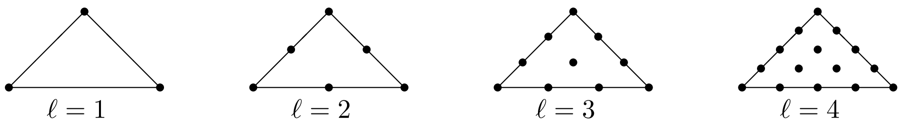
Lagrange finite element on n-simplices of type (2)

Results for n-simplices of type ℓ for $\ell > 1$ follow in an analogous fashion. In $\mathbb{R}^2$ we know that the dimension of $P_2(\mathcal{K})$ is six so we must specify a second degree polynomial at six points to uniquely determine it; the dimension of $P_3(\mathcal{K})$ is ten so that a third degree polynomial must be specified at ten points on the triangle. The most commonly chosen points are the obvious ones. These points form the ℓ^{th} *order principal lattice of an n-simplex $\mathcal{K}$* given by

$$\begin{aligned} \mathcal{L}(\ell, n) \quad = \quad & \Big\{ x = \sum_{k=1}^{n+1} \sigma_k \, z_k \; : \; \sum_{k=1}^{n+1} \sigma_k = 1, \\ & \qquad \sigma_k \in \big\{0, \frac{1}{\ell}, \frac{2}{\ell}, \cdots, \frac{\ell - 1}{\ell}, 1\big\}, 1 \le k \le n+1 \Big\} \end{aligned} \qquad (6.7)$$

{s

Figure 6.3. *ℓ^{th} order lattice for a 2-simplex*



{spaces_fig_prinlattice}

where $z_1, z_2, \ldots, z_{n+1}$ are the vertices of $\mathcal{K}$. It is easy to demonstrate that $\mathcal{L}(\ell, n)$ contains $\binom{\ell+n}{\ell}$ points. For example, in $\mathbb{R}^2$ for $\ell = 1$ $\sigma_k \in \{0, 1\}$ so that the points in $\mathcal{L}(1, 2)$ are $\{z_1, z_2, z_3\}$, i.e., the vertices of the triangle. For $\ell = 2$, $\sigma_k \in \{0, \frac{1}{2}, 1\}$ so that

$$\mathcal{L}(2,2) = \left\{ z_1, z_2, z_3, \frac{z_1 + z_2}{2}, \frac{z_1 + z_3}{2}, \frac{z_2 + z_3}{2} \right\},$$

i.e., the verticies of the triangle and the midpoints of the sides. The ℓth order principal lattice for a 2-simplex for $\ell = 1, 2, 3, 4$ is illustrated in Figure 6.3. The following proposition states that an ℓ^{th} order polynomial on an n-simplex is uniquely determined by its values at the points in the corresponding principal lattice.

paces_thm_puniquelattice}

Proposition 6.7. *Let $\mathcal{K}$ be an n-simplex in $\mathbb{R}^n$ with vertices z_k, $1 \le k \le n+1$. Then for a given integer $\ell \ge 1$, any polynomial $p \in P_\ell$ is uniquely determined by its value at the points in $\mathcal{L}(\ell, n)$ defined by (6.7).*

Proof. The proof is left to the exercies. ∎

We now know that any quadratic polynomial on an n-simplex is uniquely determined by its values at the nodes and the midpoints of the edges of the n-simplex. If we can write any $p \in P_2(\mathcal{K})$ in terms of the specified values at these nodes then we will have a basis for the space. For example, for a 2-simplex we want to write

$$p(x) = \sum_{i=1}^{3} p(z_i) q_i(x) + p(z_{12}) q_{12}(x) + p(z_{13}) q_{13}(x) + p(z_{23}) q_{23}(x)\,,$$

where q_i, $i = 1, 2, 3$, and q_{12}, q_{13}, and q_{23} are quadratic functions on $\mathcal{K}$ and z_{ij} represents the midpoint of the edge joining the nodes z_i and z_j. We use products of the linear barycentric coordinates to write these quadratic functions which serve as our basis functions with small support. First, consider the function q_1 which is a quadratic function which has the properties $q_1(z_1) = 1$ and $q_1(x) = 0$ at the five points z_2, z_3, z_{12}, z_{13}, and z_{23}. Recall that $\lambda_1(x)$ is a linear function such that $\lambda_1(z_1) = 1$, $\lambda_1(z_2) = \lambda_1(z_3) = 0$, so in barycentric coordinates the equation of the line through z_2 and z_3 is just $\lambda_1 = 0$; similarly, the equation through the midpoints z_{12} and z_{13} is $\lambda_1 = 1/2$. Since the point z_{23} lies on the line $\lambda_1 = 0$ we have that the quadratic function

$$\lambda_1(x)\left(\lambda_1(x)-\frac{1}{2}\right)$$

vanishes at the five points z_2, z_3, z_{12}, z_{13}, and z_{23}. Hence we choose $q_1(x) = C\lambda_1(x)\left(\lambda_1(x)-\frac{1}{2}\right)$ and normalize so that $q_1(z_1) = 1$. Since $\lambda_1(z_1) = 1$ we set $C = 2$. In a similar manner $q_i = \lambda_i(x)(2\lambda_i(x)-1)$, $i = 2, 3$. Now we must construct a quadratic function q_{12} which has the properties that $q_{12}(z_{12}) = 1$ and $q_{12}(x) = 0$ at the vertices and the remaining midpoints. In this case the equation of the line through z_1 and z_3 is $\lambda_2 = 0$ and the line through z_2 and z_3 is $\lambda_1 = 0$. Thus the quadratic $\lambda_1(x)\lambda_2(x)$ has the property that it is zero at the verticies and the midpoints z_{23}, z_{13} and takes on the value one-fourth at z_{12}; consequently we take $q_{12}(x) = 4\lambda_1(x)\lambda_2(x)$. In general, $q_{ij}(x) = 4\lambda_i(x)\lambda_j(x)$. Combining these results we have that for $p \in P_2(\mathcal{K})$ where $\mathcal{K}$ is a 2-simplex

$$p = \sum_{i=1}^{3} p(z_i)\lambda_i(2\lambda_i - 1) + 4p(z_{12})\lambda_1\lambda_2 + 4p(z_{13})\lambda_1\lambda_3 + 4p(z_{23})\lambda_2\lambda_3 \,.,$$

For a n-simplex

$$p = \sum_{i=1}^{n+1} p(z_i)\lambda_i(2\lambda_i - 1) + \sum_{\substack{i,k=1\\ i<j}}^{n+1} 4p(z_{ik})\lambda_i\lambda_k \quad \forall\, p \in \mathcal{P}_2(\mathcal{K})\,. \qquad (6.8)$$

{s

Recall that to determine the barycentric coordinates with respect to the points z_i, $i = 1, \ldots, n+1$ we had to solve an $(n+1)\times(n+1)$ linear system of equations.

We now define the *n-simplex of type (2)* to be an n-simplex $\mathcal{K}$ together with the space $P_2(\mathcal{K})$ and the degrees of freedom consisting of the values at the vertices and the midpoints of the edges. Properties of the n-simplex of type(2) are summarized in Table 6.1

The cases $\ell \geq 3$ can be handled in a similar manner. Their properties are summarized in Table 6.1. See the exercises for details.

6.2.4 Hermite 2-simplices

In our examples so far in this chapter we have considered Lagrange finite elements whose degrees of freedom were function values at a prescribed set of points and the resulting finite element spaces were subspaces of $H^1(\Omega)$. In the examples in this section, we consider finite elements in which some of the degrees of freedom are partial derivatives, or more generally, directional derivatives. We denote the partial derivative of a function $p(x)$ in the direction of the line segment through two points $a, b \in \mathbb{R}^n$ and evaluated at a point $x = c \in \mathbb{R}^n$ by $D_{[a,b]}p(c)$. Of course, knowledge of the directional derivatives at a point is equivalent to the knowledge of the partial derivatives.

Hermite 2-simplex of type(3)

{s

Recall that in $\mathbb{R}^1$ we used the space of cubic Hermite polynomials to construct a subspace of $H^2(\Omega)$. However, we see that in $\mathbb{R}^2$ (and also in $\mathbb{R}^3$) using Hermite cubics generates a finite element space which is only a subspace of $C^0(\overline{\Omega})$ and thus only a subspace of $H^1(\Omega)$ by Proposition 6.3. In the next section we consider an example of a $C^1(\Omega)$ triangular element in $\mathbb{R}^2$.

To uniquely determine a cubic polynomial on a triangle, we must specify ten conditions since $\dim P_3 = 10$. The following result gives ten degrees of freedom which are combinations of function and derivative values that uniquely determine a polynomial $p \in P_3$.

{spaces_thm_hermite} **Proposition 6.8.** *Let $\mathcal{K}$ be an 2-simplex with vertices z_i, $1 \le i \le 3$, and let $z_{123} = \frac{1}{3}(z_1 + z_2 + z_3)$. Then any polynomial $p(x)$ in the space $P_3(\mathcal{K})$ is uniquely determined by its value at the vertices, $p(z_i)$, $i = 1, 2, 3$ and the value of its two first partial derivatives at the vertices z_j, $1 \le j \le 3$, and its value at the point z_{123}.*

Proof. First note we are specifying 10 degrees of freedom and $dim(P_3) = 10$ in $\mathbb{R}^2$. To show uniqueness we demonstrate that if $p \in P_3(\mathcal{K})$ and ξ_i, η_{ij}, ζ are given values then the 10×10 system

$$\begin{aligned} p(z_i) &= \xi_i \quad \text{for } i = 1, 2, 3 \\ \frac{\partial p}{\partial x_j}(z_i) &= \eta_{ij} \quad \text{for } i = 1, 2, 3 \text{ and } j = 1, 2 \\ p(z_{123}) &= \zeta \end{aligned}$$

has a unique solution. An easy way to show this is to set all the given values, ξ_i, η_{ij} and ζ, to zero and prove that $p(x)$ must be identically zero.

If we show that $p \in P_3(\mathcal{K})$ is zero along each edge of the triangle, then we know that $p = \alpha\lambda_1\lambda_2\lambda_3$ for some constant α where λ_j are the barycentric coordinates defined by (6.5). Then, since $p(z_{123}) = 0$ we have that $\alpha = 0$ and thus $p(x)$ must be identically zero in $\mathcal{K}$. To demonstrate that $p \in P_3(\mathcal{K})$ is zero along each edge of the triangle we note that along the line containing the vertices z_i and z_j p is a cubic polynomial of one variable and hence we need four conditions to uniquely determine it. But $p(z_i) = p(z_j) = 0$ and that $D_{[z_i,z_j]}p(z_i) = D_{[z_i,z_j]}p(z_j) = 0$ and thus p is zero on each edge $[z_i, z_j]$. ■

We can now define the finite element which is called the *Hermite 2-simplex of type(3)* where the partial derivatives at each vertex are degrees of freedom as well as the values at the vertices and the barycenter. Since knowledge of the directional derivatives at each vertex is equivalent to the knowledge of the partial derivatives at each vertex, we can specify either as degrees of freedom. The properties of the Hermite 2-simplex of type (3) are summarized in Table 6.1. Note that in the illustration of the element in the table we indicate the partial derivative degrees of freedom at z_i by a circle centered at z_i.

We now associate a finite element space S^h with a subdivision of $\overline{\Omega} \subset \mathbb{R}^2$ into Hermite 2-simplices of type(3). Then a function $v^h \in S^h$ implies that the restriction $v^h|_{\mathcal{K}_j}$ is in the space $\mathcal{P}_{\mathcal{K}_j} = P_3(\mathcal{K}_j)$ for each $\mathcal{K}_j$ and is defined by its values at all the vertices of the subdivison, its values at the centers of gravity of all the triangles, and the values of its two first partial derivatives at all the vertices of the subdivision. If we assume that we have an admissible triangulation of our domain into 2-simplices then we are able to obtain the following result.

Lemma 6.9. *Let S^h be the finite element space associated with Hermite 2-simplices of type(3). Then the inclusion* {s

$$S^h \subset C^0(\overline{\Omega}) \cap H^1(\Omega) \tag{6.9}$$

holds.

Proof. Because of Proposition 6.3 it suffices to show that $S^h \subset C^0(\overline{\Omega})$. Along any common side of two adjacent triangles, there is a unique polynomial of degree three in one variable which takes on the prescribed values and prescribed first derviatives at the endpoints of the side yielding a total of four conditions and thus uniquely determines a cubic in one variable. ∎

It is tempting to think that the inclusion $S^h \subset C^1(\overline{\Omega})$ holds for Hermite n-simplices of type(3); however, this is not the case. Although the tangential derivative along an edge is continuous from element to element, the normal derivative is not.

Finally, we should produce a basis set consisting of functions of minimial support. As before, we can use the barycentric coordinates to write a polynomial $p \in P_3$ in terms of its values at the vertices and the barycenter, and the six values of its directional derivatives at the vertices; ultimately they are used to construct a basis for our corresponding finite element space. In particular, we want to write any $p \in P_3$ as a linear combination of appropriate cubic polynomials times the value of p and its partial derivatives at the vertices, z_i, $i = 1, 2, 3$, and its value at the barycenter z_{123}. For example, the cubic basis function at the vertex z_1 should have the property that it is one at z_1, zero at z_2, z_3, z_{123} and, in addition, its partial derivatives at nodes z_i, $i = 1, 2, 3$, should be zero. Specifically, for all $p \in P_3(\mathcal{K})$, $\mathcal{K} \subset \mathbb{R}^2$

$$\begin{aligned} p(x) &= \sum_{i=1}^{3} p(z_i)\left(-2\lambda_i^3 + 3\lambda_i^2 - 7\lambda_1\lambda_2\lambda_3\right) \\ &\quad +27p(z_{123})\lambda_1\lambda_2\lambda_3 + \sum_{i=1}^{3}\sum_{\substack{j=1\\ j\neq i}}^{3} D_{[z_i,z_j]}p(z_i)\lambda_i\lambda_j(2\lambda_i + \lambda_j - 1)\,. \end{aligned} \tag{6.10}$$ {s

It is easy to see that when we evaluate $p(x)$ given by (6.10) at z_i, $1 \leq i \leq 3$ and at z_{123} we get the corresponding function values $p(z_i)$, $1 \leq i \leq 3$, and $p(z_{123})$. It is a little more difficult to show that when we evaluate $D_{[z_i,z_j]}p(x)$ at z_i then the terms multiplying $p(z_i)$ and $p(z_{123})$ are zero and the polynomial multiplying $D_{[z_i,z_j]}p(z_i)$ is one. The proof of this is left to the exercises but basically we must show that $D_{[z_i,z_k]}(\lambda_i\lambda_j\lambda_k)(z_i) = 0$, when we differentiate the term $-2\lambda_i^3 + 3\lambda_i^2$ the terms cancel, and a relationship of the form $D_{[z_i,z_k]}\lambda_j = \delta_{jk} - \lambda_j(z_i)$, $1 \leq k \leq 3$, $k \neq i$ then if $k = j \neq i$ we get the desired result.

6.2.5 C^1 elements on n-simplices

For fourth order differential equations, the inclusion $S^h \subset H^2(\Omega)$ is needed; however, none of the examples presented so far satisfy this condition. Recall that the difficulty in the Hermite 2-simplex was the fact that the normal derivatives did not agree along an edge common to two adjacent elements.

{spaces_sec_argyris}

The Argyris triangle

The first C^1 element which we consider is the *Argyris triangle* which uses a complete polynomial of degree five. The degrees of freedom consist of function values and first and second derivatives at the vertices in addition to normal derivatives at the midpoints of the sides. It can be shown that in $\mathbb{R}^2$ any $p(x) \in P_5$ is uniquely determined by the 21 degrees of freedom given by

$$\Theta_K = \{D^\alpha p(z_i), |\alpha| \leq 2, 1 \leq i \leq 3, \frac{\partial}{\partial n_i} p(z_{jk}), 1 \leq i \leq 3\},$$

where n_i denotes the normal along the edge of the triangle formed by z_j, z_k, $j \neq k \neq i$ and z_{jk} denotes the midpoint of that edge. Note that we have used multi-index notation to denote the derivatives to simplify the statement of the degrees of freedom. The *Argyris 21-degree of freedom triangle* is illustrated in Table 6.1 where we use | to indicate normal derivatives and a circle to indicate derivatives at vertices.

A finite element space is constructed in the usual manner. Since we require the normal derivative at the midpoint of each edge to be a degree of freedom, we expect the normal derivative as well as the tangential derivative along an edge to be continuous. The following result demonstrates that the finite element space generated by using the Argyris triangle is a subspace of $H^2(\Omega)$ and thus can be used to approximate fourth order problems.

{spthmargyris} **Proposition 6.10.** *Let S^h be the finite element space associated with the Argyris triangle. Then the inclusion*

$$S^h \subset C^1(\overline{\Omega}) \cap H^2(\Omega)$$

holds.

Proof. By Proposition 6.4, it suffices to show that $S^h \subset C^1(\overline{\Omega})$. Let $\mathcal{K}_i$ and $\mathcal{K}_j$ be two adjacent triangles with a common side $[b_k, b_\ell]$ where b_k, b_ℓ denote vertices of the triangulation and let $v^h \in S^h$. Considered as functions of an abscissa t along $[b_k, b_\ell]$ the functions $v^h|_{K_i}$ and $v^h|_{\mathcal{K}_j}$ are polynomials of degreee five in the variable t. Call these polynomials q_1 and q_2. Since, by the definition of the space S^h, we have

$$q(b_k) = q'(b_k) = q''(b_k) = q(b_\ell) = q'(b_\ell) = q''(b_\ell) = 0$$

where $q = q_1 - q_2$; it then follows that $q = 0$ and hence the inclusion $S^h \subset C^0(\overline{\Omega})$ holds. Likewise, call r_1 and r_2, the restrictions to the side $[b_k, b_\ell]$ of the functions

$\dfrac{\partial}{\partial \vec{n}} v^h|_{K_i}$ and $\dfrac{\partial}{\partial \vec{n}} v^h|_{K_j}$. Then r_1 and r_2 are polynomials of degree four in the variable t and again, from the definition of S^h, we have the five conditions

$$r(b_k) = r'(b_k) = r(b_{k\ell}) = r(b_2) = r'(b_\ell) = 0$$

where $r = r_1 - r_2$ and $b_{k\ell}$ is the midpoint of the side $[b_k, b_\ell]$. Therefore, $r = 0$. We have thus shown the continuity of the normal derivative. Since $q = 0$ along $[b_k, b_\ell]$, $q' = 0$ along $[b_k, b_\ell]$ also. Therefore, the first derivatives are also continuous on $\overline{\Omega}$. ∎

One difficulty with the Argyris triangle is that there are 21 degrees of freedom. A modification to the Argyris triangle is the *Bell element* which suppresses the values of the normal slopes at the nodes at the three midpoint sides, reducing the degrees of freedom to 18. Functions in the finite element space associated with the Bell element are in a space P_B where $P_4 \subset P_B \subset P_5$. Here P_B denotes the space of all fifth degree polynomials whose normal derivatives along each side of the triangle are third degree polynomials. Note that, in general, in the Argyris triangle the normal derivative of $p \in P_5$ along each edge is a fourth degree polynomial. In this element the degrees of freedom are

$$\Theta_K = \{D^\alpha p(z_i), |\alpha| \le 2, 1 \le i \le 3\}\,.$$

The determination of the basis functions for both the Argyris and Bell triangles is somewhat involved. The reader is referred to [**?**] for details.

Hsieh-Clough-Toucher triangles

{s

In an effort to create an element which generates a finite element space that is a subspace of $H^2(\Omega)$ but which has fewer degrees of freedom, researchers have developed composite type elements commonly called *macro elements*. In the Hsieh-Clough-Tocher triangle, the triangle is first decomposed into three triangles by connecting the barycenter of the given triangle with each of its vertices. On each of the subtriangles a cubic polynomial is constructed so that the resulting function is C^1 on the original triangle. There are a total of 12 degrees of freedom which consist of the function values and first partial derivatives at the three vertices of the original triangle in addition to the normal derivative at the midpoints of the sides of the original triangle.

There is also a reduced Hsieh-Clough-Toucher triangle where the degrees of freedom have been reduced to nine. Once again, the construction of the basis functions are involved; the reader is referred to [**?**, **?**] for details.

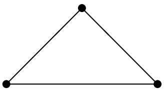
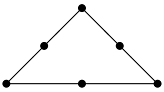
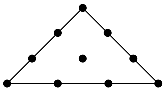
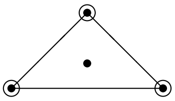
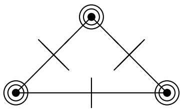
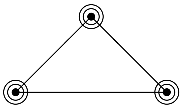
6.3 Examples of finite elements on n-rectangles

{s

In this section we assume that $\overline{\Omega} \subset \mathbb{R}^n$ is a region that can be subdivided into rectangular elements. Many of the results are analogous to those when we subdivide a polyhedral region into n-simplices.

Table 6.1. *Triangular elements*

{s

degrees of freedom	element	$\mathcal{P}_\ell(\mathcal{K})$	$\dim \mathcal{P}_\mathcal{K}$
	2-simplex of type (1)	$\mathcal{P}_1(\mathcal{K})$	3
	2-simplex of type (2)	$\mathcal{P}_2(\mathcal{K})$	6
	2-simplex of type (3)	$\mathcal{P}_3(\mathcal{K})$	10
	Hermite cubic 2-simplex	$\mathcal{P}_3(\mathcal{K})$	10
	Argyris triangle	$\mathcal{P}_5(\mathcal{K})$	21
	Bell triangle	$\mathcal{P}_B \subset \mathcal{P}_5(\mathcal{K})$	18

We let Q_ℓ, for positive integers ℓ, be the space of all polynomials of degree less than or equal to ℓ with respect to *each* of the n variables $x_1, x_2, \ldots, x_n$. For example, if $n = 2$ and $\ell = 1$ Q_1 =span$\{1, x_1, x_2, x_1x_2\}$. We note that we always have the inclusion $P_\ell \subset Q_\ell$ and in general,

{spdimrect}
$$\dim(Q_\ell) = (\ell + 1)^n \,. \tag{6.11}$$

We formally define an *n-rectangle* in $\mathbb{R}^n$ as a product of compact intervals with non-empty interiors.

Definition 6.11. *An n-rectangle, $\mathcal{K}$ in $\mathbb{R}^n$ is defined by*

{sprect}
$$\mathcal{K} = \prod_{i=1}^{n} [a_i, b_i] = \{\vec{x} = (x_1, x_2, \ldots, x_n) \ : \ a_i \le x_i \le b_i, 1 \le i \le n\} \tag{6.12}$$

for finite a_i, b_i for each $i = 1, \ldots, n$.

6.3.1 n-rectangles of type(ℓ)

{spaces_sec_nrectangles}

As in the case of n-simplices, once we have chosen the degree of Q_ℓ then we must specify points for the degrees of freedom, i.e., points where if we prescribe a polynomial of degree ℓ in the n-rectangle then the polynomial is uniquely determined. An easy way to specify the degrees of freedom is to consider a particular n-rectangle, namely the *unit hypercube* $[0,1]^n$ and specify the points on it. Then a linear mapping gives the points on an arbitrary n-rectangle. The following proposition gives a set of points which guarantees that a polynomial in Q_ℓ is uniquely determined by its values on the set.

Proposition 6.12. *A polynomial $p \in Q_\ell$ is uniquely determined by its values on the set* {s

$$\mathcal{M}(\ell,n) = \left\{ x = \left(\frac{i_1}{\ell}, \frac{i_2}{\ell}, \cdots, \frac{i_n}{\ell} \right) \in \mathbb{R}^n \ : \ i_j \in \{0,1,\cdots,\ell\}, 1 \le j \le n \right\} . \quad (6.13)$$ {s

Proof. See exercises. □

For example, in $\mathbb{R}^2$

$$\mathcal{M}(1,2) = \{(0,0),(0,1),(1,0),(1,1)\}$$

and in $\mathbb{R}^3$

$$\mathcal{M}(1,3) = \{(0,0,0),(0,1,0),(0,0,1),(1,0,0),(0,1,1),(1,0,1),(1,1,0),(1,1,1)\} .$$

Thus a 2-rectangle of type(1) consists of a rectangular element $\mathcal{K}$, the space of linear polynomials on $\mathcal{K}$, $Q_1(\mathcal{K})$, whose dimension is 4 and whose degrees of freedom consist of the values at the four vertices. Similarly a 3-rectangle of type(1) consists of a rectangular element $\mathcal{K}$, the linear polynomials on $\mathcal{K}$, $Q_1(\mathcal{K})$, whose dimension is 8 and whose degrees of freedom consist of the values at the eight vertices.

For n-rectangles of type(2)

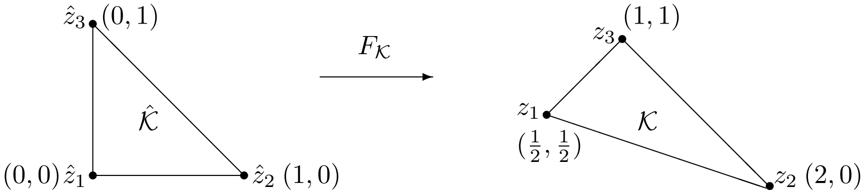
$$\mathcal{M}(2,2) = \{(0,0),(0,1),(1,0),(1,1),(0,\frac{1}{2}),(\frac{1}{2},0),(\frac{1}{2},\frac{1}{2}),(1,\frac{1}{2}),(\frac{1}{2},1)\}$$

Thus a 2-rectangle of type(2) consists of a rectangular element $\mathcal{K}$, the space of quadratic polynomials on $\mathcal{K}$, $Q_2(\mathcal{K})$, whose dimension is 9 and whose degrees of freedom consist of the values at the four vertices, the midpoints of the edges and the barycenter of the rectangle. Similar properties hold for a 3-rectangle of type(2), 2- and 3-rectangles of type(3).

6.3.2 Example of a rectangular C^1 element

{s

For fourth order problems, the inclusion $S^h \subset H^2(\Omega)$ is needed. We can easily define a rectangular element in $\mathbb{R}^2$ for which $S^h \subset H^2(\Omega)$ holds. The element is



{spaces_fig_affine} **Figure 6.4.** *Example of an affine transformation in a 2-simplex*

defined by prescribing $p(z_i)$, $\dfrac{\partial p}{\partial x_1}(z_i)$, $\dfrac{\partial p}{\partial x_2}(z_i)$, $\dfrac{\partial^2}{\partial x_1 \partial x_2}p(z_i)$ at the four vertices of the rectangular element. The resulting polynomial p is in the space $\mathcal{Q}_3$ which has dimension 16. The element is referred to as the *Bogner-Fox-Schmit rectangle.* The proof that the finite element space constructed in the usual manner using this element is a subspace of $C^1(\overline{\Omega})$ is left to the exercises.

6.4 Affine families of finite elements

In this section we want to demonstrate that for many choices of finite elements, instead of specifying a finite element discretization by the data $\mathcal{K}$, $P_{\mathcal{K}}$, and $\Theta_{\mathcal{K}}$, we can prescribe one reference finite element and the affine or linear function which maps the vertices of the reference element into the vertices of the geometric element in the admissible triangulation of the domain. We begin discussion of affine families of finite elements with an example.

We first consider the specific situation depicted in Figure 6.4 where we wish to find an affine mapping which maps the vertices of triangle $\widehat{\mathcal{K}}$ into the vertices of triangle K; i.e., we seek $F_{\mathcal{K}}$ such that $F_{\mathcal{K}}(\hat{z}_i) = z_i$, $i = 1,2,3$ where $\hat{z}_i$ are the vertices of triangle $\widehat{\mathcal{K}}$ and z_i the vertices of triangle $\mathcal{K}$. In this case, $F_{\mathcal{K}}(\hat{x})$ can be explicitly written as

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = F_{\mathcal{K}}(\hat{x}_1,\hat{x}_2) = \frac{1}{2}\begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Clearly $F_{\mathcal{K}}$ maps the vertices in the reference triangle $\hat{\mathcal{K}}$ into the corresponding vertices in triangle $\mathcal{K}$. Moreover, since the mapping is linear, $F_{\mathcal{K}}(\frac{1}{2},0) = (\frac{5}{4},\frac{1}{4})$, $F_{\mathcal{K}}(0,\frac{1}{2}) = (\frac{3}{4},\frac{3}{4})$, and $F_{\mathcal{K}}(\frac{1}{2},\frac{1}{2}) = (\frac{3}{2},\frac{1}{2})$; i.e., the midpoints are preserved under the transformation. In addition, the center of mass is preserved as well other points which we may use as degrees of freedom.

Suppose now that we choose $P_{\mathcal{K}} = P_1(\mathcal{K})$ and $P_{\hat{\mathcal{K}}} = P_1(\hat{\mathcal{K}})$ and we want to compare a basis function $\hat{\phi}_i \in P_1(\hat{\mathcal{K}})$ evaluated at a point $\hat{x}$ with the corresponding basis function in $P_{\mathcal{K}}$ evaluated at $x = F_{\mathcal{K}}(\hat{x})$. For example, the basis function $\hat{\phi}_3$ defined on $\hat{\mathcal{K}}$ which is associated with node $z_3 = (0,1)$ is $\hat{\phi}_3 = \hat{x}_2$ and the basis function ϕ_3 defined on $\mathcal{K}$ which is associated with node $z_3 = (1,1)$ is $\phi_3 = \frac{1}{2}x_1 + \frac{3}{2}x_2 - 1$. If we evaluate each basis function at, e.g., the barycenter we get the same value, i.e., $\hat{\phi}_3(\frac{1}{3},\frac{1}{3}) = \frac{1}{3}$ and $\phi_3(\frac{7}{6},\frac{1}{2}) = \frac{1}{3}$. This is because $(\frac{7}{6},\frac{1}{2}) = F_{\mathcal{K}}(\frac{1}{3},\frac{1}{3})$. Consequently, to evaluate basis functions on $\mathcal{K}$ at quadrature points on $\mathcal{K}$, we simply evaluate the corresponding basis function on the reference triangle at the corresponding quadrature point. However, this is not true when we deal with derivatives of basis functions as when we construct a stiffness matrix. For example, $\frac{\partial\hat{\phi}_3}{\partial\hat{x}_1} = 0$ and $\frac{\partial\phi_3}{\partial x_1} = \frac{1}{2}$. We shouldn't expect this to hold because we are differentiating with respect to different variables so clearly we must consider the transformation. The Jacobian of our transformation is given by

$$J = \begin{pmatrix} \frac{\partial x_1}{\partial\hat{x}_1} & \frac{\partial x_1}{\partial\hat{x}_2} \\ \frac{\partial x_2}{\partial\hat{x}_1} & \frac{\partial x_2}{\partial\hat{x}_2} \end{pmatrix} = \begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} .$$

By the chain rule we have

$$\frac{\partial\phi}{\partial\hat{x}_i} = \frac{\partial\phi}{\partial x_1}\frac{\partial x_1}{\partial\hat{x}_i} + \frac{\partial\phi}{\partial x_2}\frac{\partial x_2}{\partial\hat{x}_i}$$

so that

$$\begin{pmatrix} \frac{\partial\phi}{\partial\hat{x}_1} \\ \frac{\partial\phi}{\partial\hat{x}_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial x_1}{\partial\hat{x}_1} & \frac{\partial x_2}{\partial\hat{x}_1} \\ \frac{\partial x_1}{\partial\hat{x}_2} & \frac{\partial x_2}{\partial\hat{x}_2} \end{pmatrix} \begin{pmatrix} \frac{\partial\phi}{\partial x_1} \\ \frac{\partial\phi}{\partial x_2} \end{pmatrix} = J^T \begin{pmatrix} \frac{\partial\phi}{\partial x_1} \\ \frac{\partial\phi}{\partial x_2} \end{pmatrix} .$$

Thus

$$\begin{pmatrix} \frac{\partial\phi}{\partial x_1} \\ \frac{\partial\phi}{\partial x_2} \end{pmatrix} = J^{-T} \begin{pmatrix} \frac{\partial\phi}{\partial\hat{x}_1} \\ \frac{\partial\phi}{\partial\hat{x}_2} \end{pmatrix} .$$

For our problem this just becomes

$$\begin{pmatrix} \frac{\partial\phi}{\partial x_1} \\ \frac{\partial\phi}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} \end{pmatrix} \begin{pmatrix} \frac{\partial\phi}{\partial\hat{x}_1} \\ \frac{\partial\phi}{\partial\hat{x}_2} \end{pmatrix} .$$

so that with $\hat{\phi}_3 = \hat{x}_2$ we have

$$\frac{\partial\phi_3}{\partial x_1} = \frac{1}{2}\frac{\partial\hat{\phi}_3}{\partial\hat{x}_1} - \frac{1}{2}\frac{\partial\hat{\phi}_3}{\partial\hat{x}_2} = 0 - \frac{1}{2} = \frac{1}{2}$$

$$\frac{\partial\phi_3}{\partial x_2} = \frac{1}{2}\frac{\partial\hat{\phi}_3}{\partial\hat{x}_1} + \frac{3}{2}\frac{\partial\hat{\phi}_3}{\partial\hat{x}_2} = 0 + \frac{3}{2} = \frac{3}{2}$$

which agrees with what we would get if we differentiated $\phi_3(x_1,x_2) = \frac{1}{2}x_1 + \frac{3}{2}x_2 - 1$.

In summary, we have seen that if we have a reference element and an affine function which maps the reference element into a particular $\mathcal{K}$ of our admissible triangulation, then all of the calculations can be performed on the reference element. Moreover, using a reference element and the linear map is a simple way to describe a family of finite elements.

Consider the case where we are given a family $(\mathcal{K},\mathcal{P}_{\mathcal{K}},\Theta_{\mathcal{K}})$ of triangles of type(2) and our goal is to describe this family as simply as possible. Let $\widehat{\mathcal{K}}$ be a reference triangle with vertices $\hat{z}_i$ and edge midpoints $\hat{z}_{ij} = (\hat{z}_i + \hat{z}_j)/2$, $1 \le i < j \le 3$, and let

$$\Theta_{\widehat{\mathcal{K}}} = \{\hat{p}(\hat{z}_i), 1 \le i \le 3;\ \hat{p}(\hat{z}_{ij}), 1 \le i < j \le 3\}$$

so that the element $(\widehat{\mathcal{K}}, P_{\widehat{\mathcal{K}}}, \Theta_{\widehat{\mathcal{K}}})$ with $P_{\widehat{\mathcal{K}}} = P_2(\widehat{\mathcal{K}})$ is also a triangle of type(2). Given any finite element $\mathcal{K}$ in the family, there exists a unique invertible affine mapping

$$F_{\mathcal{K}} : \hat{x} \in \mathbb{R}^2 \to F_{\mathcal{K}}(\hat{x}) = B_{\mathcal{K}}\hat{x} + b_{\mathcal{K}}$$

such that

$$F_{\mathcal{K}}(\hat{z}_i) = z_i, \; 1 \le i \le 3\,;$$

that is, $B_{\mathcal{K}}$ is an invertible 2×2 matrix and $b_{\mathcal{K}}$ a vector in $\mathbb{R}^2$. In the previous example we constructed a specific $F_{\mathcal{K}}$ of this form. Then it automatically follows that

$$F_{\mathcal{K}}(\hat{z}_{ij}) = z_{ij} \quad 1 \le i < j \le 3$$

since the property of a point being the midpoint of a line segment is preserved under an affine mapping. Likewise the points such as $z_{ijk} = \frac{1}{3}(z_i+z_j+z_k)$, $z_{iij} = \frac{2}{3}z_i+\frac{1}{3}z_j$, etc. keep their geometrical definitions through affine transformations. Once we have established the relation $\hat{x} \in \widehat{\mathcal{K}} \to x = F_{\mathcal{K}}(\hat{x}) \in \mathcal{K}$, between the points of the sets $\mathcal{K}$ and $\widehat{\mathcal{K}}$, it is natural to associate the spaces

$$\mathcal{P}^*_{\mathcal{K}} = \{p \; : \; \mathcal{K} \to \mathbb{R}^1; \; p = \hat{p}[F_{\mathcal{K}}^{-1}(x)], \; \hat{p} \in \mathcal{P}_{\widehat{\mathcal{K}}}\}$$

with the space $\mathcal{P}_{\mathcal{K}}$. Then it follows that

$$\mathcal{P}^*_{\mathcal{K}} = \mathcal{P}_{\mathcal{K}} = \mathcal{P}_2(\mathcal{K})$$

since the mapping $F_{\mathcal{K}}$ is affine.

In other words, rather than prescribing the family by the data $\mathcal{K}, \mathcal{P}_{\mathcal{K}}, \Theta_{\mathcal{K}}$, one cas prescribe one reference finite element $(\widehat{\mathcal{K}}, \mathcal{P}_{\widehat{\mathcal{K}}}, \Theta_{\widehat{\mathcal{K}}})$ and the affine mappings $F_{\mathcal{K}}$. Then for our example of a 2-simplex of type(2), a typical element in the family $(\mathcal{K}, \mathcal{P}_{\mathcal{K}}, \Theta_{\mathcal{K}})$ is such that

$$\left\{ \begin{array}{lcl} \mathcal{K} & = & F_{\mathcal{K}}(\widehat{\mathcal{K}}) \\ \mathcal{P}_{\mathcal{K}} & = & \{p : \mathcal{K} \to \mathbb{R}^1 \; : \; p = \hat{p}[F_{\mathcal{K}}^{-1}(x)], \; \hat{p} \in \mathcal{P}_{\widehat{\mathcal{K}}}\} \\ \Theta_{\mathcal{K}} & = & \{p[F_{\mathcal{K}}(\hat{z}_i)], 1 \le i \le 3; \; p[F_{\mathcal{K}}(\hat{z}_{ij})], 1 \le i < j \le 3\}\,. \end{array} \right.$$

With this example in mind, we can now give the general definition that two finite elements $(\widehat{\mathcal{K}}, \mathcal{P}_{\widehat{\mathcal{K}}}, \Theta_{\widehat{\mathcal{K}}})$ and $(\mathcal{K}, \mathcal{P}_{\mathcal{K}}, \Theta_{\mathcal{K}})$, with degrees of freedom of the form (**??**), are said to be *affine-equivalent* if there exists an *invertile affine mapping*

$$F : \hat{x} \in \mathbb{R}^n \to F(\hat{x}) = B\hat{x} + b \in \mathbb{R}^n$$

such that the following relations hold:

$$\mathcal{K} = F(\widehat{\mathcal{K}}) \tag{6.14}$$

$$\mathcal{P}_{\mathcal{K}} = \{p : \mathcal{K} \to \mathbb{R}^1; p = \hat{p}[F^{-1}(x)], \hat{p} \in \mathcal{P}_{\widehat{\mathcal{K}}}\} \tag{6.15}$$

$$\tag{6.16}$$

whenever the nodes z_i ($\hat{z}_i$ occur in the definitions of the set $\Theta_{\mathcal{K}}$ ($\Theta_{\widehat{\mathcal{K}}}$). It is clear that two n-simplices of type(ℓ) for a given $\ell \ge 1$ are affine-equivalent. Likewise, two n-rectangles of type(ℓ) are affine-equivalent through diagonal affine transformations.

Indeed, any two identical Lagrange finite elements that we have considered are affine-equivalent. The situation for Hermite elements is less simple. For example, consider two Hermite n-simplices of type(3) with sets of degrees of freedom involving $D_{[z,z_j]}p(z_i)$. Then it is clear that they are affine-equivalent because the relations

$$z_j - z_i = F(\hat{z}_j) - F(\hat{z}_i) = B(\hat{z}_j - \hat{z}_i), 1 \leq i, j \leq n, i \neq j\,.$$

On the other hand,the Argyris 21-degee of freedom triangle, is *not, in general, affine-equivalent* unless they are equilateral triangles since the normal derivative degrees of freedom are not preserved through an affine transformation, i.e., the property of a vector that it be perpendicular to a hyperplane is not, in general, preserved through an affine mapping.

A family of finite elements is called an *affine family* if all its finite elements are affine-equivalent to a single finite element, which is called the *reference finite element* of the family. Note that the reference element, which we denote by $(\widehat{\mathcal{K}}, \mathcal{P}_{\widehat{\mathcal{K}}}\Theta_{\widehat{\mathcal{K}}})$ need not belong to the family. In the case of an affine family consisting of n-simplices, it is customary to choose the set $\widehat{\mathcal{K}}$ to be the *unit n-simplex* with vertices

$$\hat{z}_1 = (1, 0, \ldots, 0), \hat{z}_2 = (0, 1, 0, \ldots) \cdots \hat{z}_n = (0, 0, \ldots, 0, 1), \hat{z}_{n+1} = (0, 0, \ldots, 0)$$

for which the barycentric coordinates take the simple form

$$\lambda_i = x_i 1 \leq i \leq n, \text{and} \quad \lambda_{n+1} = 1 - \sum_{i=1}^{n} x_i\,.$$

In the case of an affine family of rectangular elements, the usual choice for the refence set $\widehat{\mathcal{K}}$ is either the unit hypercube $[0,1]^n$ or the hypercube $[-1,1]^n$.

The concept of affine family of finite elements is important because (i) in practical computations the calculations for the matrix entries are performed on the reference element; and (ii) for such families an elegant interpolation theory can be developed, which in turn is the basis for most of the convergence theorems concerning finite element approximations.