Finite Element Methods

Max D. Gunzburger Janet S. Peterson

January 7, 2009

Chapter 1

Introduction

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?

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

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

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

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

and

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

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:

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

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:

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

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

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

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 **??**.

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 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

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 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.

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

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

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

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

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

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. 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 it 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

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

and

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

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 $|i - j| \leq 1$. 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

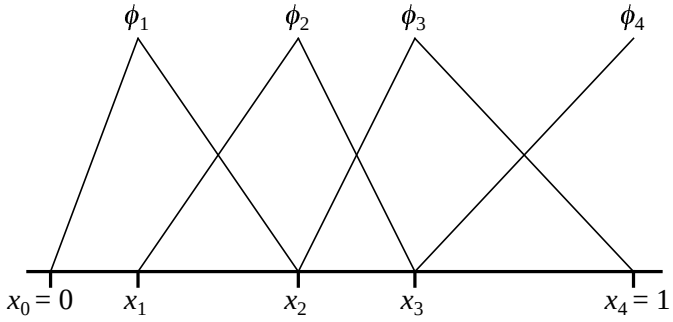

Figure 1.1. *Example of the hat basis functions for four intervals.*

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?

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

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

[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

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

so that

$$\sqrt{\int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right)^2 dx} \;\leq\; \sqrt{\int_0^1 \left(\frac{du}{dx} - \frac{dw^h}{dx}\right)^2 dx} \tag{1.13}$$

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*

$$\sqrt{\int_0^1 \left(\frac{du}{dx} - \frac{du^h}{dx}\right)^2 dx} \;\leq\; Ch \tag{1.14}$$

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. This will require formulating an weak problem which requires the introduction of appropriate function spaces and bilinear forms. In Chapter **??** we give a brief overview of selected topics from functional analysis and in Chapter **??** 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

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 $i = 0, \ldots, N$. 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 $i = 1, \ldots, N$, 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

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

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

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

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

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

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, $i = 1, \ldots, N$, is uniform with $h = x_i - x_{i-1}$ for $i = 1, \ldots, N$, 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)

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

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?

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?

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)

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

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.