14

The finite element method

14.1 Introduction: the model problem

In Chapter 13 we explored finite difference methods for the numerical solution of two-point boundary value problems. The present chapter is devoted to the foundations of the theory of finite element methods. For the sake of simplicity the exposition will be, at least initially, confined to the second-order ordinary differential equation

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right)+r(x)u=f(x)\,,\quad a<x<b\,, \tag{14.1}$$

where $p\in \mathrm{C}^1[a,b]$, $r\in \mathrm{C}[a,b]$, $f\in \mathrm{L}^2(a,b)$ and $p(x)\geq c_0>0$, $r(x)\geq 0$ for all $x\in[a,b]$, subject to the boundary conditions

$$u(a)=A\,,\quad u(b)=B\,. \tag{14.2}$$

Later on in the chapter, in Section 14.5, we shall also consider the ordinary differential equation

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right)+q(x)\frac{\mathrm{d}u}{\mathrm{d}x}+r(x)u=f(x)\,,\quad a<x<b\,, \tag{14.3}$$

subject to the boundary conditions (14.2). Indeed, much of the material discussed here can be extended to partial differential equations; for pointers to the relevant literature we refer to the Notes at the end of the chapter.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

The finite element method was proposed in a paper by Richard Courant in the early 1940s,[1] although the historical roots of the method can be traced back to earlier work by Galerkin[2] in 1915; unfortunately, the relevance of Courant's article was not recognised at the time and the idea was forgotten. In the early 1950s the method was rediscovered by engineers, but its systematic mathematical analysis began only a decade later. Since then, the finite element method has been developed into one of the most general and powerful techniques for the numerical solution of differential equations which is widely used in engineering design and analysis.

Unlike finite difference schemes which seek to approximate the unknown analytical solution to a differential equation at a finite number of selected points, the grid points or mesh points in the computational domain, the finite element method supplies an approximation to the analytical solution in the form of a piecewise polynomial function, defined over the entire computational domain. For example, in the case of the boundary value problem (14.1), (14.2), the simplest finite element method uses a linear spline, defined over the interval $[a, b]$, to approximate the analytical solution u.

We shall consider two techniques for the construction of finite element approximations: the **Rayleigh–Ritz principle** and the **Galerkin principle**. In the case of the boundary value problem (14.1), (14.2) the approximations which stem from these two principles will be seen to coincide. We note, however, that since the Rayleigh–Ritz principle relies on the fact that the boundary value problem under consideration can be restated as a variational problem involving the minimisation of a certain quadratic functional over a function space, its use is restricted to *symmetric* boundary value problems, such as (14.1), (14.2) where (14.1) does not contain a first-derivative term; for example, the Rayleigh–Ritz principle is not applicable to (14.3), (14.2) unless $q(x) \equiv 0$. The precise sense in which the word *symmetric* is to be interpreted here will be clar-

[1] R. Courant, Variational methods for the solution of problems in equilibrium and vibrations, *Bull. Amer. Math. Soc.* **49**, 1–23, 1943; Richard Courant (8 January 1888, Lublinitz, Prussia, Germany (now Lubliniec, Poland) – 27 January 1972, New Rochelle, New York, USA). For an illuminating account of the lives of Richard Courant and David Hilbert, see the book of Constance Reid: *Hilbert–Courant*, Springer, New York, 1986.

[2] Boris Grigorievich Galerkin (4 March 1871, Polotsk, Russia (now in Belarus) – 12 June 1945, Moscow, USSR) studied mathematics and engineering at the St Petersburg Technological Institute. During his studies he supported himself by private tutoring and working as a designer. His ideas on the approximate solution of differential equations were published in 1915. From 1940 until his death, Galerkin was head of the Institute of Mechanics of the Soviet Academy of Sciences.

ified later in the chapter. On the other hand, as we shall see in Section 14.5, the Galerkin principle is more generally applicable and does not require symmetry of the boundary value problem.

To make these observations rigorous, we recall from Chapter 11 the concept of **Sobolev space**.

Definition 14.1 *For a positive integer k, we define the* **Sobolev space** $\mathrm{H}^k(a,b)$ *as the set of real-valued functions v defined on $[a,b]$ such that v and all of its derivatives of order up to and including $k-1$ are absolutely continuous on $[a,b]$ and*

$$v^{(k)} = \frac{\mathrm{d}^k v}{\mathrm{d}x^k} \in \mathrm{L}^2(a,b)\,.$$

Here $\mathrm{L}^2(a,b)$ denotes the set of all functions defined on (a,b) such that

$$\|v\|_2 = \|v\|_{\mathrm{L}^2(a,b)} = \left(\int_a^b |v(x)|^2 \mathrm{d}x\right)^{1/2}$$

is finite. We equip $\mathrm{H}^k(a,b)$ with the **Sobolev norm**

$$\|v\|_{\mathrm{H}^k(a,b)} = \left(\sum_{m=0}^{k} \|v^{(m)}\|^2_{\mathrm{L}^2(a,b)}\right)^{1/2},$$

where $v^{(0)} = v$.

The Sobolev spaces $\mathrm{H}^1(a,b)$ and $\mathrm{H}^2(a,b)$ corresponding, respectively, to $k = 1$ and $k = 2$ will be particularly relevant in this chapter. The next definition introduces variants of the space $\mathrm{H}^1(a,b)$ required for the imposition of the boundary conditions (14.2).

Definition 14.2 *(i) Given that A and B are real numbers, $\mathrm{H}^1_{\mathrm{E}}(a,b)$ will denote the set of all functions $v \in \mathrm{H}^1(a,b)$ such that $v(a) = A$ and $v(b) = B$.*

(ii) $\mathrm{H}^1_0(a,b)$ will signify the set of all functions $v \in \mathrm{H}^1(a,b)$ such that $v(a) = 0$ and $v(b) = 0$.

In the next section we shall state, using Sobolev spaces, the Rayleigh–Ritz and Galerkin principles associated with the boundary value problem (14.1), (14.2), and explore their relationship.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

14.2 Rayleigh–Ritz and Galerkin principles

The Rayleigh–Ritz principle relies on converting the boundary value problem (14.1), (14.2) into a variational problem involving the minimisation of a certain quadratic functional over a function space.

Let us define the quadratic functional $\mathcal{J}\colon \mathrm{H}^1_{\mathrm{E}}(a,b) \to \mathbb{R}$ by

$$\mathcal{J}(w) = \tfrac{1}{2}\int_a^b [p(x)(w')^2 + r(x)w^2]\mathrm{d}x - \int_a^b f(x)w(x)\mathrm{d}x$$

where $w \in \mathrm{H}^1_{\mathrm{E}}(a,b)$, and consider the following *variational problem*:

$$\text{(RR)} \quad \text{find } u \in \mathrm{H}^1_{\mathrm{E}}(a,b) \text{ such that } \mathcal{J}(u) = \min_{w\in \mathrm{H}^1_{\mathrm{E}}(a,b)} \mathcal{J}(w)\,,$$

which we shall henceforth refer to as the **Rayleigh–Ritz principle**. For the sake of notational simplicity we define

$$\mathcal{A}(w,v) = \int_a^b [p(x)\,w'(x)v'(x) + r(x)\,w(x)v(x)]\mathrm{d}x$$

and recall from Chapter 9 the definition of inner product on $\mathrm{L}^2(a,b)$:

$$\langle w, v\rangle \;\; = \int_a^b w(x)v(x)\mathrm{d}x\,.$$

Using these, we can rewrite $\mathcal{J}(w)$ as follows:

$$\mathcal{J}(w) = \tfrac{1}{2}\mathcal{A}(w,w) - \langle f, w\rangle\,, \qquad w \in \mathrm{H}^1_{\mathrm{E}}(a,b)\,. \tag{14.4}$$

The mapping $\mathcal{A}\colon \mathrm{H}^1(a,b)\times \mathrm{H}^1(a,b) \to \mathbb{R}$ is a **bilinear functional** in the following sense:

❶ $\mathcal{A}(\lambda_1 w_1 + \lambda_2 w_2, v) = \lambda_1\mathcal{A}(w_1,v) + \lambda_2\mathcal{A}(w_2,v)$ for all $\lambda_1,\lambda_2 \in \mathbb{R}$ and all $w_1, w_2, v \in \mathrm{H}^1(a,b)$;

❷ $\mathcal{A}(w, \mu_1 v_1 + \mu_2 v_2) = \mu_1\mathcal{A}(w,v_1) + \mu_2\mathcal{A}(w,v_2)$ for all $\mu_1,\mu_2 \in \mathbb{R}$ and all $w, v_1, v_2 \in \mathrm{H}^1(a,b)$.

We note, in addition, that the bilinear functional $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ is **symmetric**, in that

$$\mathcal{A}(w,v) = \mathcal{A}(v,w) \qquad \forall\, w, v \in \mathrm{H}^1(a,b)\,. \tag{14.5}$$

Our next result provides an equivalent characterisation of the Rayleigh–Ritz principle; it relies on the fact that the bilinear functional $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ is symmetric in the sense of (14.5).

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

Theorem 14.1 *A function u in $\mathrm{H}^1_{\mathrm{E}}(a,b)$ minimises $\mathcal{J}(\,\cdot\,)$ over $\mathrm{H}^1_{\mathrm{E}}(a,b)$ if, and only if,*

$$\text{(G)} \qquad \mathcal{A}(u,v) = \langle f, v\rangle \qquad \forall\, v \in \mathrm{H}^1_0(a,b). \tag{14.6}$$

This identity will be referred to as the **Galerkin principle**.

Proof of theorem Suppose that $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ minimises $\mathcal{J}(\,\cdot\,)$ over $\mathrm{H}^1_{\mathrm{E}}(a,b)$; that is, $\mathcal{J}(u) \le \mathcal{J}(w)$ for all $w \in \mathrm{H}^1_{\mathrm{E}}(a,b)$. Noting that $w = u + \lambda v$ belongs to $\mathrm{H}^1_{\mathrm{E}}(a,b)$ for all $\lambda \in \mathbb{R}$ and all $v \in \mathrm{H}^1_0(a,b)$, we deduce that

$$\begin{aligned} \mathcal{J}(u) &\le \mathcal{J}(u+\lambda v) = \tfrac{1}{2}\mathcal{A}(u+\lambda v, u+\lambda v) - \langle f, u+\lambda v\rangle \\ &= \mathcal{J}(u) + \lambda[\mathcal{A}(u,v) - \langle f, v\rangle] + \tfrac{1}{2}\lambda^2 \mathcal{A}(v,v) \end{aligned} \tag{14.7}$$

for all $v \in \mathrm{H}^1_0(a,b)$ and all $\lambda \in \mathbb{R}$. Here, in the transition from the first line to the second we made use of the fact that $\mathcal{A}(u,v) = \mathcal{A}(v,u)$ for all v in $\mathrm{H}^1_0(a,b)$, which follows from (14.5). Now, (14.7) implies that

$$-\tfrac{1}{2}\lambda^2 \mathcal{A}(v,v) \le \lambda[\mathcal{A}(u,v) - \langle f, v\rangle]$$

for all $v \in \mathrm{H}^1_0(a,b)$ and all $\lambda \in \mathbb{R}$. Let us suppose that $\lambda > 0$, divide both sides of the last inequality by λ and pass to the limit $\lambda \to 0$ to deduce that

$$0 \le \mathcal{A}(u,v) - \langle f, v\rangle \qquad \forall\, v \in \mathrm{H}^1_0(a,b)\,. \tag{14.8}$$

On replacing v by $-v$ in (14.8), we have that also

$$0 \ge \mathcal{A}(u,v) - \langle f, v\rangle \qquad \forall\, v \in \mathrm{H}^1_0(a,b)\,. \tag{14.9}$$

We conclude from (14.8) and (14.9) that

$$\mathcal{A}(u,v) = \langle f, v\rangle \qquad \forall\, v \in \mathrm{H}^1_0(a,b)\,, \tag{14.10}$$

as required.

Conversely, if $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ is such that $\mathcal{A}(u,v) = \langle f, v\rangle$ for all v in $\mathrm{H}^1_0(a,b)$, then

$$\mathcal{J}(u+\lambda v) = \mathcal{J}(u) + \lambda\,[\mathcal{A}(u,v) - \langle f, v\rangle] + \tfrac{1}{2}\lambda^2 \mathcal{A}(v,v) \ge \mathcal{J}(u)$$

for all $v \in \mathrm{H}^1_0(a,b)$ and all $\lambda \in \mathbb{R}$; therefore, u minimises $\mathcal{J}(\,\cdot\,)$ over $\mathrm{H}^1_{\mathrm{E}}(a,b)$. □

Thus we have shown that, as long as $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ is a symmetric bilinear functional, $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ satisfies the Rayleigh–Ritz principle if, and only if, it satisfies the Galerkin principle.[1] Our next task is to explain the

[1] In the language of the calculus of variations, (G) is the Euler–Lagrange equation for the minimisation problem (RR).

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

relationship between (RR) and (G) on the one-hand and (14.1), (14.2) on the other. Since in the case of a symmetric bilinear functional $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ the principles (RR) and (G) are equivalent, it is sufficient to clarify the connection between (G), for example, and the boundary value problem (14.1), (14.2).

We begin with the following definition.

Definition 14.3 *If a function $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ satisfies the Galerkin principle (14.6), it is called a* **weak solution** *to the boundary value problem (14.1), (14.2), and the Galerkin principle is referred to as the* **weak formulation** *of the boundary value problem (14.1), (14.2).*

Let us justify this terminology. Suppose that $u \in \mathrm{H}^2(a,b) \cap \mathrm{H}^1_{\mathrm{E}}(a,b)$ is a solution to the boundary value problem (14.1), (14.2). Then,

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right) + r(x)u = f(x)\,, \tag{14.11}$$

for almost every $x \in (a,b)$ (see the discussion prior to Example 11.1 for a definition of **almost every**). Multiplying this equality by an arbitrary function $v \in \mathrm{H}^1_0(a,b)$, and integrating over (a,b), we conclude that

$$-\int_a^b \frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right) v\,\mathrm{d}x + \int_a^b r(x)uv\,\mathrm{d}x = \int_a^b f(x)v(x)\,\mathrm{d}x\,.$$

On integration by parts in the first term on the left-hand side,

$$-\int_a^b \frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right) v\,\mathrm{d}x = \left[p(x)\frac{\mathrm{d}u}{\mathrm{d}x}v\right]_{x=a}^{b} + \int_a^b p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x\,.$$

Since, by hypothesis, $v(a) = 0$ and $v(b) = 0$, it follows that

$$\int_a^b p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x + \int_a^b r(x)uv\,\mathrm{d}x = \int_a^b f(x)v(x)\,\mathrm{d}x$$

for all $v \in \mathrm{H}^1_0(a,b)$. Thus, we have shown the following result.

Theorem 14.2 *If $u \in \mathrm{H}^2(a,b) \cap \mathrm{H}^1_{\mathrm{E}}(a,b)$ is a solution to the boundary value problem (14.1), (14.2), then u is a weak solution to this problem; that is,*

$$\mathcal{A}(u,v) = \langle f, v\rangle \qquad \forall\, v \in \mathrm{H}^1_0(a,b)\,. \tag{14.12}$$

The converse implication, namely that any weak solution $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ of (14.1), (14.2) belongs to $\mathrm{H}^2(a,b) \cap \mathrm{H}^1_{\mathrm{E}}(a,b)$ and solves (14.1), (14.2) in the usual (pointwise) sense, is not true in general, unless the weak

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

solution can be shown to be sufficiently smooth to belong to $\mathrm{H}^2(a,b)$. It is for this reason that any function $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ satisfying (14.12) is called a *weak* solution of the original boundary value problem.

Thus, Theorem 14.1 shows that $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ is a weak solution to (14.1), (14.2) if, and only if, it minimises $\mathcal{J}(\,\cdot\,)$ over $\mathrm{H}^1_{\mathrm{E}}(a,b)$. Next, we show that if a weak solution exists then it must be unique.

Theorem 14.3 *The boundary value problem (14.1), (14.2) possesses at most one weak solution in* $\mathrm{H}^1_{\mathrm{E}}(a,b)$.

Proof The proof is by contradiction. Suppose that $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ and $\tilde{u} \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ are two weak solutions to (14.1), (14.2). Then, $u - \tilde{u}$ belongs to $\mathrm{H}^1_0(a,b)$, and

$$\mathcal{A}(u-\tilde{u},v) = \mathcal{A}(u,v) - \mathcal{A}(\tilde{u},v) = \langle f, v\rangle - \langle f, v\rangle = 0$$

for all $v \in \mathrm{H}^1_0(a,b)$. In particular,

$$\mathcal{A}(u-\tilde{u},u-\tilde{u}) = 0\,.$$

However, since $p(x) \geq c_0 > 0$ and $r(x) \geq 0$ for all x in $[a,b]$,

$$\mathcal{A}(v,v) = \int_a^b [p(x)(v')^2 + r(x)v^2]\mathrm{d}x \geq c_0 \int_a^b |v'|^2 \mathrm{d}x\,.$$

On choosing $v = u - \tilde{u}$, this implies that

$$0 = \mathcal{A}(u-\tilde{u},u-\tilde{u}) \geq c_0 \int_a^b |(u-\tilde{u})'|^2 \mathrm{d}x\,.$$

Since the right-hand side in the last inequality is nonnegative, it follows that $(u-\tilde{u})'(x) = 0$ for almost every x in (a,b); as $u - \tilde{u}$ is absolutely continuous on $[a,b]$ and $(u-\tilde{u})(a) = (u-\tilde{u})(b) = 0$, we conclude that $u = \tilde{u}$, and hence we get the desired uniqueness of a weak solution. □

It turns out that under the present hypotheses on p, q and f the *existence* of a weak solution $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ is also ensured, although the proof of this is less simple and is omitted here; the interested reader is referred to the literature listed in the Notes at the end of the chapter.

14.3 Formulation of the finite element method

In the previous section we showed that the weak solution to the boundary value problem (14.1), (14.2) minimises $\mathcal{J}(\,\cdot\,)$ over $\mathrm{H}^1_{\mathrm{E}}(a,b)$. The finite element method is based on constructing an approximate solution u^h to

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

the problem by minimising $\mathcal{J}(\,\cdot\,)$ over a finite-dimensional subset S_{E}^h of $\mathrm{H}_{\mathrm{E}}^1(a,b)$, instead.

A simple way of constructing S_{E}^h is to choose any function $\psi \in \mathrm{H}_{\mathrm{E}}^1(a,b)$, for example,

$$\psi(x) = \frac{B-A}{b-a}(x-a) + A \tag{14.13}$$

and a finite set of linearly independent functions φ_j, $j = 1, \ldots, n-1$, in $\mathrm{H}_0^1(a,b)$ for $n \geq 2$, and then define

$$\begin{aligned} S_{\mathrm{E}}^h \quad = \quad & \{v^h \in \mathrm{H}_{\mathrm{E}}^1(a,b)\colon v^h(x) = \psi(x) + \sum_{i=1}^{n-1} v_i \varphi_i(x)\,, \\ & \qquad \text{where } (v_1, \ldots, v_{n-1})^{\mathrm{T}} \in \mathbb{R}^{n-1}\}\,. \end{aligned}$$

We consider the following approximation of problem (RR):

$$(\mathrm{RR})^h \quad \text{find } u^h \in S_{\mathrm{E}}^h \text{ such that } \mathcal{J}(u^h) = \min\nolimits_{w^h \in S_{\mathrm{E}}^h} \mathcal{J}(w^h)\,.$$

Our next result is a finite-dimensional analogue of Theorem 14.1.

Theorem 14.4 *A function $u^h \in S_{\mathrm{E}}^h$ minimises $\mathcal{J}(\,\cdot\,)$ over S_{E}^h if, and only if,*

$$(\mathrm{G})^h \qquad \mathcal{A}(u^h, v^h) = \langle f, v^h \rangle \qquad \forall\, v^h \in S_0^h. \tag{14.14}$$

Here,

$$\begin{aligned} S_0^h \quad = \quad & \{v^h \in \mathrm{H}_0^1(a,b)\colon v^h(x) = \sum_{i=1}^{n-1} v_i \varphi_i(x)\,, \\ & \qquad \text{where } (v_1, \ldots, v_{n-1})^{\mathrm{T}} \in \mathbb{R}^{n-1}\}\,. \end{aligned}$$

The problem $(\mathrm{G})^h$ can be thought of as an approximation to the Galerkin principle (G), and is therefore referred to as the **Galerkin method**. For a similar reason, $(\mathrm{RR})^h$ is called the Rayleigh–Ritz method, or just **Ritz method**. Thus, in complete analogy with the equivalence of (RR) and (G) formulated in Theorem 14.1, Theorem 14.4 now expresses the equivalence of $(\mathrm{RR})^h$ and $(\mathrm{G})^h$, the approximations to (RR) and (G), respectively. Of course, as in the case of (RR) and (G), the equivalence of $(\mathrm{RR})^h$ and $(\mathrm{G})^h$ relies on the assumption that the bilinear functional $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ is symmetric. The proof is identical to that of Theorem 14.1, and is left as an exercise.

Theorem 14.4 provides no information about the existence and uniqueness of u^h that minimises $\mathcal{J}(\,\cdot\,)$ over S_{E}^h (or, equivalently, of the existence

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

and uniqueness of u^h that satisfies (14.14)). This question is settled by our next result.

Theorem 14.5 *There exists a unique function $u^h \in S_{\mathrm{E}}^h$ that minimises $\mathcal{J}(\,\cdot\,)$ over S_{E}^h; this u^h is called the* **Ritz approximation** *to u. Equivalently, there exists a unique function $u^h \in S_{\mathrm{E}}^h$ that satisfies (14.14); this u^h is called the* **Galerkin approximation** *to u. The Ritz and Galerkin approximations to u coincide.*

Proof We shall prove the second of these two equivalent statements: we shall show that there exists a unique $u^h \in S_{\mathrm{E}}^h$ that satisfies (14.14). The proof of uniqueness of $u^h \in S_{\mathrm{E}}^h$ is analogous to the proof of Theorem 14.3, with u, $\tilde{u}$, $\mathrm{H}_{\mathrm{E}}^1(a,b)$ and $\mathrm{H}_0^1(a,b)$, replaced by u^h, $\tilde{u}^h$, S_{E}^h and S_0^h, respectively. Since S_{E}^h is finite-dimensional, the uniqueness of u^h satisfying (14.14) implies its existence. □

Having shown the existence and uniqueness of u^h minimising $\mathcal{J}(\,\cdot\,)$ over S_{E}^h (or, equivalently, satisfying (14.14)), we adopt the following definition.

Definition 14.4 *The functions φ_i, $i = 1, 2, \ldots, n-1$, appearing in the definitions of S_{E}^h and S_0^h are called the* **Galerkin basis functions**.

Since any function $v^h \in S_0^h$ can be represented as a linear combination of the Galerkin basis functions φ_i, $1 \le i \le n-1$, it is clear that (14.14) is equivalent to

$$\mathcal{A}(u^h, \varphi_i) = \langle f, \varphi_i \rangle\,, \quad 1 \le i \le n-1\,. \tag{14.15}$$

As u^h belongs to S_{E}^h, it can be expressed in terms of ψ and the Galerkin basis functions as

$$u^h(x) = \psi(x) + \sum_{j=1}^{n-1} u_j \varphi_j(x)\,,$$

where $u_j \in \mathbb{R}$, $j = 1, \ldots, n-1$, are to be determined. On substituting this expansion of u^h into (14.15), we arrive at the following system of simultaneous linear equations:

$$\sum_{j=1}^{n-1} M_{ij} u_j = b_i\,, \quad 1 \le i \le n-1\,, \tag{14.16}$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

where

$$M_{ij} = \mathcal{A}(\varphi_j, \varphi_i)\,, \quad b_i = \langle f, \varphi_i \rangle - \mathcal{A}(\psi, \varphi_i)\,. \tag{14.17}$$

The coefficients u_j, $1 \le j \le n-1$, in the representation of the approximate solution are thus obtained by solving the system of linear equations (14.16). The matrix M is, clearly, symmetric (since the bilinear form $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ is symmetric by hypothesis) and positive definite, because

$$\boldsymbol{v}^{\mathrm{T}} M \boldsymbol{v} = \mathcal{A}(v, v) > 0\,,$$

where $\boldsymbol{v} = (v_1, v_2, \ldots, v_{n-1})^{\mathrm{T}} \in \mathbb{R}^{n-1}$ is any nonzero vector and $v = v_1\varphi_1 + \cdots + v_{n-1}\varphi_{n-1} \in S_0^h$.

The Ritz and Galerkin methods can be used to compute an approximation u^h to u as a linear combination of *any* finite set of linearly independent functions φ_i, $1 \le i \le n-1$, in $\mathrm{H}_0^1(a,b)$. We obtain the **Ritz finite element method** and the **Galerkin finite element method**, respectively, when we select the approximating subspaces S_{E}^h and S_0^h in the Ritz or the Galerkin method to be spaces of spline functions (see Chapter 11). Here we only consider the simplest case of linear splines, and choose the basis functions φ_i, $1 \le i \le n-1$, to be the hat functions (11.4). We begin by fixing a set of points x_k, $k = 0, 1, \ldots, n$, $n \ge 2$, in the interval $[a,b]$ such that

$$a = x_0 < x_1 < \cdots < x_n = b\,. \tag{14.18}$$

The intervals $[x_{i-1}, x_i]$, $1 \le i \le n$, are referred to as **elements**; hence the name *finite element method*. In the theory of the finite element methods (14.18) is called a **subdivision** of the computational domain $[a,b]$, and the points x_k are called **mesh points**. The function φ_i is the piecewise linear function which takes the value 0 at all the mesh points except x_i, where it takes the value 1. Thus,

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h_i & \text{if } x_{i-1} \le x \le x_i\,, \\ (x_{i+1} - x)/h_{i+1} & \text{if } x_i \le x \le x_{i+1}\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{14.19}$$

where $h_i = x_i - x_{i-1}$. The functions φ_i, $1 \le i \le n-1$, are called the (piecewise linear) **finite element basis functions** and the associated Galerkin approximation u^h is referred to as the (piecewise linear) **finite element approximation** of u. The closure of the interval (x_{i-1}, x_{i+1}) over which φ_i is nonzero is called the **support** of the function φ_i. The piecewise linear finite element basis function φ_i, $1 \le i \le n-1$, with support $[x_{i-1}, x_{i+1}]$, is depicted in Figure 14.1.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

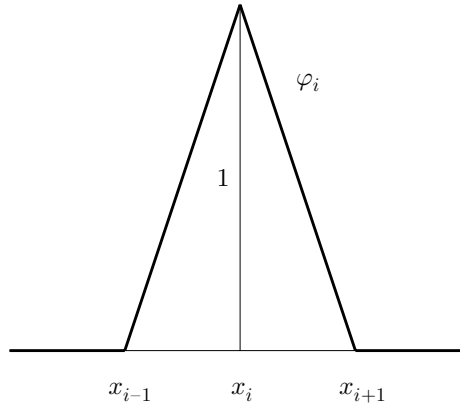

Fig. 14.1. A piecewise linear finite element basis function, φ_i, $1 \le i \le n-1$.

For the finite element method the important property of the basis functions φ_i, $1 \le i \le n-1$, is that they have *local* support, being nonzero only in one pair of adjacent intervals, $(x_{i-1}, x_i]$ and $[x_i, x_{i+1})$. This means that, in the matrix M,

$$M_{ij} = 0 \quad \text{if } |i-j| > 1\,.$$

The matrix M is, therefore, symmetric, positive definite and tridiagonal, and the associated system of linear equations can be solved very efficiently by the methods of Section 3.3, the most efficient algorithm being LU decomposition, without any use of symmetry. The fact that M is positive definite means that no interchanges are necessary.

The function ψ in (14.13), which is included in the definition of S_{E}^h to ensure that u^h satisfies the boundary conditions at $x = a$ and $x = b$, is then given by

$$\psi(x) = A\varphi_0(x) + B\varphi_n(x)\,,$$

which is also piecewise linear; clearly, $\psi(a) = A$ and $\psi(b) = B$. Here, φ_0 and φ_n are defined by setting, respectively, $i = 0$ and $i = n$ in (14.19) and restricting the resulting functions to the interval $[a, b] = [x_0, x_n]$. In (14.17) we see that the term $\mathcal{A}(\psi, \varphi_i)$ is nonzero only for $i = 1$ and $i = n-1$.

Before attempting to solve the system of linear equations we must, of course, first compute the elements of the matrix M, and the quantities on the right-hand side, b_i, $i = 1, \ldots, n-1$; see (14.16) and (14.17). The

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

matrix elements are obtained from

$$M_{ij} = \mathcal{A}(\varphi_j, \varphi_i) = \int_a^b p(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x + \int_a^b r(x)\varphi_j(x)\varphi_i(x)\mathrm{d}x\,,$$

with $1 \le i, j \le n-1$. We have written this as the sum of two terms, as the matrix M is often written in this way as the sum of two matrices which, for historical reasons, are often known as the **stiffness matrix** and the **mass matrix**, respectively. The terms M_{ij} are very simple; in fact in the first integral the derivatives φ_j' and φ_i' are piecewise constant functions over $[a, b]$.

It may be possible to compute these integrals analytically, but more generally some form of numerical quadrature will be necessary. It is then easy to show that if we use certain types of quadrature formulae we shall be led to the same system of equations as in the finite difference method of Section 13.5. Consider the particularly simple case where the mesh points are equally spaced, so that $x_j = a + jh$, $j = 0, 1, \ldots, n$, $h = (b-a)/n$. If we then approximate the integrals involved in the stiffness matrix by the midpoint rule (see Chapter 10), we obtain

$$\begin{aligned}\int_{x_{i-1}}^{x_i} p(x)\varphi_{i-1}'(x)\varphi_i'(x)\mathrm{d}x &= -(1/h^2)\int_{x_{i-1}}^{x_i} p(x)\mathrm{d}x \\ &\approx -p_{i-1/2}/h\,,\end{aligned}$$

where $p_{i-1/2} = p(x_i - h/2)$, and similarly for the other integrals involved. For the integrals in the mass matrix we use the trapezium rule, and then

$$\int_{x_{i-1}}^{x_i} r(x)\varphi_{i-1}(x)\varphi_i(x)\mathrm{d}x \approx 0\,,$$

since φ_i is zero at x_{i-1} and φ_{i-1} is zero at x_i. In the same way

$$\int_{x_{i-1}}^{x_i} r(x)[\varphi_i(x)]^2\mathrm{d}x \approx \tfrac{1}{2}hr_i\,,$$

where $r_i = r(x_i)$, since φ_i is zero at one end of the interval and unity at the other. The other part of the integral is, similarly,

$$\int_{x_i}^{x_{i+1}} r(x)[\varphi_i(x)]^2\mathrm{d}x \approx \tfrac{1}{2}hr_i\,. \tag{14.20}$$

Assuming that $f \in \mathrm{C}[a, b]$, approximating the integral on the right-hand side by the trapezium rule in the same way, and putting all the parts together, equation (14.14) now takes the approximate form

$$-\frac{p_{i-1/2}}{h}u_{i-1} + \frac{p_{i-1/2} + p_{i+1/2}}{h}u_i - \frac{p_{i+1/2}}{h}u_{i+1} + hr_iu_i = hf_i\,,$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

for $i = 1, 2, \ldots, n-1$, with the notational convention that $u_0 = A$ and $u_n = B$, and $f_i = f(x_i)$; clearly, this is the same as the finite difference equation (13.19). Of course, had we used a different set of basis functions φ_i, $1 \leq i \leq n-1$, or different numerical quadrature rules, the finite element and finite difference methods would have no longer been identical. Indeed, this example is just an illustration of the relation between the two methods; we should normally expect to compute the entries of the matrix M by using some more accurate quadrature method, such as a two-point Gauss formula.

In the next two sections we shall assess the accuracy of the finite element method. Our goal is to quantify the amount of reduction in the error $u - u^h$ as the mesh spacing h is reduced.

14.4 Error analysis of the finite element method

We begin with a fundamental result that underlies the error analysis of finite element methods.

Theorem 14.6 (Céa's Lemma) *Suppose that u is the function that minimises $\mathcal{J}(u)$ over $\mathrm{H}^1_{\mathrm{E}}(a,b)$ (or, equivalently, that u satisfies (14.6)), and that u^h is its Galerkin approximation obtained by minimising $\mathcal{J}(\,\cdot\,)$ over S^h_{E} (or, equivalently, that u^h satisfies (14.14)). Then,*

$$\mathcal{A}(u - u^h, v^h) = 0 \qquad \forall\, v^h \in S^h_0\,, \tag{14.21}$$

and

$$\mathcal{A}(u - u^h, u - u^h) = \min_{v^h \in S^h_{\mathrm{E}}} \mathcal{A}(u - v^h, u - v^h)\,. \tag{14.22}$$

The identity (14.21) is referred to as **Galerkin orthogonality**. The terminology stems from the fact that, since the bilinear functional $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ is symmetric and $\mathcal{A}(v,v) > 0$ for all $v \in \mathrm{H}^1_0(a,b) \setminus \{0\}$, $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ is an inner product in the linear space $\mathrm{H}^1_0(a,b)$. Therefore, by virtue of Definition 9.2, (14.21) means that $u - u^h$ is orthogonal to S^h_0 in $\mathrm{H}^1_0(a,b)$. A geometrical illustration of Galerkin orthogonality is given in Figure 14.2. Given that ψ is a fixed element of $\mathrm{H}^1_{\mathrm{E}}(a,b)$, the mapping

$$R^h\colon u - \psi \in \mathrm{H}^1_0(a,b) \mapsto u^h - \psi \in S^h_0$$

which assigns a $u^h \in S^h_{\mathrm{E}}$ to $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ (where u and u^h are as in Theorem 14.6) is called the **Ritz projector**.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

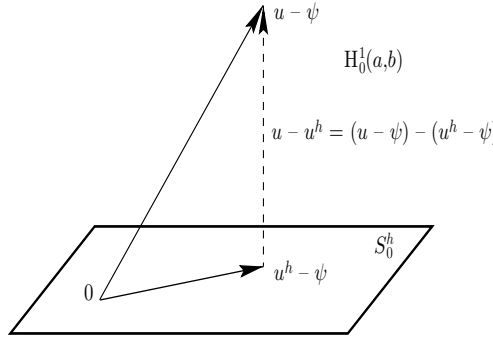

Fig. 14.2. Illustration of the Galerkin orthogonality property of the finite element method. $\mathcal{A}((u-\psi)-(u^h-\psi),v^h)=\mathcal{A}(u-u^h,v^h)=0$ for all v^h in S_0^h. Here, $\psi(x)=A\varphi_0(x)+B\varphi_n(x)$, so that $u-\psi\in\mathrm{H}_0^1(a,b)$ and $u^h-\psi\in S_0^h$. The 0 in the figure denotes the zero element of the linear space S_0^h (and, simultaneously, that of $\mathrm{H}_0^1(a,b)$), namely the function that is identically zero on the interval (a,b).

Proof of theorem By the definition of the Galerkin method $(\mathrm{G})^h$,

$$\mathcal{A}(u^h,v^h)=\langle f,v^h\rangle \qquad \forall\, v^h\in S_0^h\,.$$

On the other hand, we deduce from (G) that

$$\mathcal{A}(u,v^h)=\langle f,v^h\rangle \qquad \forall\, v^h\in S_0^h\,,$$

since $v^h\in S_0^h\subset\mathrm{H}_0^1(a,b)$. The Galerkin orthogonality property (14.21) follows by subtraction.

Now suppose that v^h is any function in S_{E}^h; then,

$$\begin{aligned}\mathcal{A}(u-v^h,u-v^h) &= \mathcal{A}(u-u^h+u^h-v^h,u-u^h+u^h-v^h)\\ &= \mathcal{A}(u-u^h,u-u^h)+\mathcal{A}(u^h-v^h,u^h-v^h)\\ &\quad +2\,\mathcal{A}(u-u^h,u^h-v^h)\\ &= \mathcal{A}(u-u^h,u-u^h)+\mathcal{A}(u^h-v^h,u^h-v^h)\,,\end{aligned}$$

by Galerkin orthogonality, given that $u^h-v^h\in S_0^h$. In the transition from the first line to the second, we made use of the fact that the bilinear functional $\mathcal{A}$ is symmetric. As the term $\mathcal{A}(u^h-v^h,u^h-v^h)$ is nonnegative, we deduce that

$$\mathcal{A}(u-u^h,u-u^h)\le\mathcal{A}(u-v^h,u-v^h) \qquad \forall\, v^h\in S_0^h\,,$$

with equality when $v^h=u^h$; hence (14.22). □

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

Motivated by the minimisation property (14.22), we define the **energy norm** $\|\cdot\|_{\mathcal{A}}$ on $\mathrm{H}^1_0(a,b)$ *via*

$$\|v\|_{\mathcal{A}} = [\mathcal{A}(v,v)]^{1/2}\,. \tag{14.23}$$

Under our hypotheses on p and q, it is easy to see that $\|\cdot\|_{\mathcal{A}}$ satisfies all axioms of norm (see Chapter 2). The result we have just proved shows that u^h is the *best approximation* from S^h_{E} to the true solution $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ of our problem, when we measure the error of the approximation in the energy norm:

$$\|u-u^h\|_{\mathcal{A}} = \min_{v^h \in S^h_{\mathrm{E}}} \|u-v^h\|_{\mathcal{A}}\,. \tag{14.24}$$

A particularly relevant question is how the error $u-u^h$ depends on the spacing h of the subdivision of the computational domain $[a,b]$. We can obtain a bound on the error $u-u^h$, measured in the energy norm, by choosing a particular function $v^h \in S^h_{\mathrm{E}}$ in (14.24) whose closeness to u is easy to assess. For this purpose, we introduce the **finite element interpolant** $\mathcal{I}^h u \in S^h_{\mathrm{E}}$ of $u \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ by

$$\mathcal{I}^h u(x) = \psi(x) + \sum_{i=1}^{n-1} u(x_i)\varphi_i(x)\,, \qquad x \in [a,b]\,.$$

Clearly,

$$\mathcal{I}^h u(x_j) = u(x_j)\,, \qquad j = 0,1,\ldots,n\,,$$

which justifies our use of the word *interpolant*.

We then deduce from (14.24) that

$$\|u-u^h\|_{\mathcal{A}} \le \|u-\mathcal{I}^h u\|_{\mathcal{A}}\,; \tag{14.25}$$

hence, in order to quantify $\|u-u^h\|_{\mathcal{A}}$, we only need to estimate the size of $\|u-\mathcal{I}^h u\|_{\mathcal{A}}$. This leads us to the next theorem.

Theorem 14.7 *Suppose that $u \in \mathrm{H}^2(a,b) \cap \mathrm{H}^1_{\mathrm{E}}(a,b)$ and let $\mathcal{I}^h u$ be the finite element interpolant of u from S^h_{E} defined above; then, the following error bounds hold:*

$$\begin{aligned} \|u-\mathcal{I}^h u\|_{\mathrm{L}^2(x_{i-1},x_i)} &\le \left(\frac{h_i}{\pi}\right)^2 \|u''\|_{\mathrm{L}^2(x_{i-1},x_i)}\,, \\ \|u'-(\mathcal{I}^h u)'\|_{\mathrm{L}^2(x_{i-1},x_i)} &\le \frac{h_i}{\pi}\|u''\|_{\mathrm{L}^2(x_{i-1},x_i)}\,, \end{aligned}$$

for $i = 1,2,\ldots,n$, where $h_i = x_i - x_{i-1}$.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

Proof Consider an element $[x_{i-1}, x_i]$, $1 \le i \le n$, and define $\zeta(x) = u(x) - \mathcal{I}^h u(x)$ for $x \in [x_{i-1}, x_i]$. Then, $\zeta \in \mathrm{H}^2(x_{i-1}, x_i)$ and $\zeta(x_{i-1}) = \zeta(x_i) = 0$. Therefore ζ can be expanded into a convergent Fourier sine-series,

$$\zeta(x) = \sum_{k=1}^{\infty} a_k \sin \frac{k\pi(x - x_{i-1})}{h_i}, \qquad x \in [x_{i-1}, x_i].$$

Here, convergence is to be understood in the norm $\|\cdot\|_{\mathrm{L}^2(x_{i-1},x_i)}$. Hence,

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} [\zeta(x)]^2 \mathrm{d}x &= \int_{x_{i-1}}^{x_i} \zeta(x)\zeta(x)\mathrm{d}x \\
&= \sum_{k,\ell=1}^{\infty} a_k a_\ell \int_{x_{i-1}}^{x_i} \sin \frac{k\pi(x - x_{i-1})}{h_i} \sin \frac{\ell\pi(x - x_{i-1})}{h_i} \mathrm{d}x \\
&= h_i \sum_{k,\ell=1}^{\infty} a_k a_\ell \int_0^1 \sin k\pi t \, \sin \ell\pi t \, \mathrm{d}t \\
&= \frac{h_i}{2} \sum_{k,\ell=1}^{\infty} a_k a_\ell \, \delta_{k\ell} \\
&= \frac{h_i}{2} \sum_{k=1}^{\infty} |a_k|^2,
\end{aligned}$$

where $\delta_{k\ell}$ is the Kronecker delta. Differentiating the Fourier sine series of ζ twice, we find that the Fourier coefficients of ζ' are $(k\pi/h_i)a_k$, while those of ζ'' are $-(k\pi/h_i)^2 a_k$. Thus, proceeding in the same way as above,

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} [\zeta'(x)]^2 \mathrm{d}x &= \frac{h_i}{2} \sum_{k=1}^{\infty} \left(\frac{k\pi}{h_i}\right)^2 |a_k|^2, \\
\int_{x_{i-1}}^{x_i} [\zeta''(x)]^2 \mathrm{d}x &= \frac{h_i}{2} \sum_{k=1}^{\infty} \left(\frac{k\pi}{h_i}\right)^4 |a_k|^2.
\end{aligned}$$

Because $k^4 \ge k^2 \ge 1$, it follows that

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} [\zeta(x)]^2 \mathrm{d}x &\le \left(\frac{h_i}{\pi}\right)^4 \int_{x_{i-1}}^{x_i} [\zeta''(x)]^2 \mathrm{d}x, \\
\int_{x_{i-1}}^{x_i} [\zeta'(x)]^2 \mathrm{d}x &\le \left(\frac{h_i}{\pi}\right)^2 \int_{x_{i-1}}^{x_i} [\zeta''(x)]^2 \mathrm{d}x.
\end{aligned}$$

However, $\zeta''(x) = u''(x) - (\mathcal{I}^h u)''(x) = u''(x)$ for $x \in (x_{i-1}, x_i)$, and hence the desired bounds on the interpolation error. □

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

Now, substituting the bounds from Theorem 14.7 into the definition of the norm $\|u - \mathcal{I}^h u\|_{\mathcal{A}}$, we arrive at the following estimate of the interpolation error in the energy norm.

Corollary 14.1 *Suppose that* $u \in \mathrm{H}^2(a,b) \cap \mathrm{H}^1_{\mathrm{E}}(a,b)$. *Then,*

$$\|u - \mathcal{I}^h u\|^2_{\mathcal{A}} \leq \sum_{i=1}^{n} \left\{ \left(\frac{h_i}{\pi}\right)^2 P_i + \left(\frac{h_i}{\pi}\right)^4 R_i \right\} \|u''\|^2_{\mathrm{L}^2(x_{i-1},x_i)} ,$$

where $P_i = \max_{x\in[x_{i-1},x_i]} p(x)$ *and* $R_i = \max_{x\in[x_{i-1},x_i]} r(x)$.

Proof Let us observe that

$$\begin{aligned} \|v\|^2_{\mathcal{A}} &= \mathcal{A}(v,v) \\ &= \int_a^b \left[p(x)|v'(x)|^2 + r(x)|v(x)|^2\right] \mathrm{d}x \\ &= \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left[p(x)|v'(x)|^2 + r(x)|v(x)|^2\right] \mathrm{d}x \\ &\leq \sum_{i=1}^{n} \left\{ P_i \|v'\|^2_{\mathrm{L}^2(x_{i-1},x_i)} + R_i \|v\|^2_{\mathrm{L}^2(x_{i-1},x_i)} \right\}. \end{aligned}$$

On letting $v = u - \mathcal{I}^h u$ and applying the preceding theorem on the right-hand side of the last inequality, with v' and v replaced by $u' - (\mathcal{I}^h u)'$ and $u - \mathcal{I}^h u$, respectively, the result follows. □

Inserting this estimate into (14.25) leads to the desired bound on the error between the analytical solution u and its finite element approximation u^h in the energy norm.

Corollary 14.2 *Suppose that* $u \in \mathrm{H}^2(a,b) \cap \mathrm{H}^1_{\mathrm{E}}(a,b)$. *Then,*

$$\|u - u^h\|^2_{\mathcal{A}} \leq \sum_{i=1}^{n} \left\{ \left(\frac{h_i}{\pi}\right)^2 P_i + \left(\frac{h_i}{\pi}\right)^4 R_i \right\} \|u''\|^2_{\mathrm{L}^2(x_{i-1},x_i)} ,$$

where $P_i = \max_{x\in[x_{i-1},x_i]} p(x)$ *and* $R_i = \max_{x\in[x_{i-1},x_i]} r(x)$. *Further,*

$$\|u - u^h\|_{\mathcal{A}} \leq \frac{h}{\pi} \left\{ P + \left(\frac{h}{\pi}\right)^2 R \right\}^{1/2} \|u''\|_{\mathrm{L}^2(a,b)} , \tag{14.26}$$

where $P = \max_{x\in[a,b]} p(x)$, $R = \max_{x\in[a,b]} r(x)$, *and* $h = \max_{1\leq i\leq n} h_i$.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

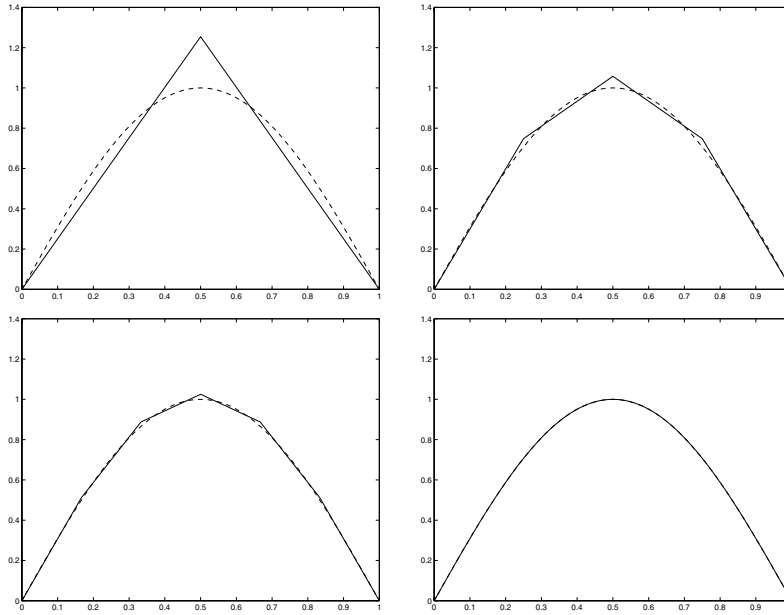

Fig. 14.3. Graph of the finite element approximation u^h to the analytical solution u of the boundary value problem (14.27) on a uniform subdivision of $[0,1]$ of spacing $h = 1/n$, with $n = 2$ (top left), $n = 4$ (top right), $n = 6$ (bottom left), and $n = 100$ (bottom right). In each of the four subfigures, the dashed curve is the graph of the analytical solution $u(x) = \sin(\pi x)$. In the last figure the approximation error is so small that u and u^h are indistinguishable.

In order to illustrate the performance of the finite element method, we consider the following example:

$$-u'' + r(x)u = f(x)\,, \qquad x \in (0,1)\,, \qquad u(0) = 0\,, \quad u(1) = 0\,. \tag{14.27}$$

If $r(x) \equiv 1$ and $f(x) = (1+\pi^2)\sin(\pi x)$, the unique solution to this problem is $u(x) = \sin(\pi x)$. Let us pretend that we do not know the analytical solution u, and solve the boundary value problem numerically, using the finite element method on a subdivision of $[0,1]$ of uniform spacing $h = 1/n$, for various values of n. The integrals $\langle f, \varphi_i \rangle$ involved in the definition of b_i in (14.17) have been approximated, on each of the elements $[x_{i-1}, x_i]$, $1 \le i \le n$, by means of the trapezium rule. The resulting approximations u^h, for $n = 2, 4, 6, 100$, are shown in Figure 14.3.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

We see from Figure 14.3 that, as the spacing h of the subdivision is reduced, the finite element solution u^h approximates the analytical solution $u(x) = \sin(\pi x)$ with increasing accuracy. Indeed, the results corresponding to $n = 2$ and $n = 4$ in Figure 14.3 indicate that as the number of intervals in the subdivision is doubled (*i.e.*, h is halved), the maximum error between $u(x)$ and $u^h(x)$ is reduced by a factor of about 4. This reduction in the error cannot be explained by Corollary 14.2 which merely implies that halving h should lead to a reduction in $\|u - u^h\|_{\mathcal{A}}$ by a factor no less than 2. If you would like to learn more about the source of the observed enhancement of accuracy, consult Exercise 5 at the end of the chapter.

14.5 *A posteriori* error analysis by duality

The bound on the error between the analytical solution u and its finite element approximation u^h formulated in Corollary 14.2 shows that, in the limit of $h \to 0$, the error $\|u - u^h\|_{\mathcal{A}}$ will tend to zero as $\mathcal{O}(h)$. This is a useful result from the theoretical point of view: it reassures us that the unknown analytical solution may be approximated arbitrarily well by making h sufficiently small. On the other hand, asymptotic error bounds of this kind are not particularly helpful for the purpose of precisely quantifying the size of the error between u and u^h for a given, *fixed*, mesh size $h > 0$: as u is unknown, it is difficult to tell just how large the right-hand side of (14.26) really is.

The aim of the present section is, therefore, to derive a computable bound on the error, and to demonstrate how such a bound may be implemented into an adaptive mesh-refinement algorithm, capable of reducing the error $u - u^h$ below a certain prescribed tolerance in an automated manner, without human intervention. The approach is based on seeking a bound on $u - u^h$ in terms of the computed solution u^h rather than in terms of norms of the unknown analytical solution u. A bound on the error in terms of u^h is referred to as an ***a posteriori*** error bound, due to the fact that it becomes *computable* only *after* the numerical solution u^h has been obtained.

In order to illuminate the key ideas while avoiding technical difficulties, we shall consider the two-point boundary value problem

$$-(p(x)u')' + q(x)u' + r(x)u = f(x), \qquad a < x < b, \tag{14.28}$$

$$u(a) = A, \quad u(b) = B, \tag{14.29}$$

where $p, q \in \mathrm{C}^1[a,b]$, $r \in \mathrm{C}[a,b]$ and $f \in \mathrm{L}^2(a,b)$. We shall assume, as

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

at the beginning of the chapter, that $p(x) \geq c_0 > 0$, $x \in [a,b]$; however, instead of supposing that $r(x) \geq 0$, we shall now demand that

$$r(x) - \frac{1}{2}q'(x) \geq c_1\,, \quad x \in [a,b]\,, \tag{14.30}$$

where c_1 is assumed to be a positive constant.[1]

Letting

$$\mathcal{A}(w,v) = \int_a^b [p(x)w'(x)v'(x) + q(x)w'(x)v(x) + r(x)w(x)v(x)]\mathrm{d}x\,,$$

the weak formulation of (14.28), (14.29) is as follows:

$$\text{find } u \in \mathrm{H}^1_{\mathrm{E}}(a,b) \text{ such that } \mathcal{A}(u,v) = \langle f, v\rangle \qquad \forall\, v \in \mathrm{H}^1_0(a,b)\,. \tag{14.31}$$

Here, the bilinear functional $\mathcal{A}(\,\cdot\,,\,\cdot\,)$ is *not* symmetric, unless $q(x) \equiv 0$: indeed, $\mathcal{A}(w,v) = \mathcal{A}(v,w)$ for all $v, w \in \mathrm{H}^1(a,b)$ if, and only if, $q \equiv 0$. Hence, in general, the boundary value problem (14.28), (14.29) cannot be assigned a Ritz principle. On the other hand, the Galerkin principle (weak formulation) (14.31) is perfectly meaningful for any choice of q.

The Galerkin finite element approximation of (14.31) is constructed by introducing a (possibly nonuniform) subdivision of the interval $[a,b]$ defined by the points

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

and considering the finite element space $S^h_{\mathrm{E}} \subset \mathrm{H}^1_{\mathrm{E}}(a,b)$ consisting of all continuous piecewise linear functions v^h on this subdivision that satisfy the boundary conditions $v^h(a) = A$ and $v^h(b) = B$. The Galerkin finite element approximation of the boundary value problem is

$$\text{find } u^h \in S^h_{\mathrm{E}} \text{ such that } \mathcal{A}(u^h, v^h) = \langle f, v^h\rangle \qquad \forall\, v^h \in S^h_0\,. \tag{14.32}$$

We let $h_i = x_i - x_{i-1}$, $i = 1, \ldots, n$, and put $h = \max_i h_i$.

We wish to derive an *a posteriori* bound on the error in the $\|\cdot\|_{\mathrm{L}^2(a,b)}$ norm; that is, our aim is to quantify the size of $\|u - u^h\|_{\mathrm{L}^2(a,b)}$ in terms of the mesh parameter h and the computed solution u^h (rather than in terms of the analytical solution u as was the case in the *a priori* error analysis developed in the previous section). For this purpose, we

[1] At the expense of slight technical complications in the subsequent discussion, the requirement that $c_1 > 0$ can be relaxed to $c_1 > -\lambda_1$, where λ_1 is the smallest (positive) eigenvalue for the Sturm–Liouville eigenvalue problem $-(p(x)w')' = \lambda w$ for $x \in (a,b)$, $w(a) = 0$, $w(b) = 0$.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

consider the auxiliary boundary value problem

$$-(p(x)z')' - (q(x)z)' + r(x)z = (u-u^h)(x)\,, \quad a<x<b\,, \tag{14.33}$$

$$z(a) = 0\,,\ z(b) = 0\,, \tag{14.34}$$

called the **dual problem** (or adjoint problem).

We begin our error analysis by noting that the definition of the dual problem and straightforward integration by parts yield (recalling that $(u-u^h)(a) = 0$, $(u-u^h)(b) = 0$)

$$\begin{aligned} \|u-u^h\|^2_{\mathrm{L}^2(a,b)} &= \langle u-u^h, u-u^h\rangle \\ &= \langle u-u^h, -(pz')' - (qz)' + rz\rangle \\ &= \mathcal{A}(u-u^h, z)\,. \end{aligned}$$

On the other hand, (14.31) and (14.32) imply the Galerkin orthogonality property

$$\mathcal{A}(u-u^h, z^h) = 0 \qquad \forall\, z^h \in S^h_0\,.$$

In particular, by choosing

$$z^h = \mathcal{I}^h z \in S^h_0\,,$$

the continuous piecewise linear interpolant of the function $z \in \mathrm{H}^1_0(a,b)$, associated with the subdivision $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$, we have that

$$\mathcal{A}(u-u^h, \mathcal{I}^h z) = 0\,.$$

Thus,

$$\begin{aligned} \|u-u^h\|^2_{\mathrm{L}^2(a,b)} &= \mathcal{A}(u-u^h, z-\mathcal{I}^h z) \\ &= \mathcal{A}(u, z-\mathcal{I}^h z) - \mathcal{A}(u^h, z-\mathcal{I}^h z) \\ &= \langle f, z-\mathcal{I}^h z\rangle - \mathcal{A}(u^h, z-\mathcal{I}^h z)\,, \end{aligned} \tag{14.35}$$

where the last transition follows from (14.31) with $v = z - \mathcal{I}^h z$.

We observe that the right-hand side no longer involves the unknown analytical solution u. Furthermore,

$$\begin{aligned} \mathcal{A}(u^h, z-\mathcal{I}^h z) &= \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} p(x)(u^h)'(x)\,(z-\mathcal{I}^h z)'(x)\,\mathrm{d}x \\ &\quad + \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} q(x)\,(u^h)'(x)\,(z-\mathcal{I}^h z)(x)\,\mathrm{d}x \\ &\quad + \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} r(x)\,u^h(x)\,(z-\mathcal{I}^h z)(x)\,\mathrm{d}x\,. \end{aligned}$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

Integrating by parts in each of the n integrals in the first sum on the right-hand side, noting that $(z - \mathcal{I}^h z)(x_i) = 0$, $i = 0, \ldots, n$, we deduce that

$$\begin{aligned}&\mathcal{A}(u^h, z - \mathcal{I}^h z)\\ &= \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left[-(p(x)(u^h)')' + q(x)(u^h)' + r(x)u^h\right](z - \mathcal{I}^h z)(x)\,\mathrm{d}x\,.\end{aligned}$$

Furthermore,

$$\langle f, z - \mathcal{I}^h z\rangle = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,(z - \mathcal{I}^h z)(x)\,\mathrm{d}x\,.$$

Substituting these two identities into (14.35), we deduce that

$$\|u - u^h\|^2_{\mathrm{L}^2(a,b)} = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} R(u^h)(x)\,(z - \mathcal{I}^h z)(x)\,\mathrm{d}x\,, \tag{14.36}$$

where, for $1 \le i \le n$, and $x \in (x_{i-1}, x_i)$,

$$R(u^h)(x) = f(x) - \left[-(p(x)(u_h)')' + q(x)(u^h)' + r(x)u^h\right].$$

The function $R(u^h)$ is called **the finite element residual**; it measures the extent to which u^h fails to satisfy the differential equation

$$-(p(x)u')' + q(x)u' + r(x)u = f(x)$$

on the union of the intervals (x_{i-1}, x_i), $i = 1, \ldots, n$. Now, applying the Cauchy–Schwarz inequality on the right-hand side of (14.36) yields

$$\|u - u^h\|^2_{\mathrm{L}^2(a,b)} \le \sum_{i=1}^{n} \|R(u^h)\|_{\mathrm{L}^2(x_{i-1},x_i)}\,\|z - \mathcal{I}^h z\|_{\mathrm{L}^2(x_{i-1},x_i)}\,.$$

Recalling from Theorem 14.7 that

$$\|z - \mathcal{I}^h z\|_{\mathrm{L}^2(x_{i-1},x_i)} \le \left(\frac{h_i}{\pi}\right)^2 \|z''\|_{\mathrm{L}^2(x_{i-1},x_i)}\,, \quad i = 1, 2, \ldots, n\,,$$

we deduce that

$$\|u - u^h\|^2_{\mathrm{L}^2(a,b)} \le \frac{1}{\pi^2} \sum_{i=1}^{n} h_i^2 \|R(u^h)\|_{\mathrm{L}^2(x_{i-1},x_i)}\,\|z''\|_{\mathrm{L}^2(x_{i-1},x_i)}\,,$$

and consequently, using the Cauchy–Schwarz inequality for finite sums,

$$\sum_{i=1}^{n} a_i b_i \le \left(\sum_{i=1}^{n} |a_i|^2\right)^{1/2} \left(\sum_{i=1}^{n} |b_i|^2\right)^{1/2}$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

with

$$a_i = h_i^2 \|R(u^h)\|_{\mathrm{L}^2(x_{i-1},x_i)} \qquad \text{and} \qquad b_i \;=\; \|z''\|_{\mathrm{L}^2(x_{i-1},x_i)}\,,$$

we find that

$$\|u-u^h\|^2_{\mathrm{L}^2(a,b)} \le \frac{1}{\pi^2}\left(\sum_{i=1}^n h_i^4 \|R(u^h)\|^2_{\mathrm{L}^2(x_{i-1},x_i)}\right)^{1/2} \|z''\|_{\mathrm{L}^2(0,1)}. \tag{14.37}$$

The rest of the discussion is aimed at eliminating $\|z''\|_{\mathrm{L}^2(a,b)}$ from the right-hand side of (14.37). The desired *a posteriori* bound on the error $\|u-u^h\|_{\mathrm{L}^2(a,b)}$ in terms of $R(u^h)$ will then follow.

Lemma 14.1 *Suppose that z is the solution of the dual problem (14.33), (14.34). Then, there exists a positive constant K, dependent only on p, q and r, such that*

$$\|z''\|_{\mathrm{L}^2(a,b)} \le K\|u-u^h\|_{\mathrm{L}^2(a,b)}\,.$$

Proof As

$$-pz'' - p'z' - qz' - q'z + rz = u - u^h\,,$$

it follows that

$$pz'' = u^h - u - (p'+q)\,z' + (r-q')\,z\,,$$

and therefore, recalling that $p(x) \ge c_0 > 0$ for $x \in [a,b]$,

$$\begin{aligned} c_0\|z''\|_{\mathrm{L}^2(a,b)} \;\le\; & \|u-u^h\|_{\mathrm{L}^2(a,b)} + \|p'+q\|_\infty \|z'\|_{\mathrm{L}^2(a,b)} \\ & + \|r-q'\|_\infty \|z\|_{\mathrm{L}^2(a,b)}\,, \end{aligned} \tag{14.38}$$

where we used the notation $\|w\|_\infty = \max_{x\in[a,b]} |w(x)|$.

We shall show that both $\|z'\|_{\mathrm{L}^2(a,b)}$ and $\|z\|_{\mathrm{L}^2(a,b)}$ can be bounded in terms of $\|u-u^h\|_{\mathrm{L}^2(a,b)}$ and then, using (14.38), we shall deduce that the same is true of $\|z''\|_{\mathrm{L}^2(a,b)}$. Let us observe that, by (14.33),

$$\langle -(pz')' - (qz)' + rz, z\rangle = \langle u-u^h, z\rangle\,. \tag{14.39}$$

Integrating by parts in the terms involving p and q and noting that $z(0)=0$ and $z(1)=0$ yields

$$\begin{aligned} \langle -(pz')' - (qz)' + rz, z\rangle &= \langle pz', z'\rangle + \langle qz, z'\rangle + \langle rz, z\rangle \\ &\ge c_0\|z'\|^2_{\mathrm{L}^2(a,b)} + \frac{1}{2}\int_a^b q(x)[z^2(x)]'\mathrm{d}x + \int_a^b r(x)[z(x)]^2\mathrm{d}x\,. \end{aligned}$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

Integrating by parts, again, in the second term on the right gives

$$\begin{aligned}\langle -(pz')' - (qz)' + rz, z\rangle &\geq c_0\|z'\|^2_{\mathrm{L}^2(a,b)} - \frac{1}{2}\int_a^b q'(x)[z^2(x)]\mathrm{d}x \\ &\quad + \int_a^b r(x)[z(x)]^2\mathrm{d}x\,.\end{aligned}$$

Hence, from (14.39),

$$c_0\|z'\|^2_{\mathrm{L}^2(a,b)} + \int_a^b \left(r(x) - \frac{1}{2}q'(x)\right)[z(x)]^2\mathrm{d}x \leq \langle u - u^h, z\rangle\,,$$

and thereby, noting (14.30) and using the Cauchy–Schwarz inequality on the right-hand side,

$$\begin{aligned}\min\{c_0, c_1\}\left(\|z'\|^2_{\mathrm{L}^2(a,b)} + \|z\|^2_{\mathrm{L}^2(a,b)}\right) &\leq \langle u - u^h, z\rangle \\ &\leq \|u - u^h\|_{\mathrm{L}^2(a,b)}\,\|z\|_{\mathrm{L}^2(a,b)}\,. \quad (14.40)\end{aligned}$$

Therefore, also

$$\min\{c_0, c_1\}\|z\|^2_{\mathrm{H}^1(a,b)} \leq \|u - u^h\|_{\mathrm{L}^2(a,b)}\,\|z\|_{\mathrm{H}^1(a,b)}\,,$$

which means that

$$\begin{aligned}\left(\|z'\|^2_{\mathrm{L}^2(a,b)} + \|z\|^2_{\mathrm{L}^2(a,b)}\right)^{1/2} &= \|z\|_{\mathrm{H}^1(a,b)} \\ &\leq \frac{1}{\min\{c_0, c_1\}}\|u - u^h\|_{\mathrm{L}^2(a,b)}\,. \quad (14.41)\end{aligned}$$

Now we substitute (14.41) into (14.38) to deduce that

$$\|z''\|_{\mathrm{L}^2(a,b)} \leq K\|u - u^h\|_{\mathrm{L}^2(a,b)}\,, \quad (14.42)$$

where

$$K = \frac{1}{c_0}\left(1 + \frac{1}{\min\{c_0, c_1\}}\left(\|p' + q\|^2_\infty + \|r - q'\|^2_\infty\right)^{1/2}\right)\,.$$

□

It is important to observe here that K involves only known quantities: the coefficients in the differential equation under consideration. Therefore K can be computed, or at least bounded above, without difficulties. On inserting (14.42) into (14.37), we arrive at our final result, the computable *a posteriori* error bound,

$$\|u - u^h\|_{\mathrm{L}^2(a,b)} \leq K_0\left(\sum_{i=1}^n h_i^4\|R(u^h)\|^2_{\mathrm{L}^2(x_{i-1},x_i)}\right)^{1/2}\,, \quad (14.43)$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

where $K_0 = K/\pi^2$.

Next we shall describe the construction of an adaptive mesh refinement algorithm based on the *a posteriori* error bound (14.43).

Suppose that TOL is a prescribed tolerance and that our aim is to compute a finite element approximation u^h to the unknown solution u so that

$$\|u - u^h\|_{\mathrm{L}^2(a,b)} \leq \mathtt{TOL}\,. \tag{14.44}$$

We shall use the *a posteriori* error bound (14.43) to achieve this goal by systematically refining the subdivision, and computing a succession of numerical solutions u^h on this sequence of subdivisions, until the inequality

$$K_0 \left(\sum_{i=1}^{n} h_i^4 \|R(u^h)\|^2_{\mathrm{L}^2(x_{i-1},x_i)} \right)^{1/2} \leq \mathtt{TOL} \tag{14.45}$$

is satisfied. Clearly, if u^h satisfies (14.45), then, by virtue of (14.43), it also satisfies (14.44).

In order for the inequality (14.45) to hold it is sufficient to ensure that, on each interval $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$, we have

$$h_i^4 \|R(u^h)\|^2_{\mathrm{L}^2(x_{i-1},x_i)} \leq \frac{1}{n} \left(\frac{\mathtt{TOL}}{K_0} \right)^2 . \tag{14.46}$$

Thus, a sufficient condition for (14.44) is that (14.46) holds for all $i = 1, 2, \ldots, n$.

The mesh adaptation algorithm, therefore, proceeds as follows:

Step 1. Choose an initial subdivision

$$\mathcal{T}_0\colon \quad a = x_0^{(0)} < x_1^{(0)} < \cdots < x_{n_0-1}^{(0)} < x_{n_0}^{(0)} = b$$

of the interval $[a, b]$, with $h_i^{(0)} = x_i^{(0)} - x_{i-1}^{(0)}$, for $i = 1, 2, \ldots, n_0$; let $h^{(0)} = \max_i h_i^{(0)}$, and consider the associated finite element space $S_{\mathrm{E}}^{h^{(0)}}$ (of dimension $n_0 - 1$);

Step 2. Compute the corresponding solution $u^{h^{(0)}} \in S_{\mathrm{E}}^{h^{(0)}}$;

Step 3. Given a computed solution $u^{h^{(m)}} \in S_{\mathrm{E}}^{h^{(m)}}$ for some $m \geq 0$, defined on a subdivision $\mathcal{T}_m$, STOP if

$$K_0 \left(\sum_{i=1}^{n_m} \left(h_i^{(m)} \right)^4 \|R(u^{h^{(m)}})\|^2_{\mathrm{L}^2(x_{i-1}^{(m)},x_i^{(m)})} \right)^{1/2} \leq \mathtt{TOL}\,; \tag{14.47}$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

`Step 4.` If not, then halve those elements $[x_{i-1}^{(m)}, x_i^{(m)}]$ in $\mathcal{T}_m$, with i in the set $\{1, 2, \ldots, n_m\}$, for which

$$\left(h_i^{(m)}\right)^4 \|R(u^{h^{(m)}})\|^2_{\mathrm{L}^2(x_{i-1}^m, x_i^m)} > \frac{1}{n_m}\left(\frac{\mathtt{TOL}}{K_0}\right)^2 , \quad (14.48)$$

denote by $\mathcal{T}_{m+1}$ the resulting subdivision of $[a, b]$ with n_{m+1} elements $[x_{i-1}^{(m+1)}, x_i^{(m+1)}]$ of respective lengths

$$h_i^{(m+1)} = x_i^{(m+1)} - x_{i-1}^{(m+1)} , \qquad i = 1, \ldots, n_{m+1} ,$$

and consider the associated finite element space $S_{\mathrm{E}}^{h^{(m+1)}}$ of dimension $n_{m+1} - 1$;

`Step 5.` Compute the finite element approximation $u^{h^{(m+1)}} \in S_{\mathrm{E}}^{h^{(m+1)}}$, increase m by 1 and return to `Step 3.`

The inequality (14.47) is called the **stopping criterion** for the mesh adaptation algorithm, and (14.48) is referred to as the **refinement criterion**. According to the *a posteriori* error bound (14.43), when the adaptive algorithm terminates, the error $\|u - u^h\|_{\mathrm{L}^2(a,b)}$ is guaranteed not to exceed the prescribed tolerance `TOL`.

We conclude the body of this chapter with a numerical experiment which illustrates the performance of the adaptive algorithm.

Example 14.1 *Let us consider the second-order ordinary differential equation*

$$-(p(x)u')' + q(x)u' + r(x)u = f(x) , \qquad x \in (0, 1) , \quad (14.49)$$

subject to the boundary conditions

$$u(0) = 0 , \qquad u(1) = 0 . \quad (14.50)$$

Suppose, for example, that

$$p(x) \equiv 1 , \qquad q(x) \equiv 20 , \qquad r(x) \equiv 10 \qquad \textit{and} \qquad f(x) \equiv 1 .$$

In this case, the analytical solution, u, can be expressed in closed form:

$$u(x) = C_1 \, \mathrm{e}^{\lambda_1 x} + C_2 \, \mathrm{e}^{\lambda_2 x} + \frac{1}{10} ,$$

where λ_1 and λ_2 are the two roots of the characteristic polynomial of the differential equation, $-\lambda^2 + 20\lambda + 10 = 0$, *i.e.*,

$$\lambda_1 = 10 + \sqrt{110} , \qquad \lambda_2 = 10 - \sqrt{110} ,$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

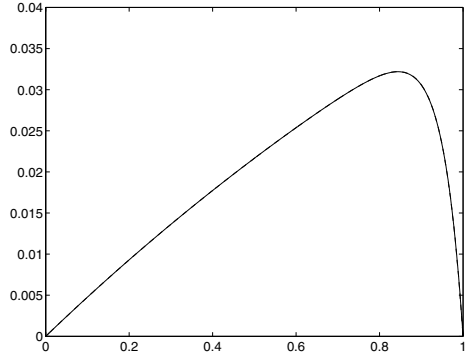

Fig. 14.4. Analytical solution of the boundary value problem (14.49), (14.50), with $p(x) \equiv 1$, $q(x) \equiv 20$, $r(x) \equiv 10$ and $f(x) \equiv 1$.

and C_1 and C_2 are constants chosen so as to ensure that $u(0) = 0$ and $u(1) = 0$; hence,

$$C_1 = \frac{\mathrm{e}^{\lambda_2} - 1}{10\,(\mathrm{e}^{\lambda_1} - \mathrm{e}^{\lambda_2})}\,, \qquad C_2 = \frac{1 - \mathrm{e}^{\lambda_1}}{10\,(\mathrm{e}^{\lambda_1} - \mathrm{e}^{\lambda_2})}\,.$$

The function u is shown in Figure 14.4.

Now, let us imagine for a moment that u is unknown, and let us compute a numerical approximation u^h to u, using the adaptive finite element algorithm described above, so that $\|u - u^h\|_{\mathrm{L}^2(0,1)} \leq \mathtt{TOL}$, where $\mathtt{TOL} = 10^{-4}$. The computation begins on a coarse subdivision of the interval $[0, 1]$ containing only 10 elements. This is then successively refined using the refinement criterion (14.48) until the stopping criterion (14.47) is satisfied; the resulting subdivisions are shown in Figure 14.5. In this example, the constant K_0 appearing in (14.43) and (14.45)–(14.48) is $(1 + \sqrt{500})/\pi^2$ (≈ 2.367).

Since we are in the fortunate (but highly idealised) position that, in addition to the numerical solution u^h, the analytical solution u is also available, we can assess the sharpness of our *a posteriori* error bound (14.43) by comparing the error $\|u - u^h\|_{\mathrm{L}^2(0,1)}$ appearing on the left-hand side of (14.43) with the computable *a posteriori* error bound on the right-hand side of (14.43). Figure 14.6 shows that the *a posteriori* bound consistently overestimates the error $\|u - u^h\|_{\mathrm{L}^2(0,1)}$ by about two orders of magnitude. By comparing the slopes of the two curves in Figure 14.6, we also see that the error and the *a posteriori* error bound decay at approximately the same rate as the number of mesh points increases in the course of mesh adaptation. ◇

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

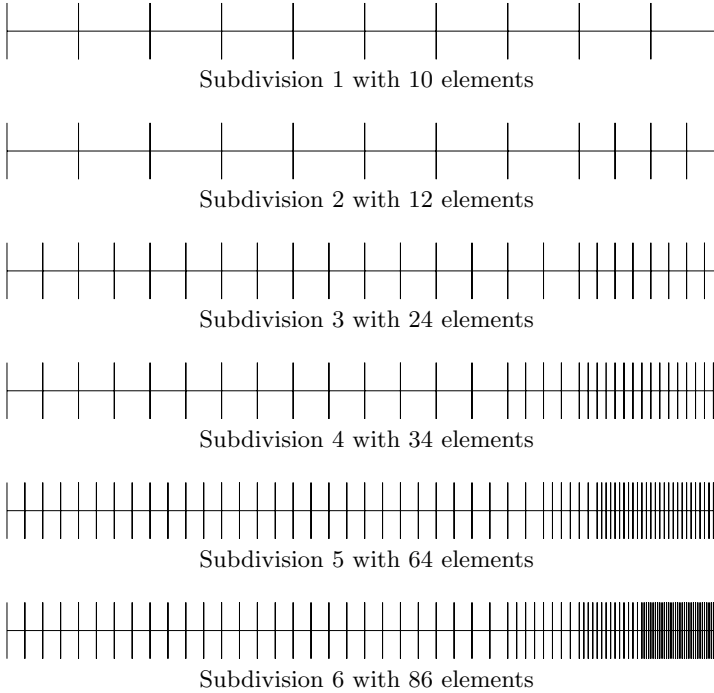

Fig. 14.5. Sequence of subdivisions of the interval $[0, 1]$ designed by the adaptive algorithm with $\texttt{TOL} = 10^{-4}$.

14.6 Notes

For further details concerning the mathematical theory and the implementation of the finite element method we refer to the following books.

- D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, 2001.
- S. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Second Edition, Springer, New York, 2002.
- C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, 1996.

For recent results on the theory of *a posteriori* error estimation for finite element approximations of differential equations, based on duality arguments, the interested reader may wish to consult the following review articles.

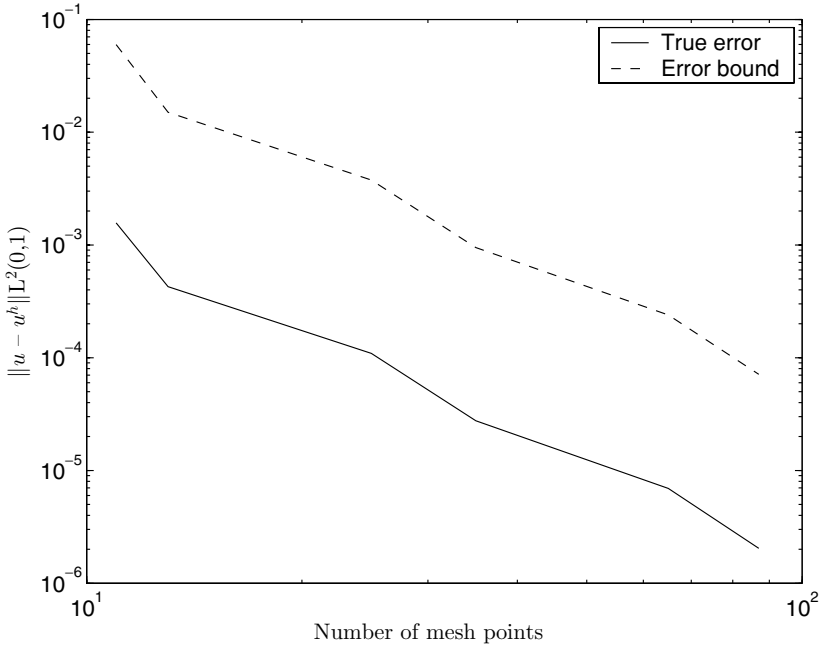

Fig. 14.6. Comparison of the true error $\|u - u^h\|_{\mathrm{L}^2(0,1)}$ (solid curve) with the *a posteriori* error bound delivered by the adaptive algorithm (dashed curve) with $\mathtt{TOL} = 10^{-4}$.

- K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, Introduction to adaptive methods for differential equations, in *Acta Numerica* **4** (A. Iserles, ed.), Cambridge University Press, Cambridge, 105–158, 1995.
- R. Becker and R. Rannacher, An optimal control approach to a-posteriori error estimation in finite element methods, in *Acta Numerica* **10** (A. Iserles, ed.), Cambridge University Press, Cambridge, 1–102, 2001.
- M.B. Giles and E. Süli, Adjoint methods for PDEs: superconvergence and adaptivity by duality, in *Acta Numerica* **11** (A. Iserles, ed.), Cambridge University Press, Cambridge, 145–236, 2002.

A detailed and general survey of the subject of *a posteriori* error estimation can be found in

- M. Ainsworth and J.T. Oden, *A posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, New York, 2000.

In this chapter we were concerned with the *a priori* error analysis of the piecewise linear finite element method in the energy norm, and its *a posteriori* error analysis in the L^2 norm. Using similar techniques, one can establish an *a priori* error bound in the L^2 norm and an *a posteriori* error bound in the energy norm. For extensions of the theory considered here to higher-order piecewise polynomial finite element approximations and generalisations to partial differential equations, the reader is referred to the books listed above.

Exercises

14.1 Given that (a, b) is an open interval of the real line, let

$$\mathrm{H}^1_{\mathrm{E}_0}(a,b) = \{v \in \mathrm{H}^1(a,b)\colon v(a) = 0\}\,.$$

(i) By writing

$$v(x) = \int_a^x v'(\xi)\mathrm{d}\xi\,,$$

for $v \in \mathrm{H}^1_{E_0}(a,b)$ and $x \in [a,b]$, show the following (**Poincaré–Friedrichs**) inequality:

$$\|v\|^2_{\mathrm{L}^2(a,b)} \leq \frac{1}{2}(b-a)^2\|v'\|^2_{\mathrm{L}^2(a,b)} \qquad \forall\, v \in \mathrm{H}^1_{\mathrm{E}_0}(a,b)\,.$$

(ii) By writing

$$[v(x)]^2 = \int_a^x \frac{\mathrm{d}}{\mathrm{d}\xi}[v(\xi)]^2\,\mathrm{d}\xi = 2\int_a^x v(\xi)v'(\xi)\mathrm{d}\xi$$

for $v \in \mathrm{H}^1_{\mathrm{E}_0}(a,b)$ and $x \in [a,b]$, show the following (**Agmon's**) inequality:

$$\max_{x\in[a,b]} |v(x)|^2 \leq 2\,\|v\|_{\mathrm{L}^2(a,b)}\|v'\|_{\mathrm{L}^2(a,b)} \quad \forall\, v \in \mathrm{H}^1_{\mathrm{E}_0}(a,b)\,.$$

14.2 Given that $f \in \mathrm{L}^2(0,1)$, state the weak formulation of each of the following boundary value problems on the interval $(0,1)$:

(a) $-u'' + u = f(x)$, $u(0) = 0$, $u(1) = 0$;
(b) $-u'' + u = f(x)$, $u(0) = 0$, $u'(1) = 1$;
(c) $-u'' + u = f(x)$, $u(0) = 0$, $u(1) + u'(1) = 2$.

In each case, show that there exists at most one weak solution.

14.3 Give a proof of Theorem 14.4.

14.4 Prove Corollary 14.2.

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

14.5 Consider the boundary value problem

$$-p_0u'' + r_0u = f(x)\,, \quad u(0) = 0\,, \;\; u(1) = 0\,,$$

on the interval $[0, 1]$, where p_0 and r_0 are positive constants and $f \in \mathrm{C}^4[0, 1]$. Using equally spaced points

$$x_i = ih\,, \qquad i = 0, 1, \ldots, n\,, \quad \text{with } h = 1/n,\;\; n \geq 2\,,$$

and the standard piecewise linear finite element basis functions (hat functions) φ_i, $i = 1, 2, \ldots, n-1$, show that the finite element equations for $u_i = u^h(x_i)$ become

$$-p_0(u_{i-1} - 2u_i + u_{i+1})/h^2 \; + r_0(u_{i-1} + 4\,u_i + u_{i+1})/6 = \frac{1}{h}\langle f, \varphi_i\rangle$$

for $i = 1, 2, \ldots, n-1$, with $u_0 = 0$ and $u_n = 0$. By expanding in Taylor series, show that

$$\frac{1}{h}\langle f, \varphi_i\rangle = f(x_i) + \tfrac{1}{12}h^2 f''(x_i) + \mathcal{O}(h^4)\,.$$

Interpreting this set of difference equations as a finite difference approximation to the boundary value problem, as in Chapter 13, show that the corresponding truncation error T_i satisfies

$$T_i = \tfrac{1}{12}h^2 r_0 u''(x_i) + \mathcal{O}(h^4)\,, \qquad i = 1, \ldots, n-1\,,$$

and use the method of Exercise 13.2 to show that

$$\max_{0 \leq i \leq n} |u(x_i) - u^h(x_i)| \leq Mh^2\,,$$

where M is a positive constant.

14.6 In the notation of Exercise 5 suppose that all the integrals involved in the calculation are approximated by the trapezium rule. Show that the system of equations becomes identical to that obtained from the central difference approximation in Chapter 13, and deduce that

$$\max_{0 \leq i \leq n} |u(x_i) - u^h(x_i)| \leq Mh^2\,,$$

where M is a positive constant.

14.7 Consider the differential equation

$$-\left(p(x)u'\right)' + r(x)u = f(x)\,, \qquad a < x < b\,,$$

with p, r and f as at the beginning of the chapter, subject to the boundary conditions

$$-p(a)u'(a) + \alpha u(a) = A\,, \qquad p(b)u'(b) + \beta u(b) = B\,,$$

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

where α and β are positive real numbers, and A and B are real numbers. Show that the weak formulation of the boundary value problem is

$$\text{find } u \in \mathrm{H}^1(a,b) \text{ such that } \mathcal{A}(u,v) = \ell(v) \text{ for all } v \in \mathrm{H}^1(a,b)\,,$$

where

$$\begin{aligned}\mathcal{A}(u,v) &= \int_a^b [p(x)u'(x)v'(x) + r(x)u(x)v(x)]\mathrm{d}x \\ &\quad + \alpha u(a)v(a) + \beta u(b)v(b)\,,\end{aligned}$$

and

$$\ell(v) = \langle f, v\rangle + Av(a) + Bv(b)\,.$$

Construct a finite element approximation of the boundary value problem based on this weak formulation using piecewise linear finite element basis functions on the subdivision

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

of the interval $[a,b]$. Show that the finite element method gives rise to a set of $n+1$ simultaneous linear equations with $n+1$ unknowns $u_i = u^h(x_i)$, $i = 0, 1, \ldots, n$. Show that this linear system has a unique solution.

Comment on the structure of the matrix $M \in \mathbb{R}^{(n+1)\times(n+1)}$ of the linear system: (a) Is M symmetric? (b) Is M positive definite? (c) Is M tridiagonal?

14.8 Given that α is a nonnegative real number, consider the differential equation

$$-u'' + u = f(x) \quad \text{for } x \in (0,1)\,,$$

subject to the boundary conditions

$$u(0) = 0\,, \qquad \alpha u(1) + u'(1) = 0\,.$$

State the weak formulation of the problem. Using continuous piecewise linear basis functions on a uniform subdivision of $[0,1]$ into elements of size $h = 1/n$, $n \geq 2$, write down the finite element approximation to this problem and show that this has a unique solution u^h. Expand u^h in terms of the standard piecewise linear finite element basis functions (hat functions) φ_i,

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

$i = 1, 2, \ldots, n$, by writing

$$u^h(x) = \sum_{i=1}^{n} U_i \varphi_i(x)$$

to obtain a system of linear equations for the vector of unknowns $(U_1, \ldots, U_n)^{\mathrm{T}}$.

Suppose that $\alpha = 0$, $f(x) \equiv 1$ and $h = 1/3$. Solve the resulting system of linear equations and compare the corresponding numerical solution $u^h(x)$ with the exact solution $u(x)$ of the boundary value problem.

14.9 Consider the differential equation

$$-(p(x)u')' + r(x)u = f(x)\,, \quad x \in (0,1)\,,$$

subject to the boundary conditions $u(0) = 0$, $u(1) = 0$, where $p(x) \geq c_0 > 0$, $r(x) \geq 0$ for all x in the closed interval $[0,1]$, with $p \in \mathrm{C}^1[0,1]$, $r \in \mathrm{C}[0,1]$ and $f \in \mathrm{L}^2(0,1)$. Given that u^h denotes the continuous piecewise linear finite element approximation to u on a uniform subdivision of $[0,1]$ into elements of size $h = 1/n$, $n \geq 2$, show that

$$\|u - u^h\|_{\mathrm{H}^1(0,1)} \leq C_1 h \|u''\|_{\mathrm{L}^2(0,1)}\,,$$

where C_1 is a positive constant that you should specify. Show further that there exists a positive constant C such that

$$\|u - u^h\|_{\mathrm{H}^1(0,1)} \leq Ch \|f\|_{\mathrm{L}^2(0,1)}\,.$$

Calculate the right-hand sides in these inequalities in the case when

$$p(x) \equiv 1\,, \qquad r(x) \equiv 0\,, \qquad f(x) \equiv 1\,,$$

for $x \in [0,1]$, and $h = 10^{-3}$.

14.10 Consider the two-point boundary value problem

$$-u'' + u = f(x)\,, \quad x \in (0,1)\,, \qquad u(0) = 0\,,\ u(1) = 0\,,$$

with $f \in \mathrm{C}^2[0,1]$. State the piecewise linear finite element approximation to this problem on a nonuniform subdivision

$$0 = x_0 < x_1 < \cdots < x_n = 1\,, \qquad n \geq 2\,,$$

with $h_i = x_i - x_{i-1}$, assuming that, for a continuous piecewise

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press

linear function v^h,

$$\int_0^1 f(x)v^h(x)\mathrm{d}x$$

has been approximated by applying the trapezium rule on each element $[x_{i-1}, x_i]$.

Verify that the following *a posteriori* bound holds for the error between u and its finite element approximation u^h:

$$\|u-u^h\|_{\mathrm{L}^2(0,1)} \leq K_0 \left(\sum_{i=1}^{n} h_i^4 \|R(u^h)\|^2_{\mathrm{L}^2(x_{i-1},x_i)} \right)^{1/2}$$
$$+ K_1 \max_{1\leq i\leq n} h_i^2 \left(\max_{x\in[x_{i-1},x_i]} |f''(x)|^2 + 4 \max_{x\in[x_{i-1},x_i]} |f'(x)| \right)^{1/2},$$

where $R(u^h) = f(x) - (-(u^h)''(x) + u^h(x))$ for $x \in (x_{i-1}, x_i)$, $i = 1, \ldots, n$, and K_0, K_1 are constants which you should specify.

How would you use this bound to compute u to within a specified tolerance `TOL`?

https://doi.org/10.1017/CBO9780511801181.015 Published online by Cambridge University Press