1

Introduction to the finite element method

This book covers the fundamentals of the finite element method in the context of numerical simulation with specific reference to the simulation of the response of structural and mechanical components to mechanical and thermal loads.

We begin with the question: what is the meaning of the term "simulation"? By its dictionary definition, simulation is the imitative representation of the functioning of one system or process by means of the functioning of another. For instance, the membrane analogy introduced by Prandtl[1] in 1903 made it possible to find the shearing stresses in bars of arbitrary cross-section, loaded by a twisting moment, through mapping the deflected shape of a thin elastic membrane. In other words, the distribution and magnitude of shearing stress in a twisted bar can be simulated by the deflected shape of an elastic membrane.

The membrane analogy exists because two unrelated phenomena can be modeled by the same partial differential equation. The physical meaning associated with the coefficients of the differential equation depends on which problem is being solved. However, the solution of one is proportional to the solution of the other: At corresponding points the shearing stress in a bar, subjected to a twisting moment, is oriented in the direction of the tangent to the contour lines of a deflected thin membrane and its magnitude is proportional to the slope of the membrane. Furthermore, the volume enclosed by the deflected membrane is proportional to the twisting moment.

In the pre-computer years the membrane analogy provided practical means for estimating shearing stresses in prismatic bars. This involved cutting the shape of the cross-section out of sheet metal or a wood panel, covering the hole with a thin elastic membrane, applying pressure to the membrane and mapping the contours of the deflected membrane. In present-day practice both problems would be formulated as mathematical problems which would then be solved by a numerical method, most likely by the finite element method.

There are many other useful analogies. For example, the same differential equations simulate the response of assemblies of mechanical components, such as linear spring-mass-viscous damper systems and assemblies of electrical components, such as capacitors, inductors and resistors. This has been exploited by the use of analogue computers. Obviously, it is much easier to build and manipulate electrical circuitry than mechanical assemblies. In present-day practice both simulation problems would be formulated as mathematical problems which would be solved by a numerical method.

At the heart of simulation of aspects of physical reality is a mathematical problem cast in a generalized form[2]. The solution of the mathematical problem is approximated by a numerical method,

1 Ludwig Prandtl 1875–1953.
2 The generalized form is also called variational form or weak form.

Finite Element Analysis: Method, Verification and Validation, Second Edition. Barna Szabó and Ivo Babuška.

Companion Website: www.wiley.com/go/szabo/finite_element_analysis

such as the finite element method, which is the subject of this book. The quantities of interest (QoI) are extracted from the approximate solution. The errors of approximation in the QoI depend on how the mathematical problem was discretized[3] and how the QoI were extracted from the numerical solution. When the errors of approximation are larger than what is considered acceptable then the discretization has to be changed either by an automated adaptive process or by action of the analyst.

Estimation and control of numerical errors are fundamentally important in numerical simulation. Consider, for example, the problem of design certification. Design rules are typically stated in the form

$$F_{\max} \leq F_{\text{all}} \tag{1.1}$$

where $F_{\max} > 0$ (resp. $F_{\text{all}} > 0$) is the maximum (resp. allowable) value of a quantity of interest, for example the first principal stress. Since in numerical simulation only an approximation to $F_{\max}$ is available, denoted by F_{num}, it is necessary to know the size of the numerical error τ:

$$|F_{\max} - F_{\text{num}}| \leq \tau F_{\max}. \tag{1.2}$$

In design and design certification the worst case scenario has to be considered, which is underestimation of $F_{\max}$, that is,

$$F_{\text{num}} = (1 - \tau)F_{\max}. \tag{1.3}$$

Therefore it has to be shown that

$$F_{\text{num}} \leq (1 - \tau)F_{\text{all}}. \tag{1.4}$$

Without a reliable estimate of the size of the numerical error it is not possible to certify design and, furthermore, numerical errors penalize design by lowering the allowable value, as indicated by eq. (1.4). Generally speaking, it is far more economical to ensure that τ is small than to accept the consequences of decreased allowable values.

We distinguish between finite element modeling and numerical simulation. As explained in greater detail in Chapter 5, finite element modeling evolved well before the theoretical basis of numerical simulation was developed. In finite element modeling a numerical problem is formulated by assembling elements from a library of finite elements that contains intuitively constructed beam, plate, shell, solid elements of various description. The numerical problem so created may not correspond to a well defined mathematical problem and therefore a solution may not even exist. For that reason it is not possible to speak of errors of approximation. Nevertheless, finite element modeling is widely practiced with success in some cases but with disappointing results in others. Such practice should be regarded as a practice of art, guided by intuition and experience, rather than a scientific activity. This is because practitioners of finite element modeling have to balance two kinds of very large errors: (a) conceptual errors in the formulation and (b) approximation errors in the numerical solution of an improperly posed mathematical problem.

In numerical simulation, on the other hand, the formulation of mathematical models is treated separately from their numerical solution. A mathematical model should be understood to be a precise statement of an idea of physical reality that permits the prediction of the occurrence, or probability of occurrence, of physical events, given certain data. The intuitive aspects of simulation are confined to the formulation of mathematical models whereas their numerical solution involves the application of well established procedures of applied mathematics. Separation of mathematical

3 The term "discretization" refers to processes by which approximating functions are defined. The most widely used discretizations will be described and illustrated by examples in this and subsequent chapters.

models from their numerical solution makes separate treatment of errors associated with the formulation of mathematical models and their numerical approximation possible. Errors associated with the formulation of mathematical models are called model form errors. Errors associated with the numerical solution of mathematical problems are called errors of approximation or errors of discretization. In the early papers and books on the finite element method no such distinction was made.

In this chapter we introduce the finite element method as a method by which the exact solution of a mathematical problem, cast in a generalized form, can be approximated. We also introduce the relevant mathematical concepts, terminology and notation in the simplest possible setting. Generalization of these concepts to two- and three-dimensional problems will be discussed in subsequent chapters.

We first consider the formulation of a second order ordinary differential equation without reference to any physical interpretation. This is to underline that once a mathematical problem was formulated, the approximation process is independent from why the mathematical problem was formulated. This important point is often missed by engineering users of legacy finite element codes because the formulation and approximation of mathematical problems is mixed in finite element libraries.

We show that the exact solution of the generalized formulation is unique. Approximation of the exact solution by the finite element method is described and various discretization strategies are explored. Efficient methods for the computation of QoIs and a posteriori error estimation are described. This chapter serves as a foundation for subsequent chapters.

We would like to assure engineering students who are not yet familiar with the concepts and notation of that branch of applied mathematics on which the finite element method is based that their investment of time and effort to master the contents of this chapter will prove to be highly rewarding.

1.1 An introductory problem

We introduce the finite element method through approximating the exact solution of the following second order ordinary differential equation

$$-(\kappa u')' + cu = f \quad \text{on the closed interval } \bar{I} = [0 \le x \le \ell] \tag{1.5}$$

with the boundary conditions

$$u(0) = u(\ell) = 0 \tag{1.6}$$

where the prime indicates differentiation with respect to x. It is assumed that $0 < \alpha \le \kappa(x) \le \beta < \infty$ where α and β are real numbers, $\kappa' < \infty$ on $\bar{I}$, $c \ge 0$ and $f = f(x)$ are defined such that the indicated operations are meaningful on I. For example, the indicated operations would not be meaningful if $(\kappa u')'$, c or f would not be finite in one or more points on the interval $0 \le x \le \ell$. The function f is called a forcing function.

We seek an approximation to u in the form:

$$u_n = \sum_{j=1}^{n} a_j \varphi_j(x), \quad \varphi_j(0) = \varphi_j(\ell) = 0 \text{ for all } j \tag{1.7}$$

where $\varphi_j(x)$ are fixed functions, called basis functions, and a_j are the coefficients of the basis functions to be determined. Note that the basis functions satisfy the zero boundary conditions.

Let us find a_j such that the integral $\mathcal{I}$ defined by

$$\mathcal{I} = \frac{1}{2}\int_0^\ell \left(\kappa(u' - u_n')^2 + c(u - u_n)^2\right)\, dx \tag{1.8}$$

is minimum. While there are other plausible criteria for selecting a_j, we will see that this criterion is fundamentally important in the finite element method. Differentiating $\mathcal{I}$ with respect to a_i and letting the derivative equal to zero, we have:

$$\frac{d\mathcal{I}}{da_i} = \int_0^\ell \left(\kappa(u' - u_n')\varphi_i' + c(u - u_n)\varphi_i\right)\, dx = 0, \quad i = 1, 2, \dots, n. \tag{1.9}$$

Using the product rule: $(\kappa u'\varphi_i)' = (\kappa u')'\varphi_i + \kappa u'\varphi_i'$ we write

$$\begin{aligned}\int_0^\ell \kappa u'\varphi_i'\, dx &= \int_0^\ell \left((\kappa u'\varphi_i)' - (\kappa u')'\varphi_i\right)\, dx \\ &= \underbrace{(\kappa u'\varphi_i)_{x=\ell}}_{=0} - \underbrace{(\kappa u'\varphi_i)_{x=0}}_{=0} - \int_0^\ell (\kappa u')'\varphi_i\, dx.\end{aligned} \tag{1.10}$$

The underbraced terms vanish on account of the boundary conditions, see eq. (1.7). On substituting this expression into eq. (1.9), we get

$$\int_0^\ell \underbrace{(-(\kappa u')' + cu)}_{=f(x)}\varphi_i\, dx - \int_0^\ell (\kappa u_n'\varphi_i' + cu_n\varphi_i)\, dx = 0$$

which will be written as

$$\int_0^\ell (\kappa u_n'\varphi_i' + cu_n\varphi_i)\, dx = \int_0^\ell f\varphi_i\, dx, \quad i = 1, 2, \dots, n. \tag{1.11}$$

We define

$$k_{ij} = \int_0^\ell \kappa\varphi_i'\varphi_j'\, dx, \quad m_{ij} = \int_0^\ell c\varphi_i\varphi_j\, dx, \quad r_i = \int_0^\ell f\varphi_i\, dx \tag{1.12}$$

and write eq. (1.11) in the following form

$$\sum_{j=1}^n (k_{ij} + m_{ij})a_j = r_i, \quad i = 1, 2, \dots, n \tag{1.13}$$

which represents n simultaneous equations in n unknowns. It is usually written in matrix form:

$$([K] + [M])\{a\} = \{r\}. \tag{1.14}$$

On solving these equations we find an approximation u_n to the exact solution u in the sense that u_n minimizes the integral $\mathcal{I}$.

Example 1.1 Let $\kappa = 1$, $c = 1$, $\ell = 2$ and

$$f = \sin(\pi x/\ell) + \sin(2\pi x/\ell).$$

With these data the exact solution of eq. (1.5) is

$$u = \frac{1}{1 + \pi^2/\ell^2}\sin(\pi x/\ell) + \frac{1}{1 + 4\pi^2/\ell^2}\sin(2\pi x/\ell).$$

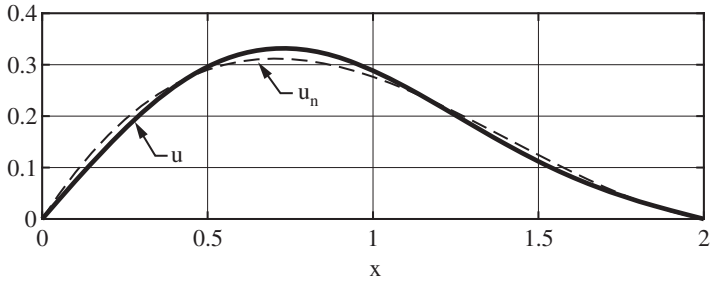

Figure 1.1 Exact and approximate solutions for the problem in Example 1.1.

We seek an approximation to u in the form:

$$u_n = u_2 = a_1 x(\ell - x) + a_2 x(\ell - x)^2.$$

On computing the elements of $[K]$, $[M]$ and $\{r\}$ we get

$$[K] = \begin{bmatrix} 2.6667 & 2.6667 \\ 2.6667 & 4.2667 \end{bmatrix} \quad [M] = \begin{bmatrix} 1.0667 & 1.0667 \\ 1.0667 & 1.2190 \end{bmatrix} \quad \{r\} = \begin{Bmatrix} 1.0320 \\ 1.4191 \end{Bmatrix}.$$

The solution of this problem is $a_1 = 0.0556$, $a_2 = 0.2209$. These coefficients, together with the basis functions, define the approximate solution u_n. The exact and approximate solutions are shown in Fig. 1.1.

The choice of basis functions

By definition, a set of functions $\varphi_j(x)$, $(j = 1, 2, \dots, n)$ are linearly independent if

$$\sum_{j=1}^{n} a_j \varphi_j(x) = 0$$

implies that $a_j = 0$ for $j = 1, 2, \dots, n$. It is left to the reader to show that if the basis functions are linearly independent then matrix $[M]$ is invertible.

Given a set of linearly independent functions $\varphi_j(x)$, $(j = 1, 2, \dots, n)$, the set of functions that can be written as

$$u_n = \sum_{j=1}^{n} a_j \varphi_j(x)$$

is called the span and $\varphi_j(x)$ are basis functions of S.

We could have defined other polynomial basis functions, for example;

$$u_n = \sum_{i=1}^{n} c_i \psi_i(x), \qquad \psi_i(x) = x^i(\ell - x). \tag{1.15}$$

When one set of basis functions $\{\varphi\} = \{\varphi_1 \ \varphi_2 \ \dots \ \varphi_n\}^T$ can be written in terms of another set $\{\psi\} = \{\psi_1 \ \psi_2 \ \dots \ \psi_n\}^T$ in the form:

$$\{\psi\} = [B]\{\varphi\} \tag{1.16}$$

where $[B]$ is an invertible matrix of constant coefficients then both sets of basis functions are said to have the same span. The following exercise demonstrates that the approximate solution depends on the span, not on the choice of basis functions.

Exercise 1.1 Solve the problem of Example 1.1 using the basis functions $\varphi_1 = x(\ell - x)$, $\varphi_2 = x^2(\ell - x)$ and show that the resulting approximate solution is identical to the approximate solution obtained in Example 1.1. The span of the basis functions in this exercise and in Example 1.1 is the same: It is the set of polynomials of degree less than or equal to 3 that vanish in the points $x = 0$ and $x = \ell$.

Summary of the main points

1. The definition of the integral $\mathcal{I}$ by eq. (1.8) made it possible to find an approximation to the exact solution u of eq. (1.5) without knowing u.
2. A formulation cannot be meaningful unless all indicated operations are defined. In the case of eq. (1.5) this means that $(\kappa u')'$ and cu are finite on the interval $0 \le x \le \ell$. In the case of eq. (1.11) the integral

$$\int_0^\ell \left(\kappa (u')^2 + cu^2\right)\, dx$$

must be finite which is a much less stringent condition. In other words, eq. (1.8) is meaningful for a larger set of functions u than eq. (1.5) is. Equation (1.5) is the strong form, whereas eq. (1.11) is the generalized or weak form of the same differential equation. When the solution of eq. (1.5) exists then u_n converges to that solution in the sense that the limit of the integral $\mathcal{I}$ is zero.
3. The error $e = u - u_n$ depends on the span and not on the choice of basis functions.

1.2 Generalized formulation

We have seen in the foregoing discussion that it is possible to approximate the exact solution u of eq. (1.5) without knowing u when $u(0) = u(\ell) = 0$. In this section the formulation is outlined for other boundary conditions.

The generalized formulation outlined in this section is the most widely implemented formulation; however, it is only one of several possible formulations. It has the properties of stability and consistency. For a discussion on the requirements of stability and consistency in numerical approximation we refer to [5].

1.2.1 The exact solution

If eq. (1.5) holds then for an arbitrary function $v = v(x)$, subject only to the restriction that all of the operations indicated in the following are properly defined, we have

$$\int_0^\ell \left((-\kappa u')' + cu - f\right) v\, dx = 0. \tag{1.17}$$

Using the product rule; $(\kappa u' v)' = (\kappa u')' v + \kappa u' v'$ we get

$$\int_0^\ell (-\kappa u')' v\, dx = -(\kappa u' v)_{x=\ell} + (\kappa u' v)_{x=0} + \int_0^\ell \kappa u' v'\, dx$$

therefore eq. (1.17) is transformed to:

$$\int_0^\ell (\kappa u' v' + cuv)\, dx = \int_0^\ell fv\, dx + (\kappa u' v)_{x=\ell} - (\kappa u' v)_{x=0}. \tag{1.18}$$

We introduce the following notation:

$$B(u,v) \overset{\text{def}}{=} \int_0^\ell (\kappa u'v' + cuv)\, dx \tag{1.19}$$

where $B(u,v)$ is a bilinear form. A bilinear form has the property that it is linear with respect to each of its two arguments. The properties of bilinear forms are listed Section A.1.3 of Appendix A. We define the linear form:

$$F(v) \overset{\text{def}}{=} \int_0^\ell fv\, dx + (\kappa u'v)_{x=\ell} - (\kappa u'v)_{x=0}. \tag{1.20}$$

The forcing function $f(x)$ may be a sum of forcing functions: $f(x) = f_1(x) + f_2(x) + \ldots$, some or all of which may be the Dirac delta function[4] multiplied by a constant. For example if $f_k(x) = F_0\delta(x_0)$ then

$$\int_0^\ell f_k(x)v\, dx = \int_0^\ell F_0\delta(x_0)v\, dx = F_0v(x_0). \tag{1.21}$$

The properties of linear forms are listed in Section A.1.2. Note that $F_0v(x_0)$ in eq. (1.21) is a linear form only if v is continuous and bounded.

The definitions of $B(u,v)$ and $F(v)$ are modified depending on the boundary conditions. Before proceeding further we need the following definitions.

1. The *energy norm* is defined by

$$\|u\|_{E(I)} \overset{\text{def}}{=} \sqrt{\frac{1}{2}B(u,u)} \tag{1.22}$$

   where I represents the open interval $I = \{x \mid 0 < x < \ell\}$. This notation should be understood to mean that $x \in I$ if and only if x satisfies the condition to the right of the bar (|). This notation may be shortened to $I = (0,\ell)$, or more generally $I = (a,b)$ where $b > a$ are real numbers. If the interval includes both boundary points then the interval is a closed interval denoted by $\bar{I} \overset{\text{def}}{=} [0,\ell]$.

   We have seen in the introductory example that the error is minimized in energy norm, that is, $\|u - u_n\|^2_{E(I)}$, equivalently $\|u - u_n\|_{E(I)}$ is minimum. The square root is introduced so that $\|\alpha u\|_{E(I)} = |\alpha|\|u\|_{E(I)}$ (where α is a constant) holds. This is one of the definitive properties of norms listed in Section A.1.1.
2. The energy space, denoted by $E(I)$, is the set of all functions u defined on I that satisfy the following condition:

$$E(I) \overset{\text{def}}{=} \{u \mid \|u\|_{E(I)} < \infty\}. \tag{1.23}$$

   Since infinitely many linearly independent functions satisfy this condition, the energy space is infinite-dimensional.
3. The *trial space*, denoted by $\tilde{E}(I)$, is a subspace of $E(I)$. When boundary conditions are prescribed on u, such as $u(0) = \hat{u}_0$ and/or $u(\ell) = \hat{u}_\ell$, then the functions that lie in $\tilde{E}(I)$ satisfy those boundary conditions. Note that when $\hat{u}_0 \neq 0$ and/or $\hat{u}_\ell \neq 0$ then $\tilde{E}(I)$ is not a linear space. This is because the condition stated under item 1 in Section A.1.1 is not satisfied. When u is prescribed on a boundary then that boundary condition is called an essential boundary condition. If no essential boundary conditions are prescribed on u then $\tilde{E}(I) = E(I)$.

4 See Definition A.5 in the appendix.

4. The *test space*, denoted by $E^0(I)$, is a subspace of $E(I)$. When boundary conditions are prescribed on u, such as $u(0) = \hat{u}_0$ and/or $u(\ell) = \hat{u}_\ell$ then the functions that lie in $E^0(I)$ are zero in those boundary points.
   If no boundary conditions are prescribed on u then $\tilde{E}(I) = E^0(I) = E(I)$. If $u(0) = \hat{u}_0$ is prescribed and $u(\ell)$ is not known then

$$\tilde{E}(I) \stackrel{\text{def}}{=} \{u \mid u \in E(I),\ u(0) = \hat{u}_0\} \tag{1.24}$$

$$E^0(I) \stackrel{\text{def}}{=} \{u \mid u \in E(I),\ u(0) = 0\}. \tag{1.25}$$

   If $u(0)$ is not known and $u(\ell) = \hat{u}_\ell$ is prescribed then

$$\tilde{E}(I) \stackrel{\text{def}}{=} \{u \mid u \in E(I),\ u(\ell) = \hat{u}_\ell\} \tag{1.26}$$

$$E^0(I) \stackrel{\text{def}}{=} \{u \mid u \in E(I),\ u(\ell) = 0\}. \tag{1.27}$$

   If $u(0) = \hat{u}_0$ and $u(\ell) = \hat{u}_\ell$ are prescribed then

$$\tilde{E}(I) \stackrel{\text{def}}{=} \{u \mid u \in E(I),\ u(0) = \hat{u}_0,\ u(\ell) = \hat{u}_\ell\} \tag{1.28}$$

$$E^0(I) \stackrel{\text{def}}{=} \{u \mid u \in E(I),\ u(0) = 0,\ u(\ell) = 0\}. \tag{1.29}$$

We are now in a position to describe the generalized formulation for various boundary conditions in a concise manner;

1. When u is prescribed on a boundary then the boundary condition is called essential or Dirichlet[5] boundary condition. Let us assume that u is prescribed on both boundary points. In this case we write $u = \overline{u} + u^\star$ where $\overline{u} \in E^0(I)$ is the function to be approximated and $u^\star \in \tilde{E}$ is an arbitrary fixed function that satisfies the boundary conditions. Substituting $\overline{u} + u^\star$ for u in eq. (1.18) we have:

$$\underbrace{\int_0^\ell (\kappa \overline{u}' v' + c\overline{u}v)\, dx}_{B(\overline{u},v)} = \underbrace{\int_0^\ell fv\, dx - \int_0^\ell (\kappa (u^\star)' v' + cu^\star v)\, dx}_{F(v)} \tag{1.30}$$

   and the generalized formulation is stated as follows: "Find $\overline{u} \in E^0(I)$ such that $B(\overline{u}, v) = F(v)$ for all $v \in E^0(I)$" where $E^0(I)$ is defined by eq. (1.29). Note that $u \in \tilde{E}(I)$ is independent of the choice of $u^\star$. Essential boundary conditions are enforced by restriction on the space of admissible functions.
2. When $\kappa u' = F$ is prescribed on a boundary then the boundary condition is called Neumann[6] boundary condition. Assume that $u(0) = \hat{u}_0$ and $(\kappa u')_{x=\ell} = F_\ell$ are prescribed. In this case

$$\underbrace{\int_0^\ell (\kappa \overline{u}' v' + c\overline{u}v)\, dx}_{B(\overline{u},v)} = \underbrace{\int_0^\ell fv\, dx + F_\ell v(\ell) - \int_0^\ell (\kappa (u^\star)' v' + cu^\star v)\, dx}_{F(v)} \tag{1.31}$$

   and the generalized formulation is: "Find $\overline{u} \in E^0(I)$ such that $B(\overline{u}, v) = F(v)$ for all $v \in E^0(I)$" where $E^0(I)$ is defined by eq. (1.25).

---

5 Peter Gustav Lejeune Dirichlet 1805–1859.
6 Carl Gottfried Neumann 1832–1925.

An important special case is when $c = 0$ and $(\kappa u')_{x=0} = F_0$ and $(\kappa u')_{x=\ell} = F_\ell$ are prescribed. In this case:

$$\underbrace{\int_0^\ell \kappa u' v' \, dx}_{B(u,v)} = \underbrace{\int_0^\ell f v \, dx - F_0 v(0) + F_\ell v(\ell)}_{F(v)} \tag{1.32}$$

and the generalized formulation is "Find $u \in E(I)$ such that $B(u, v) = F(v)$ for all $v \in E(I)$ where $E(I)$ is defined by eq. (1.23)." Since the left-hand side is zero for $v = C$ (constant) the specified data must satisfy the condition

$$\int_0^\ell f \, dx - F_0 + F_\ell = 0. \tag{1.33}$$

3. When $(\kappa u')_{x=0} = k_0(u(0) - \delta_0)$ and/or $(\kappa u')_{x=\ell} = k_\ell(\delta_\ell - u(\ell))$, where $k_0 > 0$, $k_\ell > 0$, δ_0 and δ_ℓ are given real numbers, is prescribed on a boundary then the boundary condition is called a Robin[7] boundary condition. Assume, for example, that $(\kappa u')_{x=0} = k_0(u(0) - \delta_0)$ and $(\kappa u')_{x=\ell} = F_\ell$ are prescribed. In that case

$$\underbrace{\int_0^\ell (\kappa u' v' + cuv) \, dx + k_0 u(0) v(0)}_{B(u,v)} = \underbrace{\int_0^\ell f v \, dx + F_\ell v(\ell) - k_0 \delta_0 v(0)}_{F(v)} \tag{1.34}$$

and the generalized formulation is: "Find $u \in E(I)$ such that $B(u, v) = F(v)$ for all $v \in E(I)$ where $E(I)$ is defined by eq. (1.23)."

These boundary conditions may be prescribed in any combination. The Neumann and Robin boundary conditions are called natural boundary conditions. Natural boundary conditions cannot be enforced by restriction. This is illustrated in Exercise 1.3.

The generalized formulation is stated as follows: "Find $u_{EX} \in X$ such that $B(u_{EX}, v) = F(v)$ for all $v \in Y$". The space X is called the trial space, the space Y is called the test space. We will use this notation with the understanding that the definitions of X, Y, $B(u, v)$ and $F(v)$ depend on the boundary conditions. It is essential for analysts to understand and be able to precisely state the generalized formulation for any set of boundary conditions.

Under frequently occurring special conditions the mathematical problem can be formulated on a subdomain and the solution extended to the full domain by symmetry, antisymmetry or periodicity. The symmetric, antisymmetric and periodic boundary conditions will be discussed in Chapter 2.

Theorem 1.1 The solution of the generalized formulation is unique in the energy space. The proof is by contradiction: Assume that there are two solutions u_1 and u_2 in $X \subset E(I)$ that satisfy

$$B(u_1, v) = F(v) \quad \text{for all } v \in Y$$
$$B(u_2, v) = F(v) \quad \text{for all } v \in Y.$$

Using property 1 of bilinear forms stated in the appendix, Section A.1.3, we have

$$B(u_1 - u_2, v) = 0 \quad \text{for all } v \in Y.$$

Selecting $v = u_1 - u_2$ we have $B(u_1 - u_2, u_1 - u_2) \equiv 2\|u_1 - u_2\|^2_{E(I)} = 0$. That is, $u_1 = u_2$ in energy space. Observe that when $c = 0$ and $u_1 = u_2 + C$ where C is an arbitrary constant, then $\|u_1 - u_2\|_{E(I)} = 0$.

7 Victor Gustave Robin (1855–1897).

Summary of the main points

The exact solution of the generalized formulation u_{EX} is called the generalized solution or weak solution whereas the solution that satisfies equation (1.5) is called the strong solution. The generalized formulation has the following important properties:

1. The exact solution, denoted by u_{EX}, exists for all data that satisfy the conditions $0 < \alpha \leq \kappa(x) \leq \beta < \infty$ where α and β are real numbers, $0 \leq c(x) < \infty$ and f is such that $F(v)$ satisfies the definitive properties of linear forms listed in Section A.1.2 for all $v \in E(I)$. Note that κ, c and f can be discontinuous functions.
2. The exact solution is unique in the energy space, see Theorem 1.1.
3. If the data are sufficiently smooth for the strong solution to exist then the strong and weak solutions are the same.
4. This formulation makes it possible to find approximations to u_{EX} with arbitrary accuracy. This will be addressed in detail in subsequent sections.

Exercise 1.2 Assume that $u(0) = \hat{u}_0$ and $(\kappa u')_{x=\ell} = k_\ell(\delta_\ell - u(\ell))$ are given. State the generalized formulation.

Exercise 1.3 Consider the sequence of functions $u_n(x) \in E(I)$

$$u_n(x) = \begin{cases} -x + (2\ell/n + b) & \text{for} \quad 0 \leq x \leq \ell/n \\ x + b & \text{for} \quad \ell/n < x \leq \ell \end{cases}$$

illustrated in Fig. 1.2. Show that $u_n(x)$ converges to $u(x) = x + b$ in the space $E(I)$ as $n \to \infty$. For the definition of convergence refer to Section A.2 in the appendix.

This exercise illustrates that restriction imposed on u' (or higher derivatives of u) at the boundaries will not impose a restriction on $E(I)$. Therefore natural boundary conditions cannot be enforced by restriction. Whereas all functions in $E(I)$ are continuous and bounded, the derivatives do not have to be continuous or bounded.

Exercise 1.4 Show that $F(v)$ defined on $E(I)$ by eq. (1.20) satisfies the properties of linear forms listed in Section A.1.2 if f is square integrable on I. This is a sufficient but not necessary condition for $F(v)$ to be a linear form.

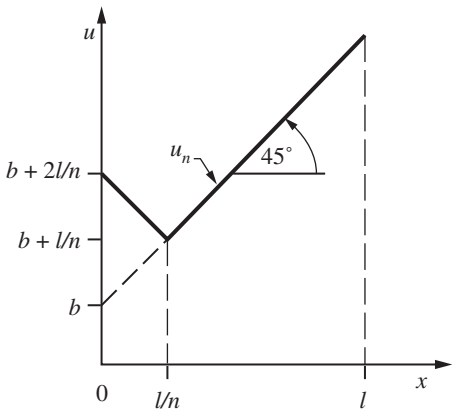

Figure 1.2 Exercise 1.3: The function $u_n(x)$.

Remark 1.1 $F(v)$ defined on $E(I)$ by eq. (1.20) satisfies the properties of linear forms listed in Section A.1.2 if the following inequality is satisfied:

$$\int_0^\ell fv\, dx < \infty \quad \text{for all } v \in E(I). \tag{1.35}$$

1.2.2 The principle of minimum potential energy

Theorem 1.2 The function $u \in \tilde{E}(I)$ that satisfies $B(u,v) = F(v)$ for all $v \in E^0(I)$ minimizes the quadratic functional[8] $\pi(u)$, called the potential energy;

$$\pi(u) \overset{\text{def}}{=} \frac{1}{2}B(u,u) - F(u) \tag{1.36}$$

on the space $\tilde{E}(I)$.

Proof: For any $v \in E^0(I)$, $\|v\|_E \neq 0$ we have:

$$\begin{aligned}\pi(u+v) &= \frac{1}{2}B(u+v,u+v) - F(u+v)\\ &= \frac{1}{2}B(u,u) + B(u,v) + \frac{1}{2}B(v,v) - F(u) - F(v)\\ &= \pi(u) + \underbrace{B(u,v) - F(v)}_{0} + \frac{1}{2}B(v,v)\end{aligned} \tag{1.37}$$

where $B(v,v) > 0$ unless $\|v\|_{E(I)} = 0$. Therefore any admissible nonzero perturbation of u will increase $\pi(u)$.

This important theorem, called the theorem or principle of minimum potential energy, will be used in Chapter 7 as our starting point in the formulation of mathematical models for beams, plates and shells.

Given the potential energy and the space of admissible functions, it is possible to determine the strong form. This is illustrated by the following example.

Example 1.2 Let us determine the strong form corresponding to the potential energy defined by

$$\pi(u) = \frac{1}{2}\int_0^\ell \left(\kappa\left(u'\right)^2 + cu^2\right)\, dx + \frac{1}{2}k_0u^2(0) - \int_0^\ell fu\, dx - k_0\delta_0u(0) \tag{1.38}$$

with $\tilde{E}(I) = \{u \mid u \in E(I),\ u(\ell) = \hat{u}_\ell\}$.

Since u minimizes $\pi(u)$, any perturbation of u by $v \in E^0(I)$ will increase $\pi(u)$. Therefore $\pi(u+\epsilon v)$ is minimum at $\epsilon = 0$ and hence

$$\left.\frac{d\pi(u+\epsilon v)}{d\epsilon}\right|_{\epsilon=0} = 0. \tag{1.39}$$

Therefore we have

$$\int_0^\ell \left(\kappa u'v' + cuv\right)\, dx - \int_0^\ell fv\, dx + \underbrace{k_0u(0)v(0) - k_0\delta_0v(0)}_{0} = 0 \tag{1.40}$$

where the last two terms are zero because $v \in E^0(I)$. Integrating the first term by parts,

$$\int_0^\ell \kappa u'v'\, dx = \underbrace{\kappa u'(\ell)v(\ell) - \kappa u'(0)v(0)}_{0} - \int_0^\ell (\kappa u')'v\, dx$$

8 A functional is a real-valued function defined on a space of functions or vectors.

and, substituting this into eq. (1.40), we get

$$\int_0^\ell \left(-(\kappa u')' + cu - f\right) v \, dx = 0. \tag{1.41}$$

Since this holds for all $v \in E^0(I)$, the bracketed expression must be zero. In other words, the solution of the differential equation

$$-(\kappa u')' + cu = f, \quad (\kappa u')_{x=0} = k_0(u(0) - \delta_0), \quad u(\ell) = \hat{u}_\ell \tag{1.42}$$

minimizes the potential energy defined by eq. (1.38). This is the strong form of the problem.

Remark 1.2 The procedure in Example 1.2 is used in the calculus of variations for identifying the differential equation, known as the Euler[9]-Lagrange[10] equation, the solution of which maximizes or minimizes a functional. In this example the solution minimizes the potential energy on the space $\tilde{E}(I)$.

Remark 1.3 Whereas the strain energy is always positive, the potential energy may be positive, negative or zero.

1.3 Approximate solutions

The trial and test spaces defined in the preceding section are infinite-dimensional, that is, they span infinitely many linearly independent functions. To find an approximate solution, we construct finite-dimensional subspaces denoted, respectively, by $S \subset X$, $V \subset Y$ and seek the function $u \in S$ that satisfies $B(u, v) = F(v)$ for all $v \in V$. Let us return to the introductory example described in Section 1.1 and define

$$u = u_n = \sum_{j=1}^{n} a_j \varphi_j, \quad v = v_n = \sum_{i=1}^{n} b_i \varphi_i$$

where φ_i $(i = 1, 2, \ldots n)$ are basis functions. Using the definitions of k_{ij} and m_{ij} given in eq. (1.12), we write the bilinear form as

$$\begin{aligned} B(u, v) \equiv \int_0^\ell (\kappa u' v' + cuv) \, dx &= \sum_{i=1}^{n}\sum_{j=1}^{n} (k_{ij} + m_{ij}) a_j b_i \\ &= \{b\}^T([K] + [M])\{a\}. \end{aligned} \tag{1.43}$$

Similarly,

$$F(v) \equiv \int_0^\ell f v \, dx = \sum_{i=1}^{n} b_i r_i = \{b\}^T\{r\} \tag{1.44}$$

where r_i is defined in eq. (1.12). Therefore we can write $B(u, v) - F(v) = 0$ in the following form:

$$\{b\}^T(([K] + [M])\{a\} - \{r\}) = 0. \tag{1.45}$$

Since this must hold for any choice of $\{b\}$, it follows that

$$([K] + [M])\{a\} = \{r\} \tag{1.46}$$

9 Leonhard Euler 1707–1783.
10 Joseph-Louis Lagrange 1736–1813.

which is the same system of linear equations we needed to solve when minimizing the integral $\mathcal{I}$, see eq. (1.14). Of course, this is not a coincidence. The solution of the generalized problem: "Find $u_n \in S$ such that $B(u_n, v) = F(v)$ for all $v \in V$", minimizes the error in the energy norm. See Theorem 1.4.

Theorem 1.3 The error e defined by $e = u - u_n$ satisfies $B(e, v) = 0$ for all $v \in S^0(I)$. This result follows directly from

$$B(u, v) = F(v) \quad \text{for all } v \in S^0(I)$$
$$B(u_n, v) = F(v) \quad \text{for all } v \in S^0(I).$$

Subtracting the second equation from the first we have,

$$B(u - u_n, v) \equiv B(e, v) = 0 \quad \text{for all } v \in S^0(I). \tag{1.47}$$

This equation is known as the Galerkin[11] orthogonality condition.

Theorem 1.4 If $u_n \in S^0(I)$ satisfies $B(u_n, v) = F(v)$ for all $v \in S^0(I)$ then u_n minimizes the error $u_{EX} - u_n$ in energy norm where u_{EX} is the exact solution:

$$\|u_{EX} - u_n\|_{E(I)} = \min_{u \in \tilde{S}} \|u_{EX} - u\|_{E(I)}. \tag{1.48}$$

Proof: Let $e = u - u_n$ and let v be an arbitrary function in $S^0(I)$. Then

$$\|e + v\|^2_{E(I)} \equiv \frac{1}{2}B(e + v, e + v) = \frac{1}{2}B(e, e) + B(e, v) + \frac{1}{2}B(v, v).$$

The first term on the right is $\|e\|^2_{E(I)}$, the second term is zero on account of Theorem 1.3, the third term is positive for any $v \neq 0$ in $S^0(I)$. Therefore $\|e\|_{E(I)}$ is minimum.

Theorem 1.4 states that the error depends on the exact solution of the problem u_{EX} and the definition of the trial space $\tilde{S}(I)$.

The finite element method is a flexible and powerful method for constructing trial spaces. The basic algorithmic structure of the finite element method is outlined in the following sections.

1.3.1 The standard polynomial space

The standard polynomial space of degree p, denoted by $S^p(I_{\text{st}})$, is spanned by the monomials $1,\ \xi,\ \xi^2,\ \ldots,\ \xi^p$ defined on the standard element

$$I_{\text{st}} = \{\xi \mid -1 < \xi < 1\}. \tag{1.49}$$

The choice of basis functions is guided by considerations of implementation, keeping the condition number of the coefficient matrices small, and personal preferences. For the symmetric positive-definite matrices considered here the condition number C is the largest eigenvalue divided by the smallest. The number of digits lost in solving a linear problem is roughly equal to $\log_{10}C$. Characterizing the condition number as being large or small should be understood in this context. In the finite element method the condition number depends on the choice of the basis functions and the mesh.

The standard polynomial basis functions, called shape functions, can be defined in various ways. We will consider shape functions based on Lagrange polynomials and Legendre[12] polynomials. We will use the same notation for both types of shape function.

11 Boris Grigoryevich Galerkin 1871–1945.
12 Adrien-Marie Legendre 1752–1833.

Lagrange shape functions

Lagrange shape functions of degree p are constructed by partitioning I_{st} into p sub-intervals. The length of the sub-intervals is typically $2/p$ but the lengths may vary. The node points are $\xi_1 = -1$, $\xi_2 = 1$ and $-1 < \xi_3 < \xi_4 < \cdots < \xi_{p+1} < 1$. The ith shape function is unity in the ith node point and is zero in the other node points:

$$N_i(\xi) = \prod_{\substack{k=1 \\ k \neq i}}^{p+1} \frac{\xi - \xi_k}{\xi_i - \xi_k}, \quad i = 1, 2, \ldots, p+1, \quad \xi \in I_{\text{st}}. \tag{1.50}$$

These shape functions have the following important properties:

$$N_i(\xi_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad \text{and} \quad \sum_{i=1}^{p+1} N_i(\xi) = 1. \tag{1.51}$$

For example, for $p = 2$ the equally spaced node points are $\xi_1 = -1$, $\xi_2 = 1$, $\xi_3 = 0$. The corresponding Lagrange shape functions are illustrated in Fig. 1.3.

Exercise 1.5 Sketch the Lagrange shape functions for $p = 3$.

Legendre shape functions

For $p = 1$ we have

$$N_1 = \frac{1 - \xi}{2}, \quad N_2 = \frac{1 + \xi}{2}. \tag{1.52}$$

For $p \geq 2$ we define the shape functions as follows:

$$N_i(\xi) = \sqrt{\frac{2i-3}{2}} \int_{-1}^{\xi} P_{i-2}(t)\, dt \quad i = 3, 4, \ldots, p+1 \tag{1.53}$$

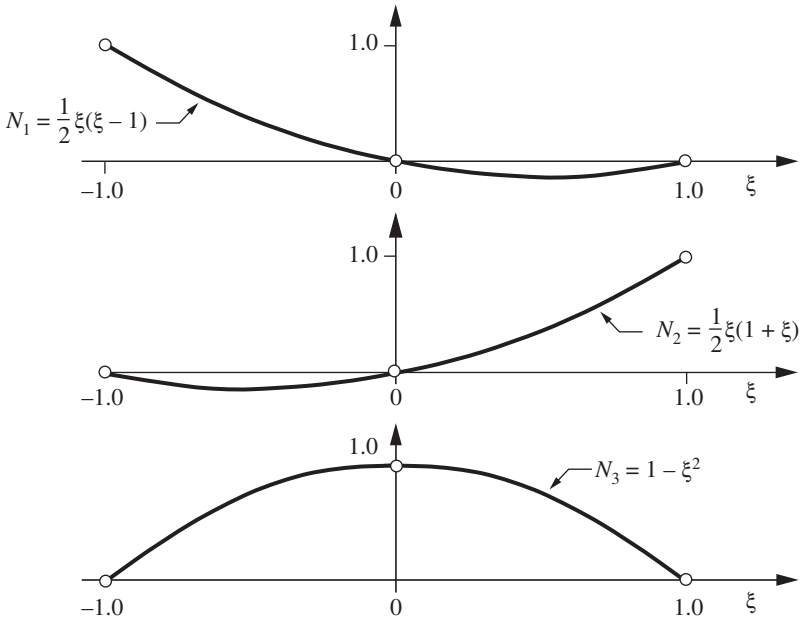

Figure 1.3 Lagrange shape functions in one dimension, $p = 2$.

where $P_i(t)$ are the Legendre polynomials. The definition of Legendre polynomials is given in Appendix D. These shape functions have the following important properties:

1. Orthogonality. For $i, j \geq 3$:

$$\int_{-1}^{+1} \frac{dN_i}{d\xi} \frac{dN_j}{d\xi}\, d\xi = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{1.54}$$

   This property follows directly from the orthogonality of Legendre polynomials, see eq. (D.13) in the appendix.
2. The set of shape functions of degree p is a subset of the set of shape functions of degree $p + 1$. Shape functions that have this property are called hierarchic shape functions.
3. These shape functions vanish at the endpoints of I_{st}: $N_i(-1) = N_i(+1) = 0$ for $i \geq 3$.

The first five hierarchic shape functions are shown in Fig. 1.4. Observe that all roots lie in I_{st}. Additional shape functions, up to $p = 8$, can be found in the appendix, Section D.1.

Exercise 1.6 Show that for the hierarchic shape functions, defined by eq. (1.53), $N_i(-1) = N_i(+1) = 0$ for $i \geq 3$.

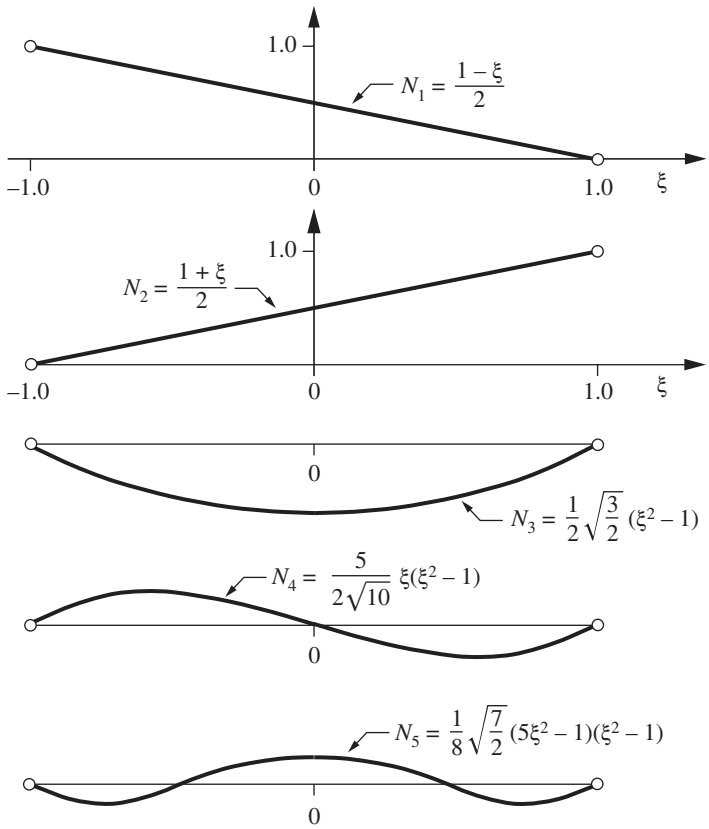

Figure 1.4 Legendre shape functions in one dimension, $p = 4$.

Exercise 1.7 Show that the hierarchic shape functions defined by eq. (1.53) can be written in the form:

$$N_i(\xi) = \frac{1}{\sqrt{2(2i-3)}}\left(P_{i-1}(\xi) - P_{i-3}(\xi)\right) \qquad i = 3, 4, \ldots \tag{1.55}$$

Hint: note that $P_n(1) = 1$ for all n and use equations (D.10) and (D.12) in Appendix D.

1.3.2 Finite element spaces in one dimension

We are now in a position to provide a precise definition of finite element spaces in one dimension.

The domain $I = \{x \mid 0 < x < \ell\}$ is partitioned into M non-overlapping intervals called finite elements. A partition, called finite element mesh, is denoted by Δ. Thus $M = M(\Delta)$. The boundary points of the elements are the node points. The coordinates of the node points, sorted in ascending order, are denoted by x_i, $(i = 1, 2, \ldots, M+1)$ where $x_1 = 0$ and $x_{M+1} = \ell$. The kth element I_k has the boundary points x_k and x_{k+1}, that is, $I_k = \{x \mid x_k < x < x_{k+1}\}$.

Various approaches are used for the construction of sequences of finite element mesh. We will consider four types of mesh design:

1. A mesh is uniform if all elements have the same size. On the interval $I = (0, \ell)$ the node points are located as follows:

$$x_k = (k-1)\ell/M(\Delta) \quad \text{for } k = 1, 2, 3, \ldots, M(\Delta)+1.$$

2. A sequence of meshes Δ_K $(K = 1, 2, \ldots)$ is quasiuniform if there exist positive constants C_1, C_2, independent of K, such that

$$C_1 \le \frac{\ell^{(K)}_{\max}}{\ell^{(K)}_{\min}} \le C_2, \quad K = 1, 2, \ldots \tag{1.56}$$

where $\ell^{(K)}_{\max}$ (resp. $\ell^{(K)}_{\min}$) is the length of the largest (resp. smallest) element in mesh Δ_K. In two and three dimensions ℓ_k is defined as the diameter of the kth element, meaning the diameter of the smallest circle or sphere that envelopes the element. For example, a sequence of quasiuniform meshes would be generated in one dimension if, starting from an arbitrary mesh, the elements would be successively halved.

3. A mesh is geometrically graded toward the point $x = 0$ on the interval $0 < x < \ell$ if the node points are located as follows:

$$x_k = \begin{cases} 0 & \text{for } k = 1 \\ q^{M(\Delta)+1-k}\ell & \text{for } k = 2, 3, \ldots, M(\Delta)+1 \end{cases} \tag{1.57}$$

where $0 < q < 1$ is called grading factor or common factor. These are called geometric meshes.

4. A mesh is a radical mesh if on the interval $0 < x < \ell$ the node points are located by

$$x_k = \left(\frac{k-1}{M(\Delta)}\right)^{\theta}\ell, \quad \theta > 1, \quad k = 1, 2, \ldots, M(\Delta)+1. \tag{1.58}$$

The question of which of these schemes is to be preferred in a particular application can be answered on the basis of a priori information concerning the regularity of the exact solution and aspects of implementation. Practical considerations that should guide the choice of the finite element mesh will be discussed in Section 1.5.2.

When the exact solution has one or more terms like $|x - x_0|^\alpha$, and $\alpha > 1/2$ is a fractional number, then the ideal mesh is a geometrically graded mesh and the polynomial degrees are assigned in such a way that the smallest elements are assigned the lowest polynomial degree, the largest elements the highest. The optimal grading factor is $q = (\sqrt{2} - 1)^2 \approx 0.17$ which is independent of α. The assigned polynomial degrees should increase at a rate of approximately 0.4 [45].

The ideal meshes are radical meshes when the same polynomial degree is assigned to each element. The optimal value of θ depends on p and α:

$$\theta = \frac{p + 1/2}{\alpha - 1/2 + (n-1)/2} \tag{1.59}$$

where n is the number of spatial dimensions. For a detailed analysis of discretization schemes in one dimension see reference [45].

The relationship between the kth element of the mesh and the standard element I_{st} is defined by the mapping function

$$x = Q_k(\xi) = \frac{1-\xi}{2}x_k + \frac{1+\xi}{2}x_{k+1}, \quad \xi \in I_{\text{st}}. \tag{1.60}$$

A finite element space S is a set of functions characterized by Δ, the assigned polynomial degrees $p_k \geq 1$ and the mapping functions $Q_k(\xi)$, $k = 1, 2, \ldots, M(\Delta)$. Specifically;

$$S = S(I, \Delta, \mathbf{p}, \mathbf{Q}) = \{u \mid u \in E(I),\ u(Q_k(\xi)) \in \mathcal{S}^{p_k}(I_{\text{st}}),\ k = 1, 2, \ldots, M(\Delta)\} \tag{1.61}$$

where $\mathbf{p}$ and $\mathbf{Q}$ represent, respectively, the arrays of the assigned polynomial degrees and the mapping functions. This should be understood to mean that $u \in S$ if and only if u satisfies the conditions on the right of the vertical bar (|). The first condition $u \in E(I)$ is that u must lie in the energy space. In one dimension this implies that u must be continuous on I. The expression $u(Q_k(\xi)) \in \mathcal{S}^{p_k}(I_{\text{st}})$ indicates that on element I_k the function $u(x)$ is mapped from the standard polynomial space $\mathcal{S}^{p_k}(I_{\text{st}})$.

The finite element test space, denoted by $S^0(I)$, is defined by the intersection $S^0(I) = S(I) \cap E(I)$, that is, $u \in S^0(I)$ is zero in those boundary points where essential boundary conditions are prescribed. The number of basis functions that span $S^0(I)$ is called the number of degrees of freedom.

The process by which the number of degrees of freedom is progressively increased by mesh refinement, with the polynomial degree fixed, is called h-extension and its implementation the h-version of the finite element method. The process by which the number of degrees of freedom is progressively increased by increasing the polynomial degree of elements, while keeping the mesh fixed, is called p-extension and its implementation the p-version of the finite element method. The process by which the number of degrees of freedom is progressively increased by concurrently refining the mesh and increasing the polynomial degrees of elements is called hp-extension and its implementation the hp-version of the finite element method.

Remark 1.4 It will be explained in Chapter 5 that the separate naming of the h, p and hp versions is related to the evolution of the finite element method rather than its theoretical foundations.

1.3.3 Computation of the coefficient matrices

The coefficient matrices are computed element by element. The numbering of the coefficients is based on the numbering of the standard shape functions, the indices range from 1 through p_{k+1}. This numbering will have to be reconciled with the requirement that each basis function must be continuous on I and must have an unique identifying number. This will be discussed separately.

Computation of the stiffness matrix

The first term of the bilinear form in eq. (1.43) is computed as a sum of integrals over the elements

$$\int_0^{\ell} \kappa(x) u_n' v_n' \, dx = \sum_{k=1}^{M(\Delta)} \int_{x_k}^{x_{k+1}} \kappa(x) u_n' v_n' \, dx. \tag{1.62}$$

We will be concerned with the evaluation of the integral on the kth element:

$$\int_{x_k}^{x_{k+1}} \kappa(x) u_n' v_n' \, dx = \int_{x_k}^{x_{k+1}} \kappa(x) \left(\sum_{j=1}^{p_k+1} a_j \frac{dN_j}{dx} \right) \left(\sum_{i=1}^{p_k+1} b_i \frac{dN_i}{dx} \right) \, dx.$$

The shape functions N_i are defined on the standard domain I_{st}. Referring to the mapping function given by eq. (1.60), we have

$$dx = \frac{x_{k+1} - x_k}{2} d\xi \equiv \frac{\ell_k}{2} \, d\xi \tag{1.63}$$

where $\ell_k \overset{\text{def}}{=} x_{k+1} - x_k$ is the length of the kth element. Also,

$$\frac{d}{dx} = \frac{d}{d\xi} \frac{d\xi}{dx} = \frac{2}{x_{k+1} - x_k} \frac{d}{d\xi} \equiv \frac{2}{\ell_k} \frac{d}{d\xi}.$$

Therefore

$$\int_{x_k}^{x_{k+1}} \kappa(x) u_n' v_n' \, dx = \frac{2}{\ell_k} \int_{-1}^{+1} \kappa(Q_k(\xi)) \left(\sum_{j=1}^{p_k+1} a_j \frac{dN_j}{d\xi} \right) \left(\sum_{i=1}^{p_k+1} b_i \frac{dN_i}{d\xi} \right) \, d\xi.$$

We define

$$k_{ij}^{(k)} = \frac{2}{\ell_k} \int_{-1}^{+1} \kappa(Q_k(\xi)) \frac{dN_i}{d\xi} \frac{dN_j}{d\xi} \, d\xi \tag{1.64}$$

and write

$$\int_{x_k}^{x_{k+1}} \kappa(x) u_n' v_n' \, dx = \sum_{i=1}^{p_k+1} \sum_{j=1}^{p_k+1} k_{ij}^{(k)} a_j b_i \equiv \{b\}^T [K^{(k)}] \{a\}. \tag{1.65}$$

The terms of the stiffness matrix $k_{ij}^{(k)}$ depend on the the mapping, the definition of the shape functions and the function $\kappa(x)$. The matrix $[K^{(k)}]$ is called the element stiffness matrix. Observe that $k_{ij}^{(k)} = k_{ji}^{(k)}$, that is, $[K^{(k)}]$ is symmetric. This follows directly from the symmetry of $B(u, v)$ and the fact that the same basis functions are used for u_n and v_n.

In the finite element method the integrals are evaluated by numerical methods. Numerical integration is discussed in Appendix E. In the important special case when $\kappa(x) = \kappa_k$ is constant on I_k, it is possible to compute $[K^{(k)}]$ once and for all. This is illustrated by the following example.

Example 1.3 When $\kappa(x) = \kappa_k$ is constant on I_k and the Legendre shape functions are used then, with the exception of the first two rows and columns, the element stiffness matrix is perfectly diagonal:

$$[K^{(k)}] = \frac{2\kappa_k}{\ell_k} \begin{bmatrix} 1/2 & -1/2 & 0 & 0 & \cdots & 0 \\ & 1/2 & 0 & 0 & & 0 \\ & & 1 & 0 & & 0 \\ & & & 1 & & 0 \\ & \text{(sym.)} & & & \ddots & \vdots \\ & & & & & 1 \end{bmatrix}. \tag{1.66}$$

Exercise 1.8 Assume that $\kappa(x) = \kappa_k$ is constant on I_k. Using the Lagrange shape functions displayed in Fig. 1.3 for $p = 2$, compute $k_{11}^{(k)}$ and $k_{13}^{(k)}$ in terms of κ_k and ℓ_k.

Computation of the Gram matrix

The second term of the bilinear form is also computed as a sum of integrals over the elements:

$$\int_0^{\ell} c(x)u_n v_n \, dx = \sum_{k=1}^{M(\Delta)} \int_{x_k}^{x_{k+1}} c(x)u_n v_n \, dx. \tag{1.67}$$

We will be concerned with evaluation of the integral

$$\begin{aligned}\int_{x_k}^{x_{k+1}} c(x)u_n v_n \, dx &= \int_{x_k}^{x_{k+1}} c(x) \left(\sum_{j=1}^{p_k+1} a_j N_j \right) \left(\sum_{i=1}^{p_k+1} b_i N_i \right) \, dx \\ &= \frac{\ell_k}{2} \int_{-1}^{+1} c(Q_k(\xi)) \left(\sum_{j=1}^{p_k+1} a_j N_j \right) \left(\sum_{i=1}^{p_k+1} b_i N_i \right) \, d\xi.\end{aligned}$$

Defining:

$$m_{ij}^{(k)} = \frac{\ell_k}{2} \int_{-1}^{1} c(Q_k(\xi)) N_i N_j \, d\xi \tag{1.68}$$

the following expression is obtained:

$$\int_{x_k}^{x_{k+1}} c(x)u_n v_n \, dx = \sum_{i=1}^{p_k+1} \sum_{j=1}^{p_k+1} m_{ij}^{(k)} a_j b_i = \{b\}^T [M^{(k)}] \{a\} \tag{1.69}$$

where $\{a\} = \{a_1 \; a_2 \; \dots \; a_{p_k+1}\}^T$, $\{b\}^T = \{b_1 \; b_2 \; \dots \; b_{p_k+1}\}$ and

$$[M^{(k)}] = \begin{bmatrix} m_{11}^{(k)} & m_{12}^{(k)} & \cdots & m_{1,p_k+1} \\ m_{21}^{(k)} & m_{22}^{(k)} & \cdots & m_{2,p_k+1} \\ \vdots & & \ddots & \vdots \\ m_{p_k+1,1}^{(k)} & m_{p_k+1,2}^{(k)} & \cdots & m_{p_k+1,p_k+1} \end{bmatrix}.$$

The terms of the coefficient matrix $m_{ij}^{(k)}$ are computable from the mapping, the definition of the shape functions and the function $c(x)$. The matrix $[M^{(k)}]$ is called the element-level Gram matrix[13] or the element-level mass matrix. Observe that $[M^{(k)}]$ is symmetric. In the important special case where $c(x) = c_k$ is constant on I_k it is possible to compute $[M^{(k)}]$ once and for all. This is illustrated by the following example.

Example 1.4 When $c(x) = c_k$ is constant on I_k and the Legendre shape functions are used then the element-level Gram matrix is strongly diagonal. For example, for $p_k = 5$ the Gram matrix is:

$$\left[M^{(k)}\right] = \frac{c_k \ell_k}{2} \begin{bmatrix} 2/3 & 1/3 & -1/\sqrt{6} & 1/3\sqrt{10} & 0 & 0 \\ & 2/3 & -1/\sqrt{6} & -1/3\sqrt{10} & 0 & 0 \\ & & 2/5 & 0 & -1/5\sqrt{21} & 0 \\ & & & 2/21 & 0 & -1/7\sqrt{45} \\ & & \text{(sym.)} & & 2/45 & 0 \\ & & & & & 2/77 \end{bmatrix} \tag{1.70}$$

13 Jörgen Pedersen Gram 1850–1916.

Remark 1.5 For $p_k \geq 2$ a simple closed form expression can be obtained for the diagonal terms and the off-diagonal terms. Using eq. (1.55) it can be shown that:

$$\begin{aligned} m_{ii}^{(k)} &= \frac{c_k \ell_k}{2} \frac{1}{2(2i-3)} \int_{-1}^{+1} (P_{i-1}(\xi) - P_{i-3}(\xi))^2 \, d\xi \\ &= \frac{c_k \ell_k}{2} \frac{2}{(2i-1)(2i-5)}, \qquad i \geq 3 \end{aligned} \tag{1.71}$$

and all off-diagonal terms are zero for $i \geq 3$, with the exceptions:

$$m_{i,i+2}^{(k)} = m_{i+2,i}^{(k)} = -\frac{c_k \ell_k}{2} \frac{1}{(2i-1)\sqrt{(2i-3)(2i+1)}}, \quad i \geq 3. \tag{1.72}$$

Remark 1.6 It has been proposed to make the Gram matrix perfectly diagonal by using Lagrange shape functions of degree p with the node points coincident with the Lobatto points. Therefore $N_i(\xi_j) = \delta_{ij}$ where δ_{ij} is the Kronecker delta[14]. Then, using $p+1$ Lobatto points, we get:

$$m_{ij}^{(k)} = \frac{c_k \ell_k}{2} \int_{-1}^{1} N_i N_j \, d\xi \approx \frac{c_k \ell_k}{2} w_i \delta_{ij}$$

where w_i is the weight of the ith Lobatto point. There is an integration error associated with this term because the integrand is a polynomial of degree $2p$. To evaluate this integral exactly $n \geq (2p+3)/2$ Lobatto points would be required (see Appendix E), whereas only $p+1$ Lobatto points are used. Throughout this book we will be concerned with errors of approximation that can be controlled by the design of mesh and the assignment of polynomial degrees. We will assume that the errors of integration and errors in mapping are negligibly small in comparison with the errors of discretization.

Exercise 1.9 Assume that $c(x) = c_k$ is constant on I_k. Using the Lagrange shape functions of degree $p = 3$, with the nodes located in the Lobatto points, compute $m_{33}^{(k)}$ numerically using 4 Lobatto points. Determine the relative error of the numerically integrated term. Refer to Remark 1.6 and Appendix E.

Exercise 1.10 Assume that $c(x) = c_k$ is constant on I_k. Using the Lagrange shape functions of degree $p = 2$, compute $m_{11}^{(k)}$ and $m_{13}^{(k)}$ in terms of c_k and ℓ_k.

1.3.4 Computation of the right hand side vector

Computation of the right hand side vector involves evaluation of the functional $F(v)$, usually by numerical means. In particular, we write:

$$F(v_n) = \int_0^{\ell} f(x) v_n \, dx = \sum_{k=1}^{M(\Delta)} \int_{x_k}^{x_{k+1}} f(x) v_n \, dx. \tag{1.73}$$

The element-level integral is computed from the definition of v_n on I_k:

$$\int_{x_k}^{x_{k+1}} f(x) v_n \, dx = \frac{\ell_k}{2} \int_{-1}^{+1} f(Q_k(\xi)) \left(\sum_{i=1}^{p_{k+1}} b_i^{(k)} N_i \right) d\xi = \sum_{i=1}^{p_{k+1}} b_i^{(k)} r_i^{(k)} \tag{1.74}$$

14 The definition of δ_{ij} is given by eq. (2.1).

where

$$r_i^{(k)} \stackrel{\text{def}}{=} \frac{\ell_k}{2} \int_{-1}^{+1} f(Q_k(\xi)) N_i(\xi)\, d\xi \tag{1.75}$$

which is computed from the given data and the shape functions.

Example 1.5 Let us assume that $f(x)$ is a linear function on I_k. In this case $f(x)$ can be written as

$$f(x) = \frac{1-\xi}{2} f(x_k) + \frac{1+\xi}{2} f(x_{k+1}) = f(x_k) N_1(\xi) + f(x_{k+1}) N_2(\xi).$$

Using the Legendre shape functions we have:

$$\begin{aligned}
r_1^{(k)} &= f(x_k) \frac{\ell_k}{2} \int_{-1}^{+1} N_1^2\, d\xi + f(x_{k+1}) \frac{\ell_k}{2} \int_{-1}^{+1} N_1 N_2\, d\xi = \frac{\ell_k}{6} \left(2f(x_k) + f(x_{k+1})\right) \\
r_2^{(k)} &= f(x_k) \frac{\ell_k}{2} \int_{-1}^{+1} N_1 N_2\, d\xi + f(x_{k+1}) \frac{\ell_k}{2} \int_{-1}^{+1} N_2^2\, d\xi = \frac{\ell_k}{6} \left(f(x_k) + 2f(x_{k+1})\right) \\
r_3^{(k)} &= f(x_k) \frac{\ell_k}{2} \int_{-1}^{+1} N_1 N_3\, d\xi + f(x_{k+1}) \frac{\ell_k}{2} \int_{-1}^{+1} N_2 N_3\, d\xi \\
&= -\frac{\ell_k}{6} \sqrt{\frac{3}{2}} \left(f(x_k) + f(x_{k+1})\right).
\end{aligned}$$

Exercise 1.11 Assume that $f(x)$ is a linear function on I_k. Using the Legendre shape functions compute $r_4^{(k)}$ and show that $r_i^{(k)} = 0$ for $i > 4$. Hint: Make use of eq. (1.55).

Exercise 1.12 Let

$$f(x) = f_k \sin \frac{x - x_k}{\ell_k} \pi, \quad x \in I_k$$

where f_k is a constant. Compute $r_5^{(k)}$ numerically in terms of f_k and ℓ_k using 3, 4 and 5 Gauss points. See Appendix E. Use the Legendre basis functions.

Exercise 1.13 Assume that $f(x)$ is a linear function on I. Using the Lagrange shape functions for $p = 2$, compute $r_1^{(k)}$.

1.3.5 Assembly

Having computed the coefficient matrices and right hand side vectors for each element, it is necessary to form the coefficient matrix and right hand side vector for the entire mesh. This process, called assembly, executes the summation in equations (1.62), (1.67) and (1.73). The local and global numbering of variables is reconciled in the assembly process. The algorithm is illustrated by the following example.

Example 1.6 Consider the three-element mesh shown in Fig. 1.5. The polynomial degrees $p_1 = 2$, $p_2 = 1$, $p_3 = 3$ are assigned to elements 1, 2, 3 respectively. The basis functions shown in Fig. 1.5 are composed of the mapped Legendre shape functions. For instance, the basis function $\varphi_2(x)$ is composed of the mapped shape function N_2 from element 1 and the mapped shape function N_1 from element 2. This basis function is zero over element 3. Basis function $\varphi_6(x)$ is the mapped shape function N_3 from element 3. This basis function is zero over elements 1 and 2.

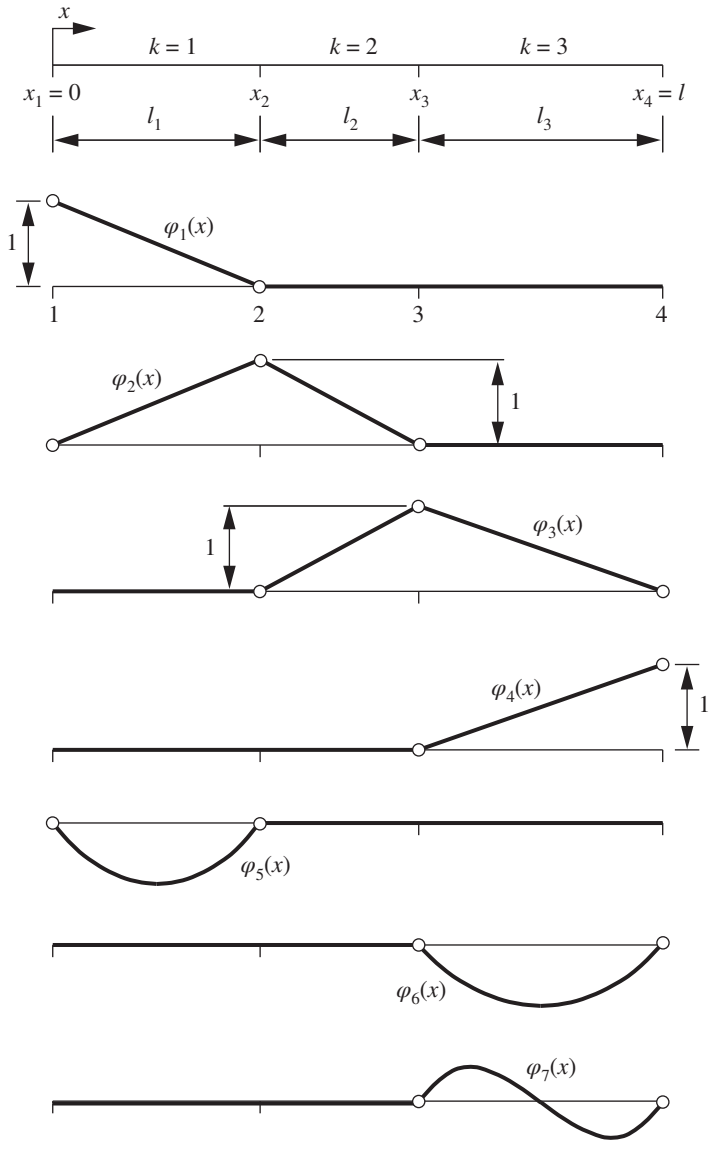

Figure 1.5 Typical finite element basis functions in one dimension.

Table 1.1 Local and global numbering in Example 1.6.

	Element number								
Numbering	**1**			**2**			**3**		
local	1	2	3	1	2	1	2	3	4
global	1	2	5	2	3	3	4	6	7

Each basis function is assigned a unique number, called a global number, and this number is associated with those element numbers and the shape function numbers from which the basis function is composed. The global and local numbers in this example are indicated in Table 1.1.

We denote $c_{ij}^{(k)} = k_{ij}^{(k)} + m_{ij}^{(k)}$ and, using equations (1.62) and (1.67), write $B(u_n, v_n)$ in the following form:

$$B(u_n, v_n) = \begin{matrix} \\ b_1 \\ b_2 \\ b_5 \end{matrix} \begin{matrix} \begin{matrix} a_1 & a_2 & a_5 \end{matrix} \\ \begin{pmatrix} c_{11}^{(1)} & c_{12}^{(1)} & c_{13}^{(1)} \\ c_{21}^{(1)} & c_{22}^{(1)} & c_{23}^{(1)} \\ c_{31}^{(1)} & c_{32}^{(1)} & c_{33}^{(1)} \end{pmatrix} \end{matrix} + \begin{matrix} \\ b_2 \\ b_3 \end{matrix} \begin{matrix} \begin{matrix} a_2 & a_3 \end{matrix} \\ \begin{pmatrix} c_{11}^{(2)} & c_{12}^{(2)} \\ c_{21}^{(2)} & c_{22}^{(2)} \end{pmatrix} \end{matrix}$$

$$+ \begin{matrix} \\ b_3 \\ b_4 \\ b_6 \\ b_7 \end{matrix} \begin{matrix} \begin{matrix} a_3 & a_4 & a_6 & a_7 \end{matrix} \\ \begin{pmatrix} c_{11}^{(3)} & c_{12}^{(3)} & c_{13}^{(3)} & c_{14}^{(3)} \\ c_{21}^{(3)} & c_{22}^{(3)} & c_{23}^{(3)} & c_{24}^{(3)} \\ c_{31}^{(3)} & c_{32}^{(3)} & c_{33}^{(3)} & c_{34}^{(3)} \\ c_{41}^{(3)} & c_{42}^{(3)} & c_{43}^{(3)} & c_{44}^{(3)} \end{pmatrix} \end{matrix}$$

where the elements within the brackets are in the local numbering system whereas the coefficients a_j and b_i outside of the brackets are in the global system. The superscripts indicate the element numbers. The terms multiplied by $a_j b_i$ are summed to obtain the elements of the assembled coefficient matrix which will be denoted by c_{ij}. For example,

$$c_{11} = c_{11}^{(1)}, \quad c_{22} = c_{22}^{(1)} + c_{11}^{(2)}, \quad c_{33} = c_{22}^{(2)} + c_{11}^{(3)}, \quad c_{77} = c_{44}^{(3)}.$$

Assuming that the boundary conditions do not include Dirichlet conditions, the bilinear form can be written in terms of the 7×7 coefficient matrix as:

$$B(u_n, v_n) = \sum_{j=1}^{7} \sum_{i=1}^{7} c_{ij} a_j b_i = \{b_1\ b_2\ \cdots\ b_7\} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{17} \\ c_{21} & c_{22} & & c_{27} \\ \vdots & & & \vdots \\ c_{71} & c_{72} & \cdots & c_{77} \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ \vdots \\ a_7 \end{Bmatrix}$$

$$\equiv \{b\}^T [C] \{a\}. \tag{1.76}$$

The treatment of Dirichlet conditions will be discussed separately in the next section.

The assembly of the right hand side vector from the element-level right hand side vectors is analogous to the procedure just described. Referring to eq. (1.73) we write $F(v_n)$ in the following form:

$$F(v_n) = \{b_1\ b_2\ b_5\} \begin{Bmatrix} r_1^{(1)} \\ r_2^{(1)} \\ r_3^{(1)} \end{Bmatrix} + \{b_2\ b_3\} \begin{Bmatrix} r_1^{(2)} \\ r_2^{(2)} \end{Bmatrix} + \{b_3\ b_4\ b_6\ b_7\} \begin{Bmatrix} r_1^{(3)} \\ r_2^{(3)} \\ r_3^{(3)} \\ r_4^{(3)} \end{Bmatrix}$$

$$= \{b_1\ b_2\ \cdots\ b_7\} \begin{Bmatrix} r_1 \\ r_2 \\ \vdots \\ r_7 \end{Bmatrix} \equiv \{b\}^T \{r\}$$

where $r_1 = r_1^{(1)}$, $r_2 = r_2^{(1)} + r_1^{(2)}$, $r_3 = r_2^{(2)} + r_1^{(3)}$, etc.

1.3.6 Condensation

Each element has $p-1$ internal basis functions. Those elements of the coefficient matrix which are associated with the internal basis functions can be eliminated at the element level. This process is called condensation.

Let us partition the coefficient matrix and right hand side vector of a finite element with $p \geq 2$ such that

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \begin{Bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{Bmatrix} = \begin{Bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{Bmatrix}$$

where the $\mathbf{a}_1 = \{a_1\ a_2\}^T$ and $\mathbf{a}_2 = \{a_3\ a_4\ \cdots\ a_{p+1}\}^T$. The coefficient matrix is symmetric therefore $\mathbf{C}_{21} = \mathbf{C}_{12}^T$. Using

$$\mathbf{a}_2 = -\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\mathbf{a}_1 + \mathbf{C}_{22}^{-1}\mathbf{r}_2 \tag{1.77}$$

we get

$$\underbrace{\left(\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\right)}_{\text{Condensed}[C]}\mathbf{a}_1 = \underbrace{\mathbf{r}_1 - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{r}_2}_{\text{Condensed}\{r\}}. \tag{1.78}$$

The condensed stiffness matrices and load vectors are assembled and the Dirichlet boundary conditions are enforced as described in the following section. Upon solving the assembled system of equations the coefficients of the internal basis functions are computed from eq. (1.77) for each element.

1.3.7 Enforcement of Dirichlet boundary conditions

When Dirichlet conditions are specified on either or both boundary points then $u \in \tilde{S}(I)$ is split into two functions; a function $\overline{u} \in S^0(I)$ and an arbitrary specific function from $\tilde{S}(I)$, denoted by $u^\star$. We then seek $\overline{u} \in S^0(I)$ such that

$$\underbrace{\int_0^\ell (\kappa \overline{u}' v' + c\overline{u}v)\, dx}_{B(\overline{u},v)} = \underbrace{\int_0^\ell fv\, dx - \int_0^\ell (\kappa (u^\star)' v' + cu^\star v)\, dx}_{F(v)} \tag{1.79}$$

for all $v \in S^0(I)$. Observe that the solution $u = \overline{u} + u^\star$ is independent of the choice of $u^\star$.

We denote the global numbers of the basis functions that are unity at $x = 0$ and $x = \ell$ by K and L respectively. For instance, in Example 1.6 $K = 1$ and $L = 4$. It is advantageous to define $u^\star$ in terms of $\varphi_K(x)$ and $\varphi_L(x)$:

$$u^\star = \hat{u}_0 \varphi_K(x) + \hat{u}_\ell \varphi_L(x) \tag{1.80}$$

as indicated in Fig. 1.6. When Dirichlet boundary condition is prescribed on only one of the boundary points then this expression is modified to include the term corresponding to that point only.

On substituting eq. (1.80) into eq. (1.79) the second term on the right-hand side of eq. (1.79) can be written as

$$\int_0^\ell (\kappa (u^\star)' v' + cu^\star v)\, dx = \sum_{i=1}^{N_u} b_i (c_{iK} + c_{iL})$$

where N_u is the number of unconstrained equations, that is, the number of equations prior to enforcement of the Dirichlet boundary conditions. (For instance, in Example 1.6 $N_u = 7$.) The coefficients c_{iK}, c_{iL} are elements of the assembled coefficient matrix.

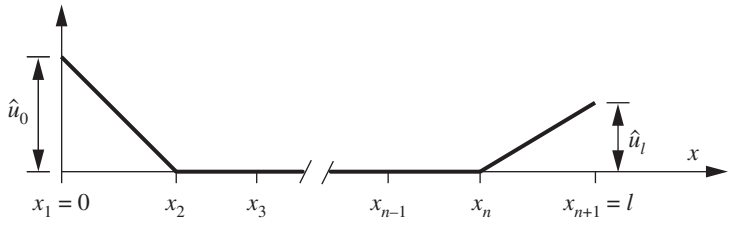

Figure 1.6 Recommended choice of the function $u^\star$ in one dimension.

Since $v \in S^0(I)$, we have $b_K = b_L = 0$ and therefore the Kth and Lth rows of matrix $[C]$ are multiplied by zero and can be deleted. The Kth and Lth columns of matrix $[C]$ are multiplied by $\hat{u}_0$ and $\hat{u}_\ell$ respectively, summed and the resulting vector is transferred to the right-hand side. The resulting coefficient matrix has the dimension N which is N_u minus the number of Dirichlet boundary conditions. The number N is called the number of degrees of freedom. It is the maximum number of linearly independent functions in $S^0(I)$.

Remark 1.7 In order to avoid having to renumber the coefficient matrix once the rows and columns corresponding to φ_K and φ_L were eliminated, all elements in the Kth and Lth rows and columns can be set to zero, with the exception of the diagonal elements, which are set to unity. The corresponding elements on the right hand side vector are set to $\hat{u}_0$ and $\hat{u}_\ell$. This is illustrated by the following example.

Example 1.7 Consider the problem

$$-u'' + 4u = 0, \quad u(0) = 1,\ u(1) = 2$$

the exact solution of which is

$$u = \frac{\exp(2) - 2}{\exp(2) - \exp(-2)} \exp(-2x) + \frac{2 - \exp(-2)}{\exp(2) - \exp(-2)} \exp(2x).$$

Using five elements of equal length on the interval $I = (0, 1)$ and $p = 1$ assigned to each element, find the finite element solution for this problem.

Referring to equations (1.66) and (1.70), the element-level coefficient matrix for each element is

$$\left[C^{(k)}\right] = \begin{bmatrix} 79/15 & -73/15 \\ -73/15 & 79/15 \end{bmatrix}, \quad k = 1, 2, \ldots 5$$

where we used $\kappa_k = 1$, $c_k = 4$, $\ell_k = 1/5$. The assembled unconstrained coefficient matrix is:

$$[C] = \begin{bmatrix} 79/15 & -73/15 & 0 & 0 & 0 & 0 \\ -73/15 & 158/15 & -73/15 & 0 & 0 & 0 \\ 0 & -73/15 & 158/15 & -73/15 & 0 & 0 \\ 0 & 0 & -73/15 & 158/15 & -73/15 & 0 \\ 0 & 0 & 0 & -73/15 & 158/15 & -73/15 \\ 0 & 0 & 0 & 0 & -73/15 & 79/15 \end{bmatrix}.$$

Upon enforcement of the Dirichlet conditions the system of equations is

$$[C] = \begin{bmatrix} 158/15 & -73/15 & 0 & 0 \\ -73/15 & 158/15 & -73/15 & 0 \\ 0 & -73/15 & 158/15 & -73/15 \\ 0 & 0 & -73/15 & 158/15 \end{bmatrix} \begin{Bmatrix} a_2 \\ a_3 \\ a_4 \\ a_5 \end{Bmatrix} = \begin{Bmatrix} 73/13 \\ 0 \\ 0 \\ 146/15 \end{Bmatrix}$$

alternatively:

$$[C] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 158/15 & -73/15 & 0 & 0 & 0 \\ 0 & -73/15 & 158/15 & -73/15 & 0 & 0 \\ 0 & 0 & -73/15 & 158/15 & -73/15 & 0 \\ 0 & 0 & 0 & -73/15 & 158/15 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 73/15 \\ 0 \\ 0 \\ 146/15 \\ 2 \end{Bmatrix}$$

where the first and sixth equations are placeholders for the boundary conditions $a_1 = 1$, $a_6 = 2$. The solution is:

$$\{a\} = \{1.0000\ 0.8784\ 0.9012\ 1.0722\ 1.4194\ 2.0000\}^T.$$

Exercise 1.14 Solve the problem in Example 1.7 with the boundary conditions $u(0) = 1$, $u'(1) = 3.6$.

Exercise 1.15 Solve the problem in Example 1.7 with the boundary conditions $u'(0) = -1$, $u(1) = 2$.

1.4 Post-solution operations

Following assembly of the coefficient matrix and enforcement of the essential boundary conditions (when applicable) the resulting system of simultaneous equations is solved by one of several methods designed to exploit the symmetry and sparsity of the coefficient matrix. The solvers are classified into two broad categories; direct and iterative solvers. Optimal choice of a solver in a particular application is based on consideration of the size of the problem and the available computational resources.

At the end of the solution process the finite element solution is available in the form

$$u_{FE} = \sum_{j=1}^{N_u} a_j \varphi_j(x) \tag{1.81}$$

where the indices reference the global numbering and N_u is the number of degrees of freedom plus the number of Dirichlet conditions.

The basis functions are decomposed into their constituent shape functions and the element-level solution records are created in the local numbering convention. Therefore the finite element solution on the kth element is available in the following form:

$$u_{FE}^{(k)} = \sum_{j=1}^{p_k+1} a_j^{(k)} N_j(\xi). \tag{1.82}$$

1.4.1 Computation of the quantities of interest

The computation of typical engineering quantities of interest (QoI) by direct and indirect methods is outlined in this section.

Computation of $u_{FE}(x_0)$

Direct computation of u_{FE} in the point $x = x_0$ involves a search to identify the element I_k in which point x_0 lies and, using the inverse of the mapping function defined by eq. (1.60), the standard

coordinate $\xi_0 \in I_{\text{st}}$ corresponding to x_0 is determined:

$$\xi_0 = Q_k^{-1}(x_0) = \frac{2x_0 - x_k - x_{k+1}}{x_{k+1} - x_k} \tag{1.83}$$

and $u_{FE}(x_0)$ is computed from

$$u_{FE}(x_0) = \sum_{j=1}^{p_k+1} a_j^{(k)} N_j(\xi_0). \tag{1.84}$$

Direct computation of $u'_{FE}(x_0)$

Direct computation of u'_{FE} in the point x_0 involves the computation of the corresponding standard coordinate $\xi_0 \in I_{\text{st}}$ using eq. (1.83) and evaluating the following expression:

$$\left(\frac{du_{FE}}{dx}\right)_{x=x_0} = \frac{2}{\ell_k}\left(\frac{du_{FE}}{d\xi}\right)_{\xi=\xi_0} = \frac{2}{\ell_k}\sum_{j=1}^{p_k+1} a_j^{(k)}\left(\frac{dN_j}{d\xi}\right)_{\xi=\xi_0} \tag{1.85}$$

where $\ell_k \overset{\text{def}}{=} x_{k+1} - x_k$. The computation of the higher derivatives is analogous.

Remark 1.8 When plotting quantities of interest such as the functions $u_{FE}(x)$ and $u'_{FE}(x)$, the data for the plotting routine are generated by subdividing the standard element into n intervals of equal length, n being the desired resolution. The QoIs corresponding to the grid-points are evaluated. This process does not involve inverse mapping. In node points information is provided from the two elements that share that node. If the computed QoI is discontinuous then the discontinuity will be visible at the nodes unless the plotting algorithm automatically averages the QoIs.

Indirect computation of $u'_{FE}(x_0)$ in node points

The first derivative in node points can be determined indirectly from the generalized formulation. For example, to compute the first derivative at node x_k from the finite element solution, we select $v = N_1(Q_k^{-1}(x))$ and use

$$\int_{x_k}^{x_{k+1}} \left(\kappa u'_{FE} v' + c u_{FE} v\right)\, dx = \int_{x_k}^{x_{k+1}} f v\, dx + \left[\kappa u'_{FE} v\right]_{x=x_{k+1}} - \left[\kappa u'_{FE} v\right]_{x=x_k}. \tag{1.86}$$

Test functions used in post-solution operations for the computation of a functional are called extraction functions. Here $v = N_1(Q_k^{-1}(x))$ is an extraction function for the functional $-\left[\kappa u'_{FE}\right]_{x=x_k}$. This is because $v(x_k) = 1$ and $v(x_{k+1}) = 0$ and hence

$$\begin{aligned} -\left[\kappa u'_{FE}\right]_{x=x_k} &= \int_{x_k}^{x_{k+1}} \left(\kappa u'_{FE} v' + c u_{FE} v\right)\, dx - \int_{x_k}^{x_{k+1}} f v\, dx \\ &= \sum_{j=1}^{p_k+1} c_{1j}^{(k)} a_j^{(k)} - r_1^{(k)} \end{aligned} \tag{1.87}$$

where, by definition; $c_{ij}^{(k)} = k_{ij}^{(k)} + m_{ij}^{(k)}$.

Example 1.8 Let us find $u'_{FE}(1)$ for the problem in Example 1.7 by the direct and indirect methods. In this case the exact solution is known from which we have $u'_{EX}(1) = 3.5978$. By direct computation:

$$u'_{FE}(1) = \frac{2}{\ell_5}\left(\frac{du_{FE}}{d\xi}\right)_{\xi=1} = 5(a_6 - a_5) = 2.9028 \quad (19.32\%\ \text{error})$$

and by indirect computation:

$$u'_{FE}(1) = -\frac{73}{15}a_5 + \frac{79}{15}a_6 = 3.6254 \quad (0.77\% \text{ error}).$$

Example 1.9 The following example illustrates that the indirect method can be used for obtaining the QoI efficiently and accurately even when the discretization was very poorly chosen. We will consider the problem

$$\int_0^\ell u'v'\,dx = \int_0^\ell \delta(x-\bar{x})v\,dx = v(\bar{x}), \quad u(0) = u(\ell) = 0$$

where δ is the delta function, see Definition A.5 in the appendix. Let us be interested in finding the approximate value of $u'(0)$. The data are $\ell = 1$ and $\bar{x} = 1/4$. We will use one finite element and $p = 2,\ 3,\ \ldots$ This is a poorly chosen discretization because the derivatives of u are discontinuous in the point $x = \bar{x}$, whereas all derivatives of the shape functions are continuous. The proper discretization would have been to use two or more finite elements with a node point in $x = \bar{x}$. Then the exact solution would be obtained at $p = 1$.

If we use the Legendre shape functions then the coefficient matrix displayed in eq. (1.66) will be perfectly diagonal. The first two rows and columns will be zero on account of the boundary conditions and the diagonal term will be 2. Referring to eq. (1.75) the right hand side vector will be

$$r_i = N_i(\bar{\xi}) \quad \text{where} \quad \bar{\xi} = Q^{-1}(\bar{x}) = -1/2.$$

Therefore the coefficients of the shape functions can be written as $a_i = r_{i+2}/2$ $(i = 1, 2, \ldots, p-1)$ where the variables are renumbered through shifting the indices to account for the boundary conditions: $a_1 = a_2 = 0$. Hence

$$u_{FE} = \frac{1}{2}\sum_{i=1}^{p-1} N_{i+2}(\bar{\xi})N_{i+2}(\xi)$$

and the QoI is:

$$u'_{FE}(0) = \sum_{i=1}^{p-1} N_{i+2}(\bar{\xi})\frac{dN_{i+2}}{d\xi}|_{\xi=-1}.$$

From the definition of N_i in eq. (1.53) we have

$$\left.\frac{dN_{i+2}}{d\xi}\right|_{\xi=-1} = \sqrt{\frac{2i+1}{2}}P_i(-1) = \sqrt{\frac{2i+1}{2}}(-1)^i$$

and the QoI can be written as

$$u'_{FE}(0) = \sum_{i=1}^{p-1} N_{i+2}(\bar{\xi})\sqrt{\frac{2i+1}{2}}(-1)^i = \frac{1}{2}\sum_{i=1}^{p-1}(-1)^i(P_{i+1}(\bar{\xi}) - P_{i-1}(\bar{\xi}))$$

where we made use of eq. (1.55). The relationships between the polynomial degree ranging from 2 to 100 and the corresponding values of the QoI computed by the direct method are displayed in Fig. 1.7. It is seen that convergence to the exact value $u'_{EX}(0) = 0.75$ is very slow.

The indirect method is based on eq. (1.18) which, applied to this example, takes the form

$$\int_0^1 u'v'\,dx = \int_0^1 \delta(\bar{x})v\,dx + (u'v)_{x=1} - (u'v)_{x=0}.$$

Selecting $v = 1 - x$ and rearranging the terms we get

$$u'(0) = v(\bar{x}) + \int_0^1 u'\,dx = v(\bar{x}) = 0.75$$

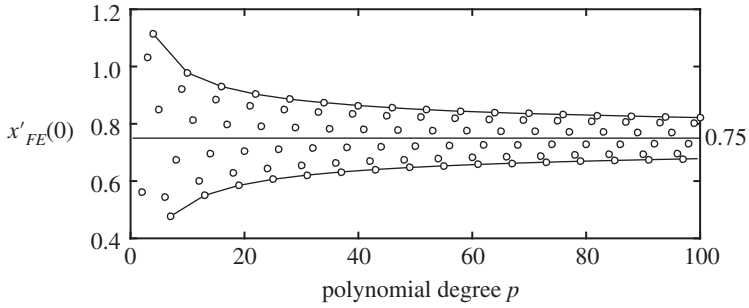

Figure 1.7 Example 1.9. Values of $u'_{FE}(0)$ computed by the direct method.

which is the exact solution. The choice $v = 1 - x$ was exceptionally fortuitous because it happens to be the Green's function (also known as the influence function) for $u'(0)$. Therefore the extracted value is independent of the solution $u \in E^0(I)$.

Let us choose $v = 1 - x^2$ for the extraction function. In this case

$$u'(0) = v(\overline{x}) - \int_0^1 u'v' \, dx = \frac{15}{16} + 2\int_0^1 u'x \, dx.$$

Substituting u'_{FE} for u':

$$\int_0^1 u'_{FE}\, x \, dx = \sum_{i=1}^{p-1} \frac{N_{i+2}(\overline{\xi})}{2}\sqrt{\frac{2i+1}{2}} \int_{-1}^1 P_i(\xi)\frac{1+\xi}{2}\, d\xi$$
$$= \frac{1}{4}\sum_{i=1}^{p-1} N_{i+2}(\overline{\xi})\sqrt{\frac{2i+1}{2}} \int_{-1}^1 P_i(\xi)(P_0(\xi) + P_1(\xi))\, d\xi = -\frac{3}{32}.$$

Taking the orthogonality of the Legendre polynomials (see eq. (D.13)) into account, the sum has to be evaluated only for $p = 2$. The extracted value of $u'_{FE}(0)$ for $p \geq 2$ is $u'_{FE}(0) = 0.5156$ (31.25% error).

An explanation of why the extraction method is much more efficient than direct computation is given in Section 1.5.4.

Exercise 1.16 Find $u'_{FE}(0)$ for the problem in Example 1.7 by the direct and indirect methods. Compute the relative errors.

Exercise 1.17 For the problem in Example 1.9 let $v = 1 - x^3$ be the extraction function. Calculate the extracted value of $u'_{FE}(0)$ for $p \geq 3$.

Nodal forces

The vector of nodal forces associated with element k, denoted by $\{f^{(k)}\}$, is defined as follows:

$$\{f^{(k)}\} = [K^{(k)}]\{a^{(k)}\} - \{\overline{r}^{(k)}\} \qquad k = 1, 2, \ldots, M(\Delta) \tag{1.88}$$

where $[K^{(k)}]$ is the stiffness matrix, $\{a^{(k)}\}$ is the solution vector and $\{\overline{r}^{(k)}\}$ is the load vector corresponding to traction forces, concentrated forces and thermal loads acting on element k.

The sign convention for nodal forces is different from the sign convention for the bar force: Whereas the bar force is positive when tensile, a nodal force is positive when acting in the direction of the positive coordinate axis.

Exercise 1.18 Assume that hierarchic basis functions based on Legendre polynomials are used. Show that when κ is constant and $c = 0$ on I_k then

$$f_1^{(k)} + f_2^{(k)} = r_1^{(k)} + r_2^{(k)}$$

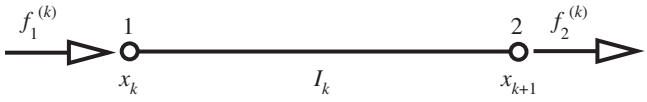

Figure 1.8 Exercise 1.18. Notation.

independently of the polynomial degree p_k. For sign convention refer to Fig. 1.8. Consider both thermal and traction loads. This exercise demonstrates that nodal forces are in equilibrium independently of the finite element solution. Therefore equilibrium of nodal forces is not an indicator of the quality of finite element solutions.

1.5 Estimation of error in energy norm

We have seen that the finite element solution minimizes the error in energy norm in the sense of eq. (1.48). It is natural therefore to use the energy norm as a measure of the error of approximation. There are two types of error estimators: (a) A priori estimators that establish the asymptotic rate of convergence of a discretization scheme, given information about the regularity (smoothness) of the exact solution and (b) a posteriori estimators that provide estimates of the error in energy norm for the finite element solution of a particular problem.

There is a very substantial body of work in the mathematical literature on the a priori estimation of the rate of convergence, given a quantitative measure of the regularity of the exact solution and a sequence of discretizations. The underlying theory is outside of the scope of this book; however, understanding the main results is important for practitioners of finite element analysis. For details we refer to [28, 45, 70, 84].

1.5.1 Regularity

Let us consider problems the exact solution of which has the functional form

$$u_{EX} = x^{\alpha}\varphi(x), \quad \alpha > 1/2, \quad x \in I = (0, \ell) \tag{1.89}$$

where $\varphi(x)$ is an analytic or piecewise analytic function, see Definition A.1 in the appendix. Our motivation for considering functions in this form is that this family of functions models the singular behavior of solutions of linear elliptic boundary value problems near vertices in polygonal and polyhedral domains. For u_{EX} to be in the energy space, its first derivative must be square integrable on I. Therefore

$$\int_0^{\ell} x^{2(\alpha-1)}\, dx > 0$$

from which it follows that α must be greater than $1/2$.

In the following we will see that when α is not an integer then the degree of difficulty associated with approximating u_{EX} by the finite element method is related to the size of $(\alpha - 1/2) > 0$. The smaller $(\alpha - 1/2)$ is, the more difficult it is to approximate u_{EX}.

If α is a fractional number then the measure of regularity used in the mathematical literature is the maximum number of square integrable derivatives, with the notion of derivative generalized to fractional numbers. See sections A.2.3 and A.2.4 in the appendix. For our purposes it is sufficient to remember that if u_{EX} has the functional form of eq. (1.89), and α is not an integer, then u_{EX} lies in the Sobolev space $H^{\alpha+1/2-\epsilon}(I)$ where $\epsilon > 0$ is arbitrarily small. This means that α must be larger than $1/2$ for the first derivative of u_{EX} to be square integrable. See, for example, [59].

If α is an integer then u_{EX} is an analytic or piecewise analytic function and the measure of regularity is the size of the derivatives of u_{EX}. Analogous definitions apply to two and three dimensions.

Remark 1.9 The kth derivative of a function $f(x)$ is a local property of $f(x)$ only when k is an integer. This is not the case for non-integer derivatives.

1.5.2 A priori estimation of the rate of convergence

Analysts are called upon to choose discretization schemes for particular problems. A sound choice of discretization is based on a priori information on the regularity of the exact solution. If we know that the exact solution lies in Sobolev space $H^k(I)$ then it is possible to say how fast the error in energy norm will approach zero as the number of degrees of freedom is increased, given a scheme by which a sequence of discretizations is generated. Index k can be inferred or estimated from the input data κ, c and f.

We define

$$h = \max_j \ell_j/\ell, \quad j = 1, 2, \dots M(\Delta) \tag{1.90}$$

where ℓ_j is the length of the jth element, ℓ is the size of the of the solution domain $I = (1, \ell)$. This is generalized to two and three dimensions where ℓ is the diameter of the domain and ℓ_j is the diameter of the jth element. In this context diameter means the diameter of the smallest circlein one and two dimensions, or sphere in three dimensions,that contains the element or domain. In two and three dimensions the solution domain is denoted by Ω.

The a priori estimate of the relative error in energy norm for $u_{EX} \in H^k(\Omega)$, quasiuniform meshes and polynomial degree p is

$$(e_r)_E \overset{\text{def}}{=} \frac{\|u_{EX} - u_{FE}\|_{E(\Omega)}}{\|u_{EX}\|_{E(\Omega)}} \leq \begin{cases} C(k)\dfrac{h^{k-1}}{p^{k-1}}\|u_{EX}\|_{H^k(\Omega)} & \text{for } k-1 \leq p \\ C(k)\dfrac{h^{p}}{p^{k-1}}\|u_{EX}\|_{H^{p+1}(\Omega)} & \text{for } k-1 > p \end{cases} \tag{1.91}$$

where $E(\Omega)$ is the energy norm, k is typically a fractional number and $C(k)$ is a positive constant that depends on k but not on h or p. This inequality gives the upper bound for the asymptotic rate of convergence of the relative error in energy norm as $h \to 0$ or $p \to \infty$ [22]. This estimate holds for one, two and three dimensions. For one and two dimensions lower bounds were proven in [13, 24] and [46] and it was shown that when singularities are located in vertex points then the rate of convergence of the p-version is twice the rate of convergence of the h-version when both are expressed in terms of the number of degrees of freedom. It is reasonable to assume that analogous results can be proven for three dimensions; however, no proofs are available at present.

We will find it convenient to write the relative error in energy norm in the following form

$$(e_r)_E \leq \frac{C}{N^\beta} \tag{1.92}$$

where N is the number of degrees of freedom and C and β are positive constants, β is called the algebraic rate of convergence. In one dimension $N \propto 1/h$ for the h-version and $N \propto p$ for the p-version. Therefore for $k - 1 < p$ we have $\beta = k - 1$. However, for the important special case when the solution has the functional form of eq. (1.89) or, more generally, has a term like $u = |x - x_0|^\lambda$ and $x_0 \in \bar{I}$ is a nodal point then $\beta = 2(k - 1)$ for the p-version: The rate of p-convergence is twice that of h-convergence [22, 84].

When the exact solution is an analytic function then $u_{EX} \in H^{\infty}(\Omega)$ and the asymptotic rate of convergence is exponential:

$$(e_r)_E \leq \frac{C}{\exp(\gamma N^{\theta})} \tag{1.93}$$

where C, γ and θ are positive constants, independent of N. In one dimension $\theta \geq 1/2$, in two dimensions $\theta \geq 1/3$, in three dimensions $\theta \geq 1/5$, see [10].

When the exact solution is a piecewise analytic function then eq. (1.93) still holds provided that the boundary points of analytic functions are nodal points, or more generally, lie on the boundaries of finite elements.

The relationship between the error $e = u_{EX} - u_{FE}$ measured in energy norm and the error in potential energy is established by the following theorem.

Theorem 1.5

$$\|e\|_E^2 = \|u_{EX} - u_{FE}\|_{E(I)}^2 = \pi(u_{FE}) - \pi(u_{EX}). \tag{1.94}$$

Proof: Writing $e = u_{EX} - u_{FE}$ and noting that $e \in E^0(I)$, from the definition of $\pi(u_{FE})$ we have:

$$\begin{aligned}\pi(u_{FE}) =& \pi(u_{EX} - e) = \frac{1}{2}B(u_{EX} - e, u_{EX} - e) - F(u_{EX} - e)\\ =& \frac{1}{2}B(u_{EX}, u_{EX}) - F(u_{EX}) \underbrace{- B(u_{EX}, e) + F(e)}_{0} + \frac{1}{2}B(e, e)\\ =& \pi(u_{EX}) + \|e\|_{E(I)}^2.\end{aligned}$$

Remark 1.10 Consider the problem given by eq. (1.5) and assume that κ and c are constants. In this case the smoothness of u depends only on the smoothness of f: If $f \in C^k(I)$ then $u \in C^{k+2}(I)$ for any $k \geq 0$. Similarly, if $f \in H^k(I)$ then $u \in H^{k+2}(I)$ for any $k \geq 0$. This is known as the shift theorem. More generally, the smoothness of u depends on the smoothness of κ, c and F. For a precise statement and proof of the shift theorem we refer to [21].

Remark 1.11 An introductory discussion on how a priori estimates are obtained under the assumption that the second derivative of the exact solution is bounded can be found in Appendix B.

1.5.3 A posteriori estimation of error

The goal of finite element computations is to estimate certain quantities of interest (QoIs) such as, for example, the maximum and minimum values of u or u' on $I = (0, \ell)$. Since finite element solutions are approximations to an exact solution, it is not sufficient to report the value of a QoI computed from the finite element solution. It is also necessary to provide an estimate of the relative error in the QoI, or present evidence that the relative error in the QoI is not greater than an acceptable value.

In this section we will use the a priori estimates described in Section 1.5.2 to obtain a posteriori estimates of error in energy norm. It is possible to obtain very accurate estimates for a large class of problems which includes most problems of practical interest.

Error estimation based on extrapolation

For most practical problems the estimate (1.92) is sufficiently sharp so that the less than or equal sign ($\leq$) can be replaced by the approximately equal sign ($\approx$) and this a priori estimate can be used in an a posteriori fashion.

The computed values of the potential energy corresponding to a sequence of finite element spaces $S_1 \subset S_2 \subset \cdots S_n$ can be used for estimating the error in energy norm by extrapolation. Sequences of finite element spaces that have this property are called hierarchic sequences. By Theorem 1.5 and eq. (1.92) we have:

$$\pi(u_{FE}) - \pi(u_{EX}) \approx \frac{C^2}{N^{2\beta}} \tag{1.95}$$

where $C \stackrel{\text{def}}{=} C\|u_{EX}\|_{E(I)}$. There are three unknowns: $\pi(u_{EX})$, C and β. Assume that we have a sequence of solutions corresponding to the hierarchic sequence of finite element spaces $S_{i-2} \subset S_{i-1} \subset S_i$. Let us denote the corresponding computed potential energy values by π_{i-2}, π_{i-1}, π_i and the degrees of freedom by N_{i-2}, N_{i-1}, N_i. We will denote the estimate for $\pi(u_{EX})$ by π_∞. With this notation we have:

$$\pi_i - \pi_\infty \approx \frac{C^2}{N_p^{2\beta}} \tag{1.96}$$

$$\pi_{i-1} - \pi_\infty \approx \frac{C^2}{N_{i-1}^{2\beta}}. \tag{1.97}$$

On dividing eq. (1.96) with eq. (1.97) and taking the logarithm we get

$$\log \frac{\pi_i - \pi_\infty}{\pi_{i-1} - \pi_\infty} \approx 2\beta \log \frac{N_{i-1}}{N_i} \tag{1.98}$$

and, repeating with $i-1$ substituted for i, it is possible to eliminate 2β to obtain:

$$\frac{\pi_i - \pi_\infty}{\pi_{i-1} - \pi_\infty} \approx \left(\frac{\pi_{i-1} - \pi_\infty}{\pi_{i-2} - \pi_\infty} \right)^Q \tag{1.99}$$

where

$$Q = \log \frac{N_{i-1}}{N_i} \left(\log \frac{N_{i-2}}{N_{i-1}} \right)^{-1}.$$

Equation (1.99) can be solved for π_∞ to obtain an estimate for the exact value of the potential energy.

The relative error in energy norm corresponding to the ith finite element solution in the sequence is estimated from

$$e_i \approx \left(\frac{\pi_i - \pi_\infty}{|\pi_\infty|} \right)^{1/2}. \tag{1.100}$$

Usually the percent relative error is reported. This estimator has been tested against the known exact solution of many problems of various smoothness. The results have shown that it works well for a wide range of problems, including most problems of practical interest; however, it cannot be guaranteed to work well for all conceivable problems. For example, this method would fail if the exact solution would happen to be energy-orthogonal to all basis functions associated with (say) odd values of i.

Remark 1.12 From equation (1.92) we get

$$\log (e_r)_E \approx \log C - \beta \log N. \tag{1.101}$$

On plotting $(e_r)_E$ vs. N on log-log scale a straight line with the slope $-\beta$ will be seen for sufficiently large N. The estimated value of β, corresponding to the ith solution in the sequence, is denoted by β_i. It is computed from eq. (1.98):

$$\beta_i = \frac{1}{2} \frac{\log(\pi_i - \pi) - \log(\pi_{i-1} - \pi)}{\log N_{i-1} - \log N_i}. \tag{1.102}$$

Examples

The properties of the finite element solution with reference to a family of model problems is discussed in the following. The problems are stated as follows: Find $u_{FE} \in S^0(I)$ such that

$$\int_0^\ell \left(\kappa u'_{FE} v' + c u_{FE} v\right)\, dx = F(v) \quad \text{for all } v \in S^0(I) \tag{1.103}$$

where κ and c are constants and $F(v)$ is defined such that the exact solution is:

$$u_{EX} = x^\alpha(\ell - x), \quad \text{on } I = (0, \ell), \quad \alpha > 1/2. \tag{1.104}$$

As explained in Section 1.5.1, when α is not an integer, the case considered in the following, then this solution lies in the space $H^{\alpha+1/2-\epsilon}(I)$. Therefore the asymptotic rate of h-convergence on uniform meshes, predicted by eq. (1.92), is $\beta = \alpha - 1/2$ and the asymptotic rate of p-convergence on a fixed mesh is $\beta = 2\alpha - 1$.

We selected this problem because it is representative of the singular part of the exact solutions of two-and three-dimensional elliptic boundary value problems.

Referring to Theorem 1.3, we have $B(u_{EX} - u_{FE}, v) = 0$ for all $v \in S^0(I)$ therefore $F(v) = B(u_{EX}, v)$. Consequently for the kth element the load vector in the local numbering convention is:

$$r_i^{(k)} = \int_{x_k}^{x_{k+1}} (\kappa u'_{EX} \varphi'_i + c u_{EX} \varphi_i)\, dx, \quad i = 1, 2, \dots, p_k + 1 \tag{1.105}$$

where by definition $\varphi_i(Q_k(\xi)) = N_i(\xi)$.

When $1/2 < \alpha < 1$ then the first derivative of u_{EX} is infinity in the point $x = 0$. To avoid having u'_{EX} in the integrand, the first term in eq. (1.105) is integrated by parts:

$$\int_{x_k}^{x_{k+1}} \kappa u'_{EX} \varphi'_i\, dx = \left(\kappa u_{EX} \varphi'_i\right)_{x_k}^{x_{k+1}} - \int_{x_k}^{x_{k+1}} \kappa u_{EX} \varphi''_i\, dx.$$

Since $\varphi''_i = 0$ for $i = 1$ and $i = 2$, we have:

$$r_1^{(k)} = -\frac{1}{\ell_k}\left(\kappa u_{EX}\right)_{x=x_{k+1}} + \frac{1}{\ell_k}\left(\kappa u_{EX}\right)_{x=x_k} + \frac{\ell_k}{2}\int_{-1}^{1} (c u_{EX})_{x=Q_k(\xi)} N_1\, d\xi$$

$$r_2^{(k)} = \frac{1}{\ell_k}\left(\kappa u_{EX}\right)_{x=x_{k+1}} - \frac{1}{\ell_k}\left(\kappa u_{EX}\right)_{x=x_k} + \frac{\ell_k}{2}\int_{-1}^{1} (c u_{EX})_{x=Q_k(\xi)} N_2\, d\xi$$

and for $i \geq 3$ we have:

$$r_i^{(k)} = \sqrt{\frac{2i-3}{2}}\frac{2}{\ell_k}\left((\kappa u_{EX})_{x=x_{k+1}} - (-1)^i (\kappa u_{EX})_{x=x_k} - \int_{-1}^{1} (\kappa u_{EX})_{x=Q_k(\xi)} \frac{dP_{i-2}}{d\xi}\, d\xi\right) + \frac{\ell_k}{2}\int_{-1}^{1} (c u_{EX})_{x=Q_k(\xi)} N_i\, d\xi \tag{1.106}$$

where $P_{i-2}(\xi)$ is the Legendre polynomial of degree $i-2$ and eq. (D.10) was used.

Since the exact solution is known, the exact value of the potential energy can be determined for any set of values of α, κ, c and ℓ. When κ and c are both constants then

$$\pi(u_{EX}) = -\frac{1}{2}\left[\kappa\left(\frac{\alpha^2}{2\alpha-1}\ell^{2\alpha-1} - (\alpha+1)\ell^{2\alpha} + \frac{(\alpha+1)^2}{2\alpha+1}\ell^{2\alpha+1}\right) + c\left(\frac{1}{2\alpha+1}\ell^{2\alpha+1} - \frac{1}{\alpha+1}\ell^{2(\alpha+1)} + \frac{1}{2\alpha+3}\ell^{2\alpha+3}\right)\right]. \tag{1.107}$$

The exact values of the potential energy for the data $\kappa = 1$, $c = 50$ and $\ell = 1$ and various values of α are shown in Table 1.2.

Table 1.2 Exact values of the potential energy for $\kappa = 1$, $c = 50$ and $\ell = 1$.

α	$\pi(u_{EX})$	α	$\pi(u_{EX})$
0.600	−2.3728354978	1.000	−1.0000000000
0.700	−1.7571858289	1.500	−0.5104166667
0.800	−1.4176885916	2.000	−0.3047619048
0.900	−1.1799028822	3.000	−0.1420634921

When α is a fractional number then derivatives higher than α will not be finite in $x = 0$. In the range $0.5 < \alpha < 1$ the first derivative in the point $x = 0$ is infinity. This range of α has considerable practical importance because the exact solutions of two- and three-dimensional problems often have analogous terms.

When α is an integer then all derivatives of u_{EX} are finite. Therefore u_{EX} can be approximated by Taylor series about any point of the domain $\bar{I} = [0, \ell]$. It is known that the error term of a Taylor series truncated at polynomial degree p is bounded by the $(p+1)$th derivative of u_{EX}:

$$\max |u_{FE} - u_{EX}| \leq \frac{\ell^{p+1}}{(p+1)!} \max_{x \in \bar{I}} \left| \frac{d^{p+1} u_{EX}}{dx^{p+1}} \right|. \tag{1.108}$$

In the special case when α is an integer and $p_{\min} \geq \alpha + 1$ then $u_{FE} = u_{EX}$.

Exercise 1.19 Show how eq. (1.106) is obtained from eq. (1.105). Provide details.

Example 1.10 Let us consider model problems in the form of eq. (1.103) with the following data: $\ell = 1$, $\kappa = 1$, $c = 50$ and exact solutions in the form of eq. (1.104) corresponding to $\alpha = 0.6,\ 0.7,\ 0.8,\ 0.9$. We will use a sequence of uniform finite element meshes with $M(\Delta) = 10,\ 100,\ 1000$ and $p_k = p = 2$ assigned to all elements. We are interested in the relationship between the estimated and true relative errors. The computed values of the potential energy and their estimated limit values computed by means of eq. (1.99) are listed in Table 1.3. These are comparable to the exact values of the potential energy listed in Table 1.2. The estimated limit values of the potential energy are denoted by $\pi_{M(\Delta)\to\infty}$.

With the information provided in Tables 1.2 and 1.3 it is possible to compare the estimated and exact values of the relative error. For example, using eq. (1.100) and

$$\|(u_{FE})_{M(\Delta)}\|^2_{E(I)} = |\pi_{M(\Delta)}|$$

Table 1.3 Example: Computed and estimated values of the potential energy π. Uniform mesh refinement, $p_k = p = 2$ for all elements.

$M(\Delta)$	N	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
10	19	−2.17753673	−1.73038992	−1.41382648	−1.17955239
100	199	−2.25079984	−1.74673700	−1.41675042	−1.17984996
1000	1999	−2.29589857	−1.75303348	−1.41745363	−1.17989453
$\pi_{M(\Delta)\to\infty}$		−2.37254083	−1.75716094	−1.41768637	−1.17990276

the estimated relative error in energy norm for $M(\Delta) = 10$, $\alpha = 0.8$ is:

$$(e_r^*)_E = \sqrt{\frac{\pi(u_{FE}) - \pi_{M(\Delta)\to\infty}}{|\pi_{M(\Delta)\to\infty}|}} = \sqrt{\frac{-1.41382648 + 1.41768637}{1.41768637}} = 0.0522$$

or 5.22%. When using the exact value of the potential energy for reference then the relative error is the same as the estimated relative error to within three digits of accuracy:

$$(e_r)_E = \sqrt{\frac{\pi(u_{FE}) - \pi(u_{EX})}{\|u_{EX}\|^2_{E(I)}}} = \sqrt{\frac{-1.41382648 + 1.41768859}{1.41768859}} = 0.0522.$$

Exercise 1.20 Compare the estimated and exact values of the relative error in energy norm for the problem in Example 1.10 for $M(\Delta) = 100$, $\alpha = 0.7$.

Example 1.11 Let us consider once again model problems in the form of eq. (1.103) with the data: $\ell = 1$, $\kappa = 1$, $c = 50$ and exact solutions corresponding to $\alpha = 0.6,\ 0.7,\ 0.8,\ 0.9$, see eq. (1.104). Using a sequence of uniform finite element meshes with $M(\Delta) = 10,\ 100,\ 1000,\ 10{,}000$ and $p = 2$ assigned to each element, the results shown in Fig. 1.9 are obtained. The values of β were computed by linear regression using eq. (1.101). We observe that $\beta = \alpha - 1/2$. This is consistent with the asymptotic estimate given by eq. (1.91).

Example 1.12 Let us consider model problems in the form of eq. (1.103) with the data: $\ell = 1$, $\kappa = 1$, $c = 50$ and exact solutions corresponding to $\alpha = 0.6,\ 0.7,\ 0.8,\ 0.9$, see eq. (1.104). Using a uniform finite element mesh with $M(\Delta) = 10$ and $p = 2,\ 3,\ 4,\ 5$ assigned to each element, the results shown in Fig. 1.10 are obtained. The values of β were computed by linear regression using eq. (1.101). We observe that $\beta = 2(\alpha - 1/2)$, that is, the rate of convergence is twice that in Example 1.11. This is consistent with the theoretical results in [22, 84]: The rate of p-convergence is at least twice the rate of h-convergence when the singular point is a nodal point.

1.5.4 Error in the extracted QoI

In Example 1.9 it was demonstrated that the QoI can be extracted from the finite element solution efficiently and accurately even when the discretization was very poorly chosen. Let us consider a quantity of interest $\Phi(u)$ and the corresponding extraction function $w \in E(I)$. The extracted value of the QoI is

$$\Phi(u_{FE}) = F(w) - B(u_{FE}, w) \tag{1.109}$$

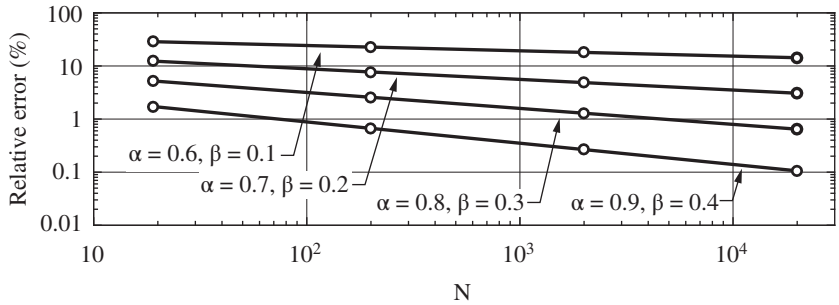

Figure 1.9 Relative error in energy norm. $M(\Delta) = 10,\ 100,\ 1000,\ 10000$, $p = 2$.

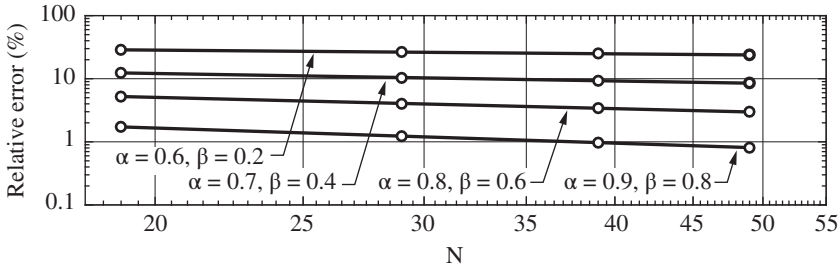

Figure 1.10 Relative error in energy norm. $M(\Delta) = 10$, $p = 2,\ 3,\ 4,\ 5$.

and the exact value of the QoI is

$$\Phi(u_{EX}) = F(w) - B(u_{EX}, w). \tag{1.110}$$

Subtracting eq. (1.109) from eq. (1.110) we get

$$\Phi(u_{EX}) - \Phi(u_{FE}) = -B(u_{EX} - u_{FE}, w). \tag{1.111}$$

We define a function $z_{EX} \in E^0(I)$ such that

$$B(z_{EX}, v) = B(w, v) \quad \text{for all } v \in E^0(I). \tag{1.112}$$

This operation projects $w \in E(I)$ onto the space $E^0(I)$. Letting $v = u_{EX} - u_{FE}$ we get:

$$B(z_{EX}, u_{EX} - u_{FE}) = B(w, u_{EX} - u_{FE}) \quad \text{for all } v \in E^0(I).$$

We will write this as

$$B(u_{EX} - u_{FE}, w) = B(u_{EX} - u_{FE}, z_{EX}). \tag{1.113}$$

Next we define $z_{FE} \in S^0(I)$ such that

$$B(z_{EX}, v) = B(z_{FE}, v) \quad \text{for all } v \in S^0(I). \tag{1.114}$$

This operation projects $z_{EX} \in E^0(I)$ onto the space $S^0(I)$. By Galerkin's orthogonality condition (see Theorem 1.3) we have

$$B(u_{EX} - u_{FE}, v) = 0 \quad \text{for all } v \in S^0(I).$$

Therefore, letting $v = z_{FE}$, we write eq. (1.113) as

$$B(u_{EX} - u_{FE}, w) = B(u_{EX} - u_{FE}, z_{EX} - z_{FE}) \tag{1.115}$$

and we can write eq. (1.111) as

$$\Phi(u_{EX}) - \Phi(u_{FE}) = -B(u_{EX} - u_{FE}, z_{EX} - z_{FE}). \tag{1.116}$$

Therefore the error in the extracted data is

$$\begin{aligned} |\Phi(u_{EX}) - \Phi(u_{FE})| &= |B(u_{EX} - u_{FE}, z_{EX} - z_{FE})| \\ &\leq 2\|u_{EX} - u_{FE}\|_{E(I)} \|z_{EX} - z_{FE}\|_{E(I)} \end{aligned} \tag{1.117}$$

where we used the Schwarz inequality, see Section A.3 in the appendix.

The function z_{FE} made it possible to write the error in the QoI in this form. It does not have to be computed.

Inequality (1.117) serves to explain why the error in the extracted data can converge to zero faster than the error in energy norm: If $\|z_{EX} - z_{FE}\|_{E(I)}$ is of comparable magnitude to $\|u_{EX} - u_{FE}\|_{E(I)}$

then the error in the extracted data is of comparable magnitude to the error in the strain energy, that is, the error in energy norm squared. But, as seen in Example 1.9, where w was much smoother than u_{EX}, it can be much smaller. In the exceptional case when the extraction function is Green's function, the error is zero.

1.6 The choice of discretization in 1D

In an ideal discretization the error (in energy norm) associated with each element would be the same. This ideal discretization can be approximated by automated adaptive methods in which the discretization is modified based on feedback information from previously obtained finite element solutions. Alternatively, based on a general understanding of the relationship between regularity and discretization, and understanding the strengths and limitations of the software tools available to them, analysts can formulate very efficient discretization schemes.

1.6.1 The exact solution lies in $H^k(I)$, $k - 1 > p$

When the solution is smooth then the most efficient finite element discretization scheme is uniform mesh and high polynomial degree. However, all implementations of finite element analysis software have limitations on how high the polynomial degree is allowed to be and therefore it may not be possible to increase the polynomial degree sufficiently to achieve the desired accuracy. In such cases the mesh has to be refined. Uniform refinement may not be optimal in all cases, however. Consider, for example, the following problem:

$$-\epsilon^2 u'' + cu = f(x), \quad u(0) = u'(\ell) = 0 \tag{1.118}$$

where $\epsilon << c$, and f is a smooth function. Intuitively, when ϵ^2 is small then the solution will be close to $u = f/c$ however, because of the boundary condition $u(0) = 0$, has to be satisfied, the function $u(x)$ will change sharply over some interval $0 < x < d(\epsilon) << \ell$.

Letting $c = 1$ and $f(x) = 1$ the exact solution of this problem is

$$u_{EX}(x) = 1 - \cosh x/\epsilon + \tanh(\ell/\epsilon) \sinh x/\epsilon \tag{1.119}$$

which is plotted for various values of ϵ on the interval $0 < x/\ell < 0.20$ in Fig. 1.11. It is seen that the gradient at $x = 0$ rapidly increases with respect to decreasing values of ϵ.

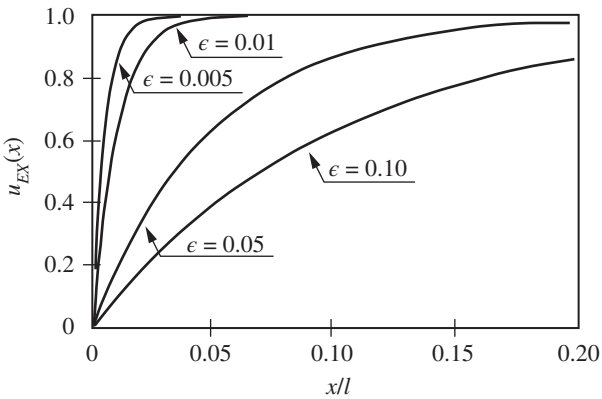

Figure 1.11 The solution $u_{EX}(x)$, given by eq. (1.119), in the neighborhood of $x = 0$ for various values of ϵ.

This is a simple example of boundary layer problems that arise in models of plates, shells and fluid flow. Despite the fact that u_{EX} is an analytic function, it may require unrealistically high polynomial degrees to obtain a close approximation to the solution when ϵ is small.

The optimal discretization scheme for problems with boundary layers is discussed in the context of the *hp*-version in [85]. The results of analysis indicate that the size of the element at the boundary is proportional to the product of the polynomial degree p and the parameter ϵ. Specifically, for the problem discussed here, the optimal mesh consists of two elements with the node points located at $x_1 = 0, x_2 = d, x_3 = \ell$, where $d = Cp\epsilon$ with $0 < C < 4/e$.

A practical approach to problems like this is to create an element at the boundary (in higher dimensions a layer of elements) the size of which is controlled by a parameter. The optimal value of that parameter is then selected adaptively.

1.6.2 The exact solution lies in $H^k(I)$, $k - 1 \leq p$

In this section we consider a special case of the problem stated in eq. (1.103):

$$\int_0^\ell u'v'\,dx = F(v), \quad \text{for all } v \in E^0(I) \tag{1.120}$$

with the data $u(0) = u(\ell) = 0$, $\ell = 1$ and $F(v)$ defined such that the exact solution is

$$u_{EX} = x^\alpha(1-x), \quad \alpha > 1/2, \quad 0 < x < 1 \tag{1.121}$$

that is,

$$F(v) = \int_0^\ell (\alpha x^{\alpha-1} - (\alpha+1)x^\alpha)\, v'\,dx. \tag{1.122}$$

On integrating by parts, we get the following expression which is better suited for numerical evaluation:

$$F(v) = -\int_0^\ell u_{EX}v''dx. \tag{1.123}$$

We address the following questions: (a) How does the error in energy norm depend on the parameter α, the mesh Δ and the p-distribution $\mathbf{p}$? and (b) How is this error distributed among the elements? Understanding these relationships is necessary for making sound choices of discretization based on a priori information concerning the regularity of the exact solution.

We compute the potential energy of the difference between the exact solution and its linear interpolant for the kth element:

$$\overline{\pi}_{EX}^{(k)} = \frac{1}{2}\int_{x_k}^{x_{k+1}} \left(u'_{EX} - \frac{u_{EX}(x_{k+1}) - u_{EX}(x_k)}{x_{k+1} - x_k}\right)^2 dx.$$

The exact solution for $\alpha = 0.75$ and its linear interpolant for $M(\Delta) = 5$, uniform mesh, are shown in Fig. 1.12.

To obtain the potential energy of the difference between the exact solution and its linear interpolant for the kth element, denoted by $\overline{\pi}_{FE}^{(k)}$, we need to solve:

$$\frac{2}{\ell_k}\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{Bmatrix} a_3^{(k)} \\ a_4^{(k)} \\ \vdots \\ a_{p_k+1}^{(k)} \end{Bmatrix} = \begin{Bmatrix} r_3^{(k)} \\ r_4^{(k)} \\ \vdots \\ r_{p_k+1}^{(k)} \end{Bmatrix}. \tag{1.124}$$

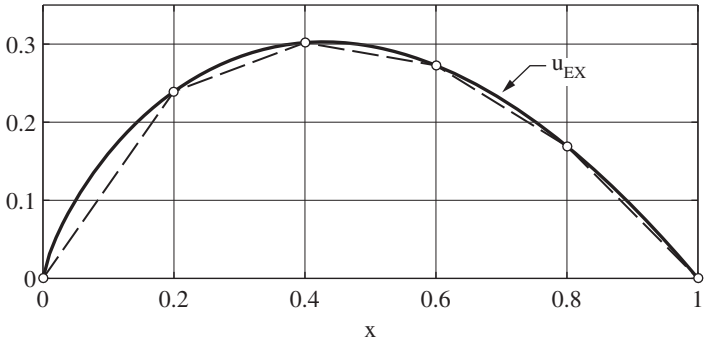

Figure 1.12 The exact solution for $\alpha = 0.75$ and its linear interpolant for $M(\Delta) = 5$, uniform mesh.

The solution is

$$a_i^{(k)} = \frac{\ell_k}{2} r_i^{(k)}, \quad i = 3, 4, \ldots, p_k + 1. \tag{1.125}$$

Using eq. (1.106) we get

$$r_i^{(k)} = -\sqrt{\frac{2i-3}{2}} \frac{2}{\ell_k} \int_{-1}^{1} \left(\kappa \, \tilde{u}_{EX}^{(k)} \right)_{x=Q_k(\xi)} \frac{dP_{i-2}}{d\xi} \, d\xi, \quad i = 3, 4, \ldots, p_k + 1 \tag{1.126}$$

where $\tilde{u}_{EX}^{(k)}$ is the difference between u_{EX} and its linear interpolant:

$$\tilde{u}_{EX}^{(k)} = (u_{EX})_{x=Q_k(\xi)} - \left(\frac{1-\xi}{2} u_{EX}(x_k) + \frac{1+\xi}{2} u_{EX}(x_{k+1}) \right) \tag{1.127}$$

and compute

$$\overline{\pi}_{FE}^{(k)} = -\frac{1}{2} \sum_{i=3}^{p_k+1} a_i^{(k)} r_i^{(k)}.$$

Referring to Theorem 1.5, the error in energy norm associated with the kth element is

$$\| e_k \|_{E(I_k)} = \sqrt{\overline{\pi}_{FE}^{(k)} - \overline{\pi}_{EX}^{(k)}} \tag{1.128}$$

and the relative error in energy norm associated with the kth element is:

$$(e_r^{(k)})_E = \frac{\| e_k \|_{E(I_k)}}{\sqrt{\left| \overline{\pi}_{EX}^{(k)} \right|}}. \tag{1.129}$$

The error of approximation over the entire domain is:

$$\| u_{EX} - u_{FE} \|_{E(I)} = \left(\sum_{k=1}^{M(\Delta)} \| e_k \|_{E(I_k)}^2 \right)^{1/2} = \sqrt{\pi_{FE} - \pi_{EX}} \tag{1.130}$$

By Theorem 1.2, the exact value of the potential energy is

$$\pi(u_{EX}) = -\frac{1}{2} \int_0^1 (u'_{EX})^2 dx = -\frac{1}{2} \left(\frac{\alpha^2}{2\alpha - 1} - (\alpha + 1) + \frac{(\alpha+1)^2}{2\alpha + 1} \right) \tag{1.131}$$

and the relative error in energy norm on the entire domain is:

$$(e_r)_E = \left(\frac{\pi_{FE} - \pi_{EX}}{|\pi(u_{EX})|} \right)^{1/2}. \tag{1.132}$$

Remark 1.13 In estimating the local error we used $u_{FE}(x_k) = u_{EX}(x_k)$. It can be shown that in the special case of this problem ($c = 0$) this relationship holds and therefore using the equal sign in eq. (1.128) is justified. In the general case ($c \neq 0$) however, $u_{FE}(x_k) \neq u_{EX}(x_k)$ and eq. (1.128) will be an estimate of the local error in the finite element solution. Therefore the equal sign in eq. (1.128) has to be replaced by the approximately equal ($\approx$) sign and the first equal sign in eq. (1.130) has to be replaced with the less or equal ($\leq$) sign.

Example 1.13 This example illustrates the distribution of the relative error among the elements for a fixed mesh and polynomial degree for selected fractional values of α. Uniform mesh on the domain $(0, 1)$ with $M(\Delta) = 5$ and $p_k = 2$ for $k = 1, 2, \ldots, 5$ is used. The exact solution for $\alpha = 0.75$ is shown in Fig. 1.12. The percent relative error in energy norm associated with the kth element, given by eq. (1.129), is shown in Table 1.4 and the relative error for the entire domain is shown in the last column.

It is seen that for all values of α the maximum error is associated with the first element.

Example 1.14 This example illustrates the distribution of the relative error among the elements for a fixed mesh and polynomial degree for selected integer values of α. Uniform mesh on the domain $(0, 1)$ with $M(\Delta) = 5$ and $p_k = 2$ for $k = 1, 2, \ldots, 5$ is used. The percent relative error in energy norm associated with the kth element, given by eq. (1.129), is shown in Table 1.5 and the relative error for the entire domain is shown in the last column.

The error of approximation for $\alpha = 1$ is zero. This follows directly from Theorem 1.4: The exact solution is a polynomial of degree 2. Therefore it lies in the finite element space and hence the finite element solution is the same as the exact solution.

Remark 1.14 In the foregoing discussion it was tacitly assumed that all data computed by numerical integration were accurate and the coefficient matrices of the linear equations were such that small changes in the right-hand-side vector produce small changes in the solution vector. This happens when the condition number of the coefficient matrix is reasonably small. In the finite element method the condition number depends on the choice of the shape functions, the mapping functions and the mesh. In one-dimensional setting the mapping is linear and the shape functions

Table 1.4 Example: Element-by-element and total relative errors in energy norm (percent) for selected fractional values of α.

	Element number					
α	1	2	3	4	5	$(e_r)_E$
1.25	79.49	7.50	2.80	1.63	1.12	4.80
1.15	99.52	4.06	1.63	0.97	0.67	3.92
1.05	29.56	1.24	0.53	0.32	0.22	1.77
0.95	18.89	1.16	0.52	0.32	0.22	2.41
0.85	42.94	3.26	1.52	0.94	0.67	9.84
0.75	60.39	5.14	2.47	1.56	1.11	22.37
0.65	76.07	6.86	3.39	2.16	1.56	42.91
0.55	91.80	8.44	4.28	2.76	2.00	76.22

Table 1.5 Example: Element-by-element and total relative errors in energy norm (percent) for selected integer values of α.

	Element number					
α	1	2	3	4	5	$(e_r)_E$
1	0	0	0	0	0	0
2	11.00	61.24	15.31	7.02	4.55	2.45
3	20.09	4.69	98.83	16.16	9.24	4.62
4	36.98	8.41	17.24	30.13	14.02	8.00

are energy-orthogonal, therefore round-off errors are not significant. This is not the case in two and three dimensions, however.

Errors in numerical integration can be particularly damaging. The reader should be mindful of this when applying the concepts and procedures discussed in this chapter to higher dimensions.

1.7 Eigenvalue problems

The following problem is a prototype of an important class of engineering problems which includes the undamped vibration of elastic structures:

$$-(\kappa u')' + cu = -\mu \frac{\partial^2 u}{\partial t^2}, \quad x \in (0, \ell), \quad t \in (0, \infty) \tag{1.133}$$

where the primes represent differentiation with respect to x. For example, we may think of an elastic bar of length ℓ, cross-section A, modulus of elasticity E, in which case $\kappa \equiv AE > 0$ given in units of Newton (N) or equivalent, the parameter $c \geq 0$ is the coefficient of distributed springs (N/mm^2) and the parameter $\mu > 0$ is mass per unit length (kg/m $= 10^{-6}$Ns^2/mm^2). The bar is vibrating in its longitudinal direction.

The boundary conditions are:

$$u(0, t) = 0, \quad u(\ell, t) = 0$$

and the initial conditions are

$$u(x, 0) = f(x), \quad \frac{\partial u}{\partial t}|_{(x,0)} = g(x)$$

where $f(x)$ and $g(x)$ are given functions in $L^2(I)$. Here we consider homogeneous Dirichlet boundary conditions. However, the boundary conditions can be homogeneous Neumann or homogeneous Robin conditions, or any combination of those.

The generalized form is obtained by multiplying eq. (1.133) by a test function $v \in E^0(I)$ and integrating by parts:

$$\int_0^\ell (\kappa u' v' + cuv)\, dx = -\int_0^\ell \mu \frac{\partial^2 u}{\partial t^2} v\, dx. \tag{1.134}$$

We now introduce $u = U(x)T(t)$ where $U \in E^0(I)$, $T \in C^2(0, \infty)$. This is known as separation of variables. Therefore we get

$$T \int_0^\ell (\kappa U' v' + cUv)\, dx = -\frac{\partial^2 T}{\partial t^2} \int_0^\ell \mu U v\, dx \tag{1.135}$$

which can be written as

$$\frac{\int_0^\ell (\kappa U' v' + cUv)\, dx}{\int_0^\ell \mu U v\, dx} = -\frac{1}{T}\frac{\partial^2 T}{\partial t^2} = \omega^2. \tag{1.136}$$

Since the functions on the left are independent of t, the function T depends only on t, both expressions must equal some positive constant denoted by ω^2. That constant has to be positive because the expression on the left holds for all $v \in E^0(I)$ and if we select $v = U$ then the expression on the left is positive.

The function $T(t)$ satisfies the ordinary differential equation

$$\frac{\partial^2 T}{\partial t^2} + \omega^2 T = 0 \tag{1.137}$$

the solution of which is

$$T = a\cos(\omega t) + b\sin(\omega t) \tag{1.138}$$

where ω is the angular velocity (rad/s). Alternatively ω is written as $\omega = 2\pi f$ where f is the frequency (Hz).

To find ω and U we have to solve the problem

$$\int_0^\ell (\kappa U' v' + cUv)\, dx - \omega^2 \int_0^\ell \mu U v\, dx = 0 \quad \text{for all } v \in E^0(I) \tag{1.139}$$

which will be abbreviated as

$$B(U, v) - \omega^2 D(U, v) = 0 \quad \text{for all } v \in E^0(I). \tag{1.140}$$

There are infinitely many solutions called eigenpairs (ω_i, U_i), $i = 1, 2, \dots, \infty$. The set of eigenvalues is called the spectrum. If U_i is an eigenfunction and α is a real number then αU_i is also an eigenfunction. In the following we assume that the eigenfunctions have been normalized so that

$$D(U_i, U_i) \equiv \int_0^\ell \mu U_i^2\, dx = 1.$$

If the eigenvalues are distinct then the corresponding eigenfunctions are orthogonal: Let (ω_i, U_i) and (ω_j, U_j) be eigenpairs, $i \neq j$. Then from eq. (1.140) we have

$$\begin{aligned} B(U_i, U_j) - \omega_i^2 D(U_i, U_j) &= 0 \\ B(U_j, U_i) - \omega_j^2 D(U_j, U_i) &= 0. \end{aligned}$$

Subtracting the second equation from the first we see that if $\omega_i \neq \omega_j$ then U_i and U_j are orthogonal functions:

$$D(U_i, U_j) \equiv \int_0^\ell \mu U_i U_j\, dx = 0 \tag{1.141}$$

and hence $B(U_i, U_j) = 0$.

Importantly, it can be shown that any function $f \in E^0(I)$ can be written as a linear combination of the eigenfunctions:

$$\left\| f - \sum_{i=1}^{\infty} a_i U_i(x) \right\|_{L^2(I)} = 0 \tag{1.142}$$

where

$$a_i = \int_0^{\ell} \mu f U_i \, dx. \tag{1.143}$$

The Rayleigh[15] quotient is defined by

$$R(u) = \frac{B(u,u)}{D(u,u)}. \tag{1.144}$$

Eigenvalues are usually numbered in ascending order. Following that convention,

$$\omega_1^2 \equiv \omega_{\min}^2 = \min_{u \in E^0(I)} R(u) = R(U_1) \tag{1.145}$$

that is, the smallest eigenvalue is the minimum of the Rayleigh quotient and the corresponding eigenfunction is the minimizer of $R(u)$ on $E^0(I)$. This follows directly from eq. (1.140). The kth eigenvalue minimizes $R(u)$ on the space $E_k^0(I)$

$$\omega_k^2 = \min_{u \in E_k^0(I)} R(u) = R(U_k) \tag{1.146}$$

where

$$E_k^0(I) = \{u \mid u \in E^0(I),\ B(u, U_i) = 0,\ i = 1, 2, \dots, k-1\}. \tag{1.147}$$

When the eigenvalues are computed numerically then the minimum of the Rayleigh quotient is sought on the finite-dimensional space $S^0(I)$. We see from the definition $R(u)$ that the error of approximation in the natural frequencies will depend on how well the eigenfunctions are approximated in energy norm, in the space $S^0(I)$.

The following example illustrates that in a sequence of numerically computed eigenvalues only the lower eigenvalues will be approximated well. It is possible, however, at least in principle, to obtain good approximation for any eigenvalue by suitably enlarging the space $S^0(I)$.

Example 1.15 Let us consider the eigenvalue problem

$$\kappa \frac{\partial^2 u}{\partial x^2} = \mu \frac{\partial^2 u}{\partial t^2}, \qquad u(0) = u(\ell) = 0,\ t \geq 0. \tag{1.148}$$

This equation models (among other things) the free vibration (natural frequencies and mode shapes) of a string of length ℓ stretched horizontally by the force $\kappa > 0$ (N) under the assumptions that the displacements are infinitesimal and confined to one plane, the plane of vibration, and the ends of the string are fixed. The mass per unit length is $\mu > 0$ (kg/m). We assume that κ and μ are constants. It is left to the reader to verify that the function u defined by

$$u = \sum_{i=1}^{\infty} \left(a_i \cos(\omega_i t) + b_i \sin(\omega_i t)\right) \sin(\lambda_i x) \tag{1.149}$$

where a_i, b_i are coefficients determined from the initial conditions and

$$\lambda_i = i\frac{\pi}{\ell}, \quad \omega_i = \lambda_i \sqrt{\frac{\kappa}{\mu}} \tag{1.150}$$

satisfies eq. (1.148).

If we approximate the eigenfunctions using uniform mesh, $p = 2$ and plot the ratio $(\omega_{FE}/\omega_{EX})_n$ against n/N, where n is the nth eigenvalue, then we get the curves shown in Fig. 1.13. The curves show that somewhat more than 20% of the numerically computed eigenvalues will be accurate.

15 John William Strutt, 3rd Baron Rayleigh 1842–1919.

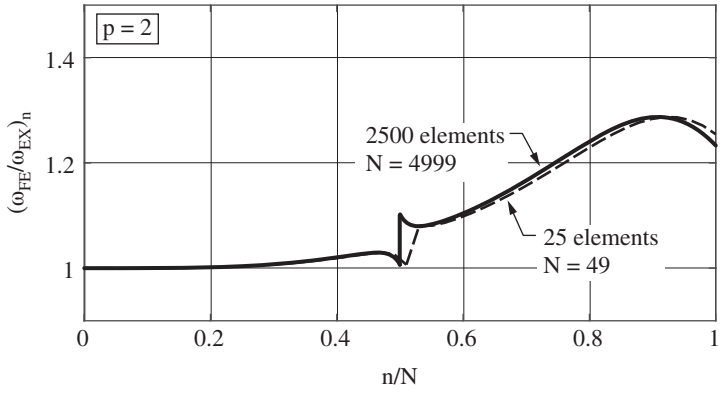

Figure 1.13 The ratio $(\omega_{FE}/\omega_{EX})_n$ corresponding to the *h* version, $p = 2$.

The higher eigenvalues cannot be well approximated in the space $S^0(I)$. The existence of the jump seen at $n/N = 0.5$ is a feature of numerically approximated eigenvalues by means of standard finite element spaces using the *h*-version [2]. The location of the jump depends on the polynomial degree of elements. There is no jump when $p = 1$.

If we approximate the eigenfunctions using a uniform mesh consisting of 5 elements, and increase the polynomial degrees uniformly then we get the curves shown in Fig. 1.14. The curves show that only about 40% of the numerically computed eigenvalues will be accurate. The error increases monotonically for the higher eigenvalues and the size of the error is virtually independent of *p*.

It is possible to reduce this error by enforcing the continuity of derivatives. Examples are available in [32]. There is a tradeoff, however: Enforcing continuity of derivatives on the basis functions reduces the number of degrees of freedom but entails a substantial programming burden because an adaptive scheme has to be devised for the general case to ensure that the proper degree of continuity is enforced. If, for example, μ would be a piecewise constant function then the continuity of the first and higher derivatives must not be enforced in those points where μ is discontinuous.

From the perspective of designing a finite element software, it is advantageous to design the software in such a way that it will work well for a broad class of problems. In the formulation presented in this chapter C^0 continuity is a requirement. Functions that lie in $C^k(I)$ where $k > 0$

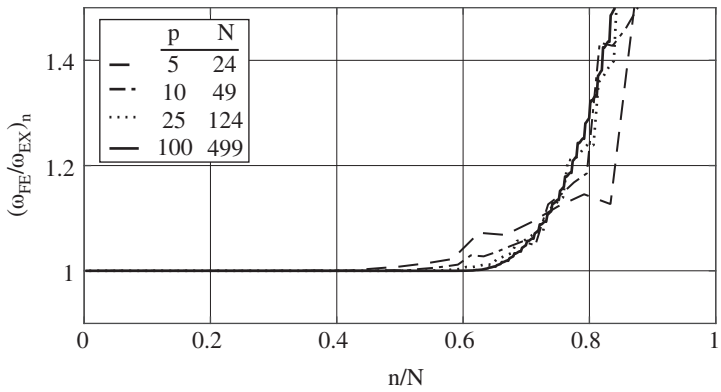

Figure 1.14 The ratio $(\omega_{FE}/\omega_{EX})_n$ corresponding to the *p* version. Uniform mesh, 5 elements.

Table 1.6 Example: p-Convergence of the 24th eigenvalue in Example 1.16.

p	5	10	15	20
ω_{24}	194.296	100.787	98.312	98.312

are also in $C^0(I)$. In other words, the space $C^k(I)$ is embedded in the space $C^0(I)$. Symbolically: $C^k(I) \subset C^0(I)$. The exact eigenfunctions in this example are in $C^\infty(I)$.

Example 1.16 Let us consider the problem in Example 1.15 modified so that μ is a piecewise constant function defined on a uniform mesh of 5 elements such that $\mu = 1$ on elements 1, 3 and 5, $\mu = 0.2$ on elements 2 and 4. In this case the exact eigenfunctions are not smooth and the exact eigenvalues are not known explicitly.

At $p = 5$ there are 24 degrees of freedom. Suppose that the 24th eigenvalue is of interest. If we increase p uniformly then this eigenvalue converges to 98.312. The results of computation are shown in Table 1.6.

Any eigenvalue can be approximated to an arbitrary degree of precision on a suitably defined mesh and uniform increase in the degrees of freedom. When κ and/or μ are discontinuous functions then the points of discontinuity must be node points.

Observe that the numerically computed eigenvalues converge monotonically from above. This follows directly from the fact that the eigenfunctions are minimizers of the Rayleigh quotient.

Exercise 1.21 Prove eq. (1.143).

Exercise 1.22 Find the eigenvalues for the problem of Example 1.15 using the generalized formulation and the basis functions $\varphi_n(x) = \sin(n\pi x/\ell)$, $(n = 1, 2, \dots, N)$. Assume that κ and μ are constants and $\mu/\kappa = 1$. Let $\ell = 10$. Explain what makes this choice of basis functions very special. Hint: Owing to the orthogonality of the basis functions, only hand calculations are involved.

1.8 Other finite element methods

Up to this point we have been concerned with the finite element method based on the generalized formulation, called the principle of virtual work. There are many other finite element methods. All finite element methods share the following attributes:

1. Formulation. A bilinear form $B(u, v)$ is defined on the normed linear spaces X, Y (i.e. $u \in X$, $v \in Y$) and the functional $F(v)$ is defined on Y. The exact solution u_{EX} lies in X and satisfies:

$$B(u_{EX}, v) = F(v) \quad \text{for all } v \in Y \tag{1.151}$$

   The normed linear spaces, X, Y, the linear functional F and the bilinear form B satisfy the respective properties listed in sections A.1.1 and A.1.2.
2. Finite element spaces. The finite-dimensional subspaces $S_i \subset X, V_i \subset Y$ $(i = 1, 2, \dots)$ are defined and it is assumed that there are $\hat{u}_i \in S_i$ such that the sequence of functions $\hat{u}_i$ $(i = 1, 2, \dots)$ converges in the space X to u_{EX}, that is:

$$\|u_{EX} - \hat{u}_i\|_X \leq \epsilon_i \quad \epsilon_i \to 0 \text{ as } i \to \infty. \tag{1.152}$$

   The functions $\hat{u}_i$ are not the finite element solutions in general.

3. The finite element solution. The finite element solution $u_{i|FE} \in S_i$ satisfies:

$$B(u_{i|FE}, v) = F(v) \quad \text{for all } v \in V_i. \tag{1.153}$$

4. The stability criterion. The finite element method is said to be stable if

$$\|u_{i|FE} - \hat{u}_i\|_X \leq C\|U - \hat{u}_i\|_X \quad i = 1, 2, \ldots \tag{1.154}$$

for all possible $U \in X$. The necessary and sufficient condition for a finite element method to be stable is that for every $u \in S_i$ there is a $v \in V_i$ so that

$$|B(u, v)| \geq C \ \|u\|_X \|v\|_Y \tag{1.155}$$

where $C > 0$ is a constant, independent of i, or for every $v \in V_i$ there is a $u \in S_i$ so that this inequality holds. This inequality is known as the Babuška-Brezzi condition, usually abbreviated to "the BB condition". This condition was formulated by Babuška in 1971 [9] and independently by Brezzi in 1974 [29].

If the Babuška-Brezzi condition is not satisfied then there will be at least some $u_{EX} \in X$ for which $\|u_{EX} - u_{i|FE}\|_X \nrightarrow 0$ as $i \to \infty$ even though there may be $u_{EX} \in X$ for which $\|u_{EX} - u_{i|FE}\|_X \to 0$ as $i \to \infty$. Examples are presented in [6]. In general it is difficult, or may even be impossible, to separate those u_{EX} for which the method works well from those for which it does not. The Babuška-Brezzi condition guarantees that the condition number of the stiffness matrix will not become too large as i increases.

Remark 1.15 Any implementation of the finite element method must be shown to satisfy the Babuška-Brezzi condition otherwise there will be some input data for which the method will fail even though it may work well for other input data. The formulation based on the principle of virtual work satisfies the Babuška-Brezzi condition.

Exercise 1.23 Show that the finite element method based on the principle of virtual work satisfies the Babuška-Brezzi condition.

1.8.1 The mixed method

Consider writing eq. (1.5) in the following form:

$$\kappa u' - F = 0 \tag{1.156}$$

$$-F' + cu = f \tag{1.157}$$

and assume that the boundary conditions are $u(0) = u(\ell) = 0$.

In the following we will use the one-dimensional equivalent of the notation introduced in sections A.2.2 and A.2.3. Multiply eq. (1.156) by $G \in L^2(I)$ and eq. (1.157) by $v \in H^1(I)$, integrate by parts and sum the resulting equations to obtain:

$$\int_0^\ell \left(\kappa \frac{du}{dx} G - FG\right) dx + \int_0^\ell \left(F \frac{dv}{dx} + cuv\right) dx = \int_0^\ell Tv dx. \tag{1.158}$$

We define the bilinear form:

$$B(u, F; v, G) \stackrel{\text{def}}{=} \int_0^\ell \left(\kappa u' G - FG\right) dx + \int_0^\ell \left(Fv' + cuv\right) dx \tag{1.159}$$

and the linear form

$$F(v) \stackrel{\text{def}}{=} \int_0^{\ell} f v \, dx. \tag{1.160}$$

The problem is now stated as follows: Find $u_{EX} \in H_0^1(I), F_{EX} \in L^2(I)$ such that

$$B\left(u_{EX}, F_{EX}; v, G\right) = F(v) \quad \text{for all } v \in H_0^1(I),\ G \in L^2(I). \tag{1.161}$$

The finite element problem is formulated as follows: Find $u_{FE} \in S^0(I)$ where $S^0(I)$ is a subspace of $H_0^1(I)$ and $F_{FE} \in V(I)$ where $V(I)$ is a subspace of $L^2(I)$ such that

$$B\left(u_{FE}, F_{FE}; v, G\right) = F(v) \quad \text{for all } v \in S^0(I),\ G \in V(I). \tag{1.162}$$

We now ask: In what sense will $\left(u_{FE}, F_{FE}\right)$ be close to $\left(u_{EX}, F_{EX}\right)$? The answer is that there is a constant C, independent of the finite element mesh and $\left(u_{EX}, F_{EX}\right)$, such that

$$\begin{aligned} &\|u_{EX} - u_{FE}\|_{H^1(I)} + \|F_{EX} - F_{FE}\|_{L^2(I)} \\ &\qquad \le C\left[\min \|u_{EX} - u\|_{H^1(I)} + \min \|F_{EX} - F\|_{L^2(I)}\right] \end{aligned} \tag{1.163}$$

provided, however, that $S^0(I)$ and $V(I)$ were properly selected.

For example, let S be the space defined in eq. (1.61) with $p_k = 1, k = 1, 2 \ldots M(\Delta)$. The space $S^0(I)$ has the dimension $M(\Delta) - 1$. For $V(I)$ consider three choices:

1. $V_1(I)$ is the set of functions which are constant on each finite element. $V_1(I)$ has the dimension $M(\Delta)$.
2. $V_2(I)$ is the space S defined in (3.11) with $p_k = 1, k = 1, 2, \ldots, M(\Delta)$ (dimension $M(\Delta) + 1$).
3. $V_3(I)$ is the set of functions which are linear on every element and discontinuous at the nodes (dimension $2M(\Delta)$).

For these choices of $V(I)$ the mixed formulation leads to systems of linear equations with $2M(\Delta) - 1$, $2M(\Delta)$ and $3M(\Delta) - 1$ unknowns, respectively. In the cases $V = V_1$ and $V = V_3$, a constant C exists such that the inequality (1.163) holds for all u_{EX}, and F_{EX}. In the case of $V = V_2$, however, such a constant does not exist. This means that no matter how large C is, there exist some $u_{EX} \in H_0^1(I)$ and $F_{EX} \in L^2(I)$ and mesh Δ so that the inequality (1.163) is not satisfied. On the other hand, there will be $u_{EX} \in H_0^1(I)$ and $F_{EX} \in L^2(I)$ for which the inequality is satisfied and therefore the finite element solutions will converge to the underlying exact solution.

1.8.2 Nitsche's method

Nitsche's method[16] allows the treatment of essential boundary conditions as natural boundary conditions. This has certain advantages in two and three dimensions. An outline of the algorithmic aspects of the method is presented in the following. For additional details we refer to [51].

Consider the problem:

$$-u'' + cu = f(x), \qquad x \in (0,\ \ell) \tag{1.164}$$

with the boundary conditions $u'(0) = 0$ and $u(\ell) = \hat{u}_\ell$. However, at $x = \ell$ we substitute the natural boundary condition:

$$u'(\ell) = \frac{1}{\epsilon}(\hat{u}_\ell - u(\ell)) \tag{1.165}$$

16 Joachim Nitsche 1926–1996.

where ϵ is a small positive number, $1/\epsilon$ is called penalty parameter. The role of the penalty parameter becomes clearly visible if we consider the potential energy

$$\Pi(u) = \frac{1}{2}\int_0^\ell \left((u')^2 + cu^2\right)\,dx + \frac{1}{2\epsilon}(u(\ell) - \hat{u}_\ell)^2 - \int_0^\ell f(x)u\,dx. \tag{1.166}$$

Letting $\epsilon \to 0$, the minimizer of the potential energy converges to the solution of the Dirichlet problem; however, the numerical problem becomes ill-conditioned. Nitsche's method stabilizes the numerical problem making it possible to solve it for the full range of boundary conditions, including $\epsilon = 0$.

Stabilization

On multiplying eq. (1.164) by v and integrating by parts we get

$$-u'(\ell)v(\ell) + \int_0^\ell (u'v' + cuv)\,dx = \int_0^\ell f(x)v\,dx. \tag{1.167}$$

We introduce the stability parameter γ and multiply eq. (1.165) by $v(\ell)\epsilon/(\epsilon + \gamma\ell)$ to get

$$\frac{1}{\epsilon + \gamma\ell}\left(\epsilon u'(\ell)v(\ell) + u(\ell)v(\ell)\right) = \frac{1}{\epsilon + \gamma\ell}\hat{u}_\ell v(\ell). \tag{1.168}$$

Adding eq. (1.167) and eq. (1.168) we get

$$\begin{aligned}\int_0^\ell (u'v' + cuv)\,dx - \frac{\gamma\ell}{\epsilon + \gamma\ell}u'(\ell)v(\ell) + \frac{1}{\epsilon + \gamma\ell}u(\ell)v(\ell)\\ = \int_0^\ell f(x)v\,dx + \frac{1}{\epsilon + \gamma\ell}\hat{u}_\ell v(\ell)\end{aligned} \tag{1.169}$$

and, multiplying eq. (1.165) by $v'(\ell)\epsilon\gamma\ell/(\epsilon + \gamma\ell)$, we have

$$\frac{\epsilon\gamma\ell}{\epsilon + \gamma\ell}u'(\ell)v'(\ell) + \frac{\gamma\ell}{\epsilon + \gamma\ell}u(\ell)v'(\ell) = \frac{\gamma\ell}{\epsilon + \gamma\ell}\hat{u}_\ell v'(\ell). \tag{1.170}$$

Subtracting eq. (1.170) from eq. (1.169) we obtain the generalized formulation:

$$\begin{aligned}\int_0^\ell (u'v' + cuv)\,dx - \frac{\gamma\ell}{\epsilon + \gamma\ell}\left(u'(\ell)v(\ell) + u(\ell)v'(\ell)\right)\\ + \frac{1}{\epsilon + \gamma\ell}u(\ell)v(\ell) - \frac{\epsilon\gamma\ell}{\epsilon + \gamma\ell}u'(\ell)v'(\ell)\\ = \int_0^\ell f(x)v\,dx + \frac{1}{\epsilon + \gamma\ell}\hat{u}_\ell v(\ell) - \frac{\gamma\ell}{\epsilon + \gamma\ell}\hat{u}_\ell v'(\ell).\end{aligned} \tag{1.171}$$

Letting $\epsilon = 0$ in eq. (1.171) we get the stabilized method proposed by Nitsche [67]:

$$\begin{aligned}\int_0^\ell (u'v' + cuv)\,dx - \left(u'(\ell)v(\ell) + u(\ell)v'(\ell)\right) + \frac{1}{\gamma\ell}u(\ell)v(\ell)\\ = \int_0^\ell f(x)v\,dx + \frac{1}{\gamma\ell}\hat{u}_\ell v(\ell) - \hat{u}_\ell v'(\ell).\end{aligned} \tag{1.172}$$

Numerical example

Letting $c = 1$, $f(x) = 1$, $\ell = 10$ and $\hat{u}_\ell = 0.25$ we construct the numerical problem using one element and the hierarchic shape functions defined in Section 1.3.1. By definition:

$$u = \sum_{j=1}^{p+1} a_j N_j(\xi), \qquad v = \sum_{i=1}^{p+1} b_i N_i(\xi) \tag{1.173}$$

Table 1.7 The computed values of $u(\ell)$.

γ	10^{-3}	10^{-6}	10^{-9}	10^{-12}	10^{-15}
$u(\ell)$	0.2540348	0.2500004	$0.25(0)_6 4$	$0.25(0)_9 4$	$0.25(0)_{12} 4$

where p is the polynomial degree. Therefore $u(\ell) = a_2$ and $v(\ell) = b_2$ and, using the Legendre shape functions, for $p = 3$ the unconstrained coefficient matrix, without the modifications of Nitsche, is

$$[M] = \frac{2}{\ell}\begin{bmatrix} 1/2 & -1/2 & 0 & 0 \\ & 1/2 & 0 & 0 \\ & \text{(sym.)} & 1 & 0 \\ & & & 1 \end{bmatrix} + \frac{c\ell}{2}\begin{bmatrix} 2/3 & 1/3 & -1/\sqrt{6} & 1/3\sqrt{10} \\ & 2/3 & -1/\sqrt{6} & -1/3\sqrt{10} \\ & \text{(sym.)} & 2/5 & 0 \\ & & & 2/21 \end{bmatrix}$$

Referring to eq. (1.172), the coefficient matrix is modified by the application of Nitsche's method. Those modifications in the present case are:

$$[N] = \begin{bmatrix} 0 & 1/\ell & 0 & 0 \\ & -1/\ell + 1/(\gamma\ell) & -\sqrt{12}/\ell & -\sqrt{18}/\ell \\ & \text{(sym.)} & 0 & 0 \\ & & & 0 \end{bmatrix}.$$

The unconstrained right hand side vector without the modifications of Nitsche is:

$$\{r\} = \{\ell/2 \quad \ell/2 - (\ell/2)\sqrt{2/3} \quad 0\}^T$$

and with the modifications of Nitsche it is:

$$\{r_N\} = \{\hat{u}_\ell/\ell \quad \hat{u}_\ell/(\gamma\ell) - \hat{u}_\ell/\ell - \hat{u}_\ell\sqrt{12}/\ell - \hat{u}_\ell\sqrt{18}/\ell\}^T.$$

The numerical results shown in Table 1.7 indicate that the stabilized formulation is remarkably robust. The notation $(0)_n$ indicates that there are n zeros.